JCTC Journal of Chemical Theory and Computation

Estimation and Inference of Diffusion Coefficients in Complex Biomolecular Environments

Christopher P. Calderon*,‡

Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005-1892, United States

Received September 9, 2010

Abstract: The 1-D diffusion coefficient associated with a charged atom fluctuating in an ion-channel binding pocket is statistically analyzed. More specifically, unconstrained and constrained molecular dynamics simulations of potassium in gramicidin A are studied. Time domain transition density based inference methods are used to fit simple stochastic differential equations and also to carry out frequentist goodness of fit tests. Particular attention is paid to varying the time between adjacent time series observations due to the well-known "non-Markovian noise" that can appear in this system due to inertia and other unresolved coordinates influencing the dynamics. Different types of non-Markovian noise are shown by the goodness of fit tests to be statistically significant on vastly different time scales. On intermediate scales, a Markovian model is not rejected by the tests; models calibrated at these intermediate scales demonstrate a predictive capability for some physical quantities. However, in this intermediate regime, ergodic sampling does not occur over the length of a time series, but a *local* diffusion coefficient is deemed statistically acceptable for the observed raw data. It is demonstrated that a linear mixed effects model can be used to summarize the variation induced by slow unresolved degrees of freedom acting as a non-Markovian noise source. The utility of quantitative criteria for assessing low-dimensional stochastic models calibrated from time series generated by high-dimensional biomolecular systems is briefly discussed. Less coarse-grained data summaries of this type show promise for better understanding the kinetic signature of unresolved degrees of freedom in time series coming from simulations and single-molecule experiments.

1. Introduction

Computations of the effective diffusion coefficient associated with a given order parameter are of interest for various reasons in complex biological systems.[1−6] For example, if the assumptions required by transition state type theories are met, then this information can be used to estimate mean first passage times. Another use of the effective (local or global) diffusion coefficient is in summarizing the statistics of unresolved forces in simulations and experiments.[7] The ability to quantify unresolved forces is particularly relevant to single-molecule experiments where interesting events occur on time scales below the temporal resolution of the measurement device.[7−11] However, in both experiments and simulations, several factors complicate unambiguous estimation of the diffusion coefficient, e.g., inertial effects,[12] measurement apparatus noise,[13,15] and nonergodic sampling of phase space.[14,16,17] Artifacts of these types of factors are sometimes reflected in a dependence of the estimated diffusion coefficient on the spacing between adjacent observations.[3,12]

In this study, simulations of a potassium ion diffusing in the binding pocket of a narrow ion channel, gramicidin A (gA), are analyzed. The ion is allowed to fluctuate in the primary binding pocket of the channel with and without external forces influencing the dynamics. Stochastic differential equation (SDE) models are fit to time series coming from these simulations using time domain methods.[18] The gA system is well-studied[1,4,17,19−22] and is of interest due

* E-mail: Chris.Calderon@numerica.us.

‡ Present Address: Numerica Corporation, Loveland, Colorado 80538, United States.

10.1021/ct1004966
Published on Web 01/07/2011

to the fact that a "memory kernel " is measurable on O(fs−ps) time scales. However, solvation effects, channel undulations, and other phenomena occurring at a broad range of time scales complicate estimating a single global diffusion coefficient from O(ns) times series.[17,18,22] Particular attention is paid to the dependence of the global diffusion coefficient on the time between adjacent observations (this time is known as the "subsampling" or "downsampling" parameter[23−25]) and on the dependence of the local diffusion coefficient on initial conditions. In cases where the latter effect is found to be statistically significant, mixed effects models[26,27] are used to provide a less coarse-grained description of the data. In all cases, goodness of fit tests[12,28] are used to assess the suitability of using a 1-D SDE to describe data arising from a high dimensional complex system. Beyond demonstrating methods that provide quantitive summaries of noisy trajectories, it is also shown how the goodness of fit tests can be utilized to help in determining which models will have predictive power for quantities of interest (such as the sum of squared displacements vs time). The techniques shown are also applicable to experimental time series where "thermal" and instrument noise exist.[13,15] The basic motivation is to efficiently infer information from experimentally accessible quantities (like force and position time series) generated by a complex system where many other degrees of freedom are not directly resolved but their influence may be detected indirectly by kinetic signatures contained in the data.[12−14,17,29]

The remainder of the article is organized as follows: Section 2 reviews the SDE model, summarizes the salient features of the statistical tools used, and provides the molecular dynamics (MD) simulation details. Section 3 presents the results and discussion, and section 4 concludes the article. Supporting Information containing additional mathematical details, a descriptive outline of the fitting procedure, and additional plots are available online.

2. Background and Methods

Computing the asymptotic slope of the mean square displacement of a freely diffusing tagged particle in a homogeneous medium plotted against time is one classic approach to defining the diffusion coefficient. An equation summarizing this idea reads

$$D \equiv \lim_{\Delta t \to \infty} \frac{\langle (z(\Delta t) - z(0))^2 \rangle}{2\Delta t} \tag{1}$$

where D denotes the "classic" global diffusion coefficient, z represents the order parameter (here, the position of the molecule evolving in 1-D), angled brackets denote an ensemble average, and Δt is the time elapsed since the initial observation $z(0)$. Ergodic sampling is usually explicitly or implicitly assumed.[24,30] There are several complications associated with applying this approach to time series coming from biomolecules; e.g., z is often confined by a nontrivial potential, there are unresolved degrees of freedom which make the dynamics "non-Markovian", the medium is not homogeneous, ergodic sampling is difficult to ensure, etc. Several approaches in the physical sciences have attempted to deal with some of these complications.[3,4,30,31] For example, under the assumption of stationary ergodic sampling of phase space, one can utilize the autocorrelation function along with an estimate of the "instantaneous variance" of the observable being monitored to obtain an estimate of the global diffusion coefficient.[3,30,31] However, rigorous unambiguous statistical methods for testing the potential sources of model misspecification given time series data and an assumed continuous time stochastic model are often not employed;[32] this issue will be expanded on in section 3.

An alternative approach to quantifying the fluctuations and computing the "local diffusion coefficient" is to use likelihood-based techniques. For simplicity, an Ornstein−Uhlenbeck SDE is considered in this article as a surrogate for the dynamics. In statistical physics, this SDE is often written as

$$\begin{aligned} \frac{dz}{dt} &= \kappa(\alpha - z) + \eta_t \\ \langle \eta_t \eta_s \rangle &= \delta(t - s) 2\tilde{D} \end{aligned} \tag{2}$$

where here the *local* diffusion coefficient is denoted by $\tilde{D}$. The tilde is used to emphasize that this is a local diffusion coefficient associated with a given SDE. η_t represents the value taken by a mean zero Gaussian process drawn at time t, and $\delta(...)$ is the Dirac δ function, which is meant to suggest that the "random force" increments are statistically uncorrelated. The parameters α and κ can be interpreted as the process mean and effective spring constant, respectively. Defining $\tilde{D}$ *does not* necessarily require one to appeal to ensemble quantities (such as a stationary autocorrelation function[3,4]) of the system observable(s), as is often the case with D. The value of $\tilde{D}$ can be estimated along individual trajectories which may not "ergodically" explore phase space.[12,14,17,33,34] However, physically interpreting $\tilde{D}$ can require care even if the model is judged statistically acceptable. $\tilde{D}$ can potentially contain signatures of unresolved degrees of freedom.[14,17] Several methods for estimating $\tilde{D}$ and quantitatively assessing various assumptions behind a proposed 1-D SDE calibrated from time series of more complex processes (e.g., in our case, the data are generated by a high-dimensional MD simulation) have recently appeared in the mathematical statistics and stochastic processes communities; some examples can be found in refs 18, 23, 35, and 36. In this body of literature, the SDE in eq 2 is often denoted by the following:

$$dz_t = \kappa(\alpha - z_t)dt + \sqrt{2\tilde{D}}dB_t \tag{3}$$

where B_t represents the standard Brownian motion process,[37] the subscript denotes the time index, and $\theta \equiv (\alpha,\kappa,\tilde{D})$ is a vector denoting the parameters needed to specify the process. [All stochastic integrals and SDEs used for modeling are of the Itô type.] For a given discretely observed time series, $\{z_i\}_{i=0}^{N}$, the maximum likelihood estimate (MLE) of the parameter, denoted by $\hat{\theta}$, can be found explicitly for the SDE in eq 3.[18] A guideline outlining some basic recommendations for fitting more general SDEs to trajectories can be found in the Supporting Information. A major advantage of utilizing modern SDE inference tools[18,23,35−37] is that unambiguous

statistical quantities can be computed and various assumptions behind a proposed surrogate SDE (possibly more involved than eq 3) evolution equation can be tested given data arising from a high-dimensional biomolecular system.[12−28] To illustrate the relevance of such statistical inference tools, consider the following: In the narrow gramicidin A channel studied in this article, it is known that inertial memory can complicate using a simple SDE like that given in eq 3 for accurately approximating/summarizing the dynamics of how an ion diffuses along the axis of the channel. If data are sampled every femtosecond, the complex statistical (temporal) dependence in the time series $\{z_i\}_{i=0}^{N}$ would not permit an SDE driven by a standard Brownian motion to approximate the dynamics. In an attempt to "average out" short time non-Markovian noise and attempt to estimate an SDE a statistically acceptable proxy, one can introduce a parameter, n, which subsamples (a.k.a. downsamples) observations.[23−25,33] For example, one can use the series $\{z_{i\times n}\}_{i=0}^{N\times n}$ to obtain $\hat{\theta}^{(n)}$, where the superscript stresses the subsampling parameter. As n increases, the influence of inertia and other fast scale motion decreases, and a process driven by Brownian motion becomes intuitively more plausible.[12]

One set of results in this article focuses on varying n and using the data coming from a high-dimensional biomolecular simulation to determine the goodness of fit of a simple SDE model. For frequently sampled data (corresponding to low n), the results are as expected; in the case of "coarsely" sampled data (corresponding to high n), slow scale unresolved motions will be shown to complicate the use of an SDE of the form given in eq 3 (or 2) to approximate statistics of the underlying complex system. Note that $N \times n$ is selected to be relatively small compared to the total time series size generated by the MD simulation. Hypothesis testing machinery, with adequate power in small samples, is useful for quantitatively determining when a simple Markovian SDE governing the dynamics is statistically acceptable given data coming from a more complex system.

The primary mathematical equations utilized in the statistical analysis are deferred to the Supporting Information; however, the basic idea behind the goodness of fit test statistics is sketched here. The time series arising from the system of interest, $\{z_{i\times n}\}_{i=0}^{N\times n}$, likely possess nontrivial temporal dependencies.[14,17,33,34] Carrying out goodness of fit tests that reliably check for temporal dependencies not implied by the assumed surrogate model class can be problematic.[28,38,39] However, if the data generating process is posited explicitly, it is possible to introduce a transformation utilizing information about all moments assumed by the proposed model which maps a correlated, stationary or nonstationary, time series to a new series of random variables, $\{Z_{i\times n}\}_{i=0}^{N\times n}$ (the transformed series is denoted by a capital letter). Under correct model specification, the Z_i's are independent and identically distributed (iid) random variables with a uniform $U[0,1]$ distribution regardless of the dependence structure.[35] The transformation with these properties does not require asymptotic arguments, and hypothesis tests can be established which simultaneously check if the transformed Z_i's are iid and have the $U[0,1]$ shape. [However, it is emphasized that the transformation requires that the data generating process be exactly known for the precise results to hold; if a parameter(s) needs to be estimated from data, then some technical complications are encountered.[35,38]]. Deviance from either condition suggests the surrogate SDE is not faithful to the data. Hong and Li proposed the so-called "omnibus" Q test statistic (relevant equations reported in Supporting Information) which jointly checks both the iid and $U[0,1]$ shape assumption. Such "omnibus" tests can sacrifice some power,[40] but tests which focus more on the independence assumption (and loosely "focus" on non-Markovian errors) can be employed. The M test in ref 35 is one such test. It does not check for the $U[0,1]$ shape but instead focuses on autocorrelations in moments of the Z_i's. The utility of both test statistics in analyzing time series possessing noise coming from many time scales will be presented.

It should be stressed that increasing n does not necessarily guarantee that a single SDE of the type given in eq 3 will provide a statistically acceptable model of the stochastic dynamics of an ion in a binding pocket. This will be demonstrated explicitly in the Results and Discussion section. For cases where ergodic sampling does not occur, but a local SDE model is deemed appropriate by some criterion at a given time scale, quantification of the influence of initial conditions on the estimated local diffusion is of interest. To accomplish this, the framework of mixed effects models will be used.[26] The setup is as follows:

$$\begin{gathered}\tilde{D}_{i,j} = \mu^{\tilde{D}} + b_i^{\tilde{D}} + \varepsilon_{i,j} \\ i = 1, 2, ..., N_{\mathrm{IC}} \\ j = 1, 2, ..., N_{\mathrm{Rep}}\end{gathered} \tag{4}$$

The more technical details of the model are deferred to the Supporting Information; the physical motivation for the terms above is described here. $\mu^{\tilde{D}}$ denotes a fixed-effect population mean of the local diffusion coefficients, and $b_i^{\tilde{D}}$ denotes a random-effect specific to initial condition i. N_{IC} represents the number of initial configurations (ICs) analyzed, where an IC is defined by the position of all atoms in a simulation. N_{Rep} denotes the number of repeat experiments for a given fixed IC (the position of all atoms is fixed, but different initial velocities are used), and $\varepsilon_{i,j}$ represents "sampling noise". The significance of the random term (i.e., dependence on initial conditions) is tested using both the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Because ergodic sampling may not occur, the random-effect, i.e., variation due to the initial conditions, may be statistically significant.

2.1. Simulation Details. The NAMD[41] simulation package with parameters used originally in ref 5 and then in ref 17 is used. The temperature was set to 310 K, the pressure was maintained at 1 atm. The only significant difference in the simulations reported here is that the harmonic guiding potential is not used to "steer" the ion in a time-dependent fashion. A configuration where a single potassium ion was located in the binding pocket was used as an initial condition. This initial configuration was equilibrated for 1 ns of simulation time (without external force). After this equilibration, an ensemble of production runs carried out for 6 ns

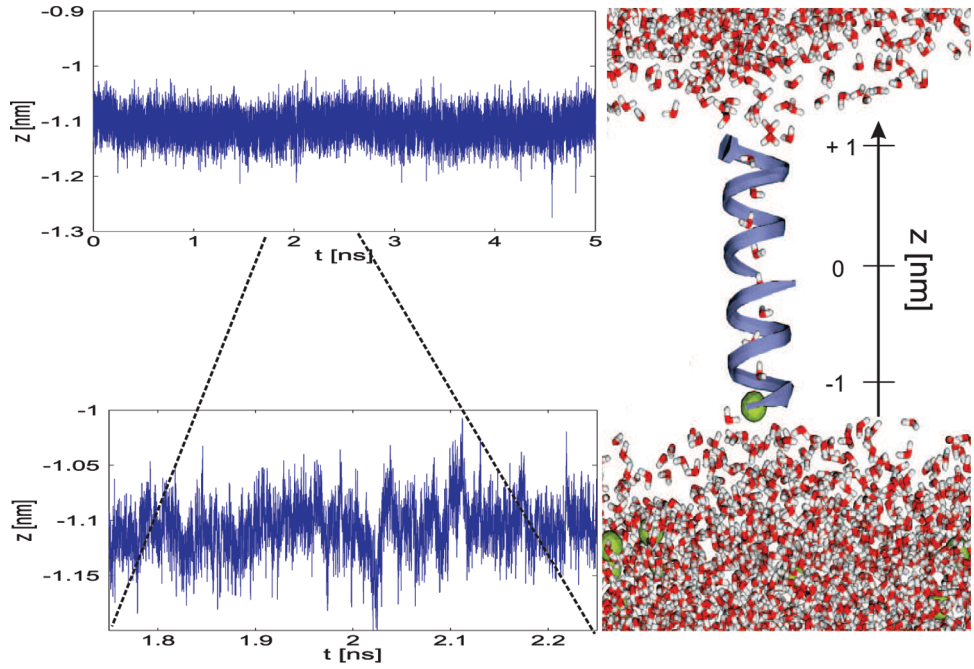

Figure 1. Sample trajectories coming from MD simulation. The bottom panel zooms in on the top panel time series to emphasize that the local mean changes in a nontrivial fashion (i.e., the potential is not a single harmonic well). In this article, we focus on the binding pocket near −1 nm. The system snapshot was generated with VMD.[48]

was used to generate additional "equilibrated" initial conditions (the tagged ion remained in the binding pocket for this time period). Every 10 fs, the position of the ion along the channel was output to disk; this sampling frequency corresponds to the parameter $n = 1$ in the expression $\{z_{i\times n}\}_{i=0}^{N\times n}$. The constrained runs used the same initial conditions as the unconstrained runs.

3. Results and Discussion

A representative trajectory obtained while monitoring the ion's position along the axis channel is plotted in the top left panel of Figure 1; the bottom left panel zooms in on a segment to show the fine temporal structure. The illustration to the right is a snapshot of the channel (lipid molecules omitted from the plot for clarity). The center of the dimer channel defines the zero of the z coordinate. The binding pocket near $z \approx -1$ nm is relatively shallow (it is computed to be $\approx 5k_BT$) but deep enough to allow the ion to be trapped for a substantial amount of MD simulation time. The 1-D potential of mean force (see Supporting Information Figure 1) used to describe z is clearly not a perfect harmonic potential nor does the PMF capture all of the information needed to understand the rich dynamics,[17,20,22] but it does describe the average location of a trapped ion fairly well over a 1−9 ns window. Note that the primary interest throughout is in the unconstrained case, but simulations utilizing a harmonic guiding potential are also studied to demonstrate that the findings are not solely an artifact of an anharmonic potential.

Figure 2 plots $\tilde{D}$ as a function of n for time series batches each consisting of $N = 100$ observations (in this plot, a total of 6×10^5 time series entries were analyzed). If n increases, while at the same time the number of (uniformly sampled) time series observations, N, is fixed, this clearly implies that

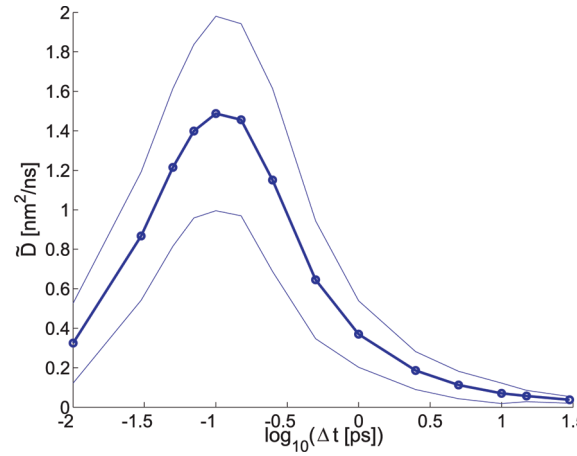

Figure 2. Estimated local diffusion coefficient for different "downsampling or subsampling" intervals. The average MLE computed from a time series appears as a symbol for a given Δt value. The solid line represents the sample mean ±2 times the sample standard. It should be noted that every observation was utilized, similar to the approach used in ref 23, but correlation in the time series does make these confidence intervals suspect. However, the trends observed in the MLE confidence bands computed are similar to those expected asymptotically if the underlying SDE was the data generating process.[18]

the time series batches constructed using a larger n are associated with a larger time between observations (and also a larger final time horizon). The logarithm of the time between adjacent entries of the (uniformly spaced) time series, denoted by $\Delta t \equiv n\delta t$, serves as the x axis in Figure 2. δt corresponds to the spacing associated with $n = 1$ ($\delta t = 10$ fs throughout). The local effective diffusion coefficient, $\tilde{D}$, estimated in this fashion displays a nontrivial trend with the "coarse-graining" parameter Δt. Dependence of $\tilde{D}$ on Δt

can either be a sign of "memory" due to inertial effects[4] and/or a sign of a "poor reaction coordinate".[2] Note that for larger Δt, $\tilde{D}$ appears to level off to a value of $\approx 3.0 \pm 1.9$ Å^2/ns ($\approx 0.030 \pm 0.019$ nm^2/ns); this "convergence" may seem to suggest that diffusive motion is an adequate approximation of the dynamics on this coarser time scale. It is interesting to also observe that the *global* diffusion coefficient estimated using a method[3] appealing to an integration of the empirically measured autocorrelation (obtained using a 9 ns MD trajectory sampled every 10 fs) and an empirical estimate of the variance is 2.7 Å^2/ns. This is in agreement with the range predicted by the MLE estimate of $\tilde{D}$ obtained using a simple Ornstein−Uhlenbeck SDE. Autocorrelation and/or memory is used traditionally in molecular dynamics.[3,4,32] Such methods usually implicitly assume that various moments are stationary and adequately sampled.

The apparent convergence of $\tilde{D}$ at larger Δt and the consistency of the local diffusion coefficient and the global diffusion coefficient (estimated using vastly different methods) might lead one to conclude that this diffusion coefficient is a reasonable summary of the dynamics which can be used for predictive purposes (such as computing the sum of squares of increments or a mean first passage time). The strong dependence of $\tilde{D}$ on the Δt for Δt in the $\approx 0.05-0.20$ ps range would also seem to suggest that this diffusion coefficient is physically meaningless (or at least nontrivial to interpret in terms of classical statistical mechanics).

However, the results shown in Figure 3 provide results suggesting that the above intuition is misleading. Here, results obtained by computing the "Q" and "M" goodness of fit tests reported in ref 35 (the relevant equations are reproduced in the Supporting Information) are plotted using the MLE obtained using observational data and the assumed SDE model hoping to approximate the high dimensional process generating the time series. If the model is correct, then it can be shown that both statistics asymptotically converge to mean zero standard normals.[35] Some simple finite sample size correction to the test statistic distribution[28] can be made (one is discussed in the caption of Figure 3); more sophisticated approaches are discussed in ref 38. Models inconsistent with the observed data (i.e., model mis-specification) will result in large values of the test statistic if there is enough statistical evidence of model inadequacy. Recall that the omnibus Q test aims to simultaneously check that the increments of the discretely observed time series follow the expected distribution shape *and* have the correlation structure consistent with the assumed model. In the surrogate SDEs considered in this paper, the dynamics of z need to be approximately Markovian for the assumed proxy to be statistically acceptable. Somewhat surprisingly, the most plausible Markovian SDE (as judged by *both* the M and Q) test statistic occurs at an intermediate Δt. Various items related to this observation are explored in the results that follow.

Figure 4 plots the empirically determined autocorrelation (AC) function of the force and position taken from three different MD simulations each spanning 3 ns. The force AC demonstrates an oscillatory behavior that decays after a fairly

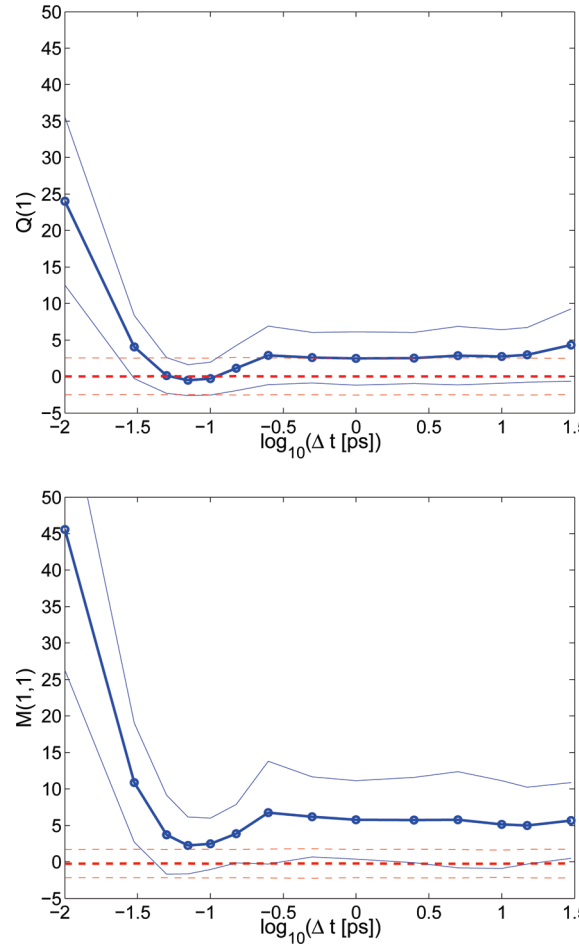

Figure 3. Goodness of fit tests. Similar to the previous figure except the corresponding M and Q statistics are plotted. The dotted lines denote the sample mean ±2 times the sample standard deviation of the test statistic one can expect under a correctly specified model. The dotted lines were generated using an iid $U[0,1]$ random variable sequence possessing the same size as the time series and evaluating Q and M with this reference sequence; the iid $U[0,1]$ reference sequence used the same nonparametric estimators as the MD data and the corresponding computed generalized residuals.

short MD timespan (≈ 0.5 ps). This nontrivial, but quickly decaying, AC in the force is the motivation for introducing a "memory kernel" in studies analyzing this channel, e.g., ref 4. It is worth noting that in each of the ACs there is a point near 0.1 ps corresponding to zero temporal correlation. This point also corresponds to the Δt (or n) where a Markovian SDE provides the best fit in regards to the goodness of fit test statistics studied. Loosely speaking, the assumptions that inertia can be ignored and the net effect of unresolved degrees of freedom can be modeled as a mean zero "random bath force" which can be approximated by a Brownian motion process become most plausible in this regime. This also might explain why the Q test has less power than the M test in this situation: the former focuses on the shape and statistical dependence between the generalized residuals computed from z_i and $z_{i\times n}$, whereas the M statistic focuses on the full AC of moments of the generalized residuals computed from the observed data and the assumed model. If inertia and other unresolved forces were truly

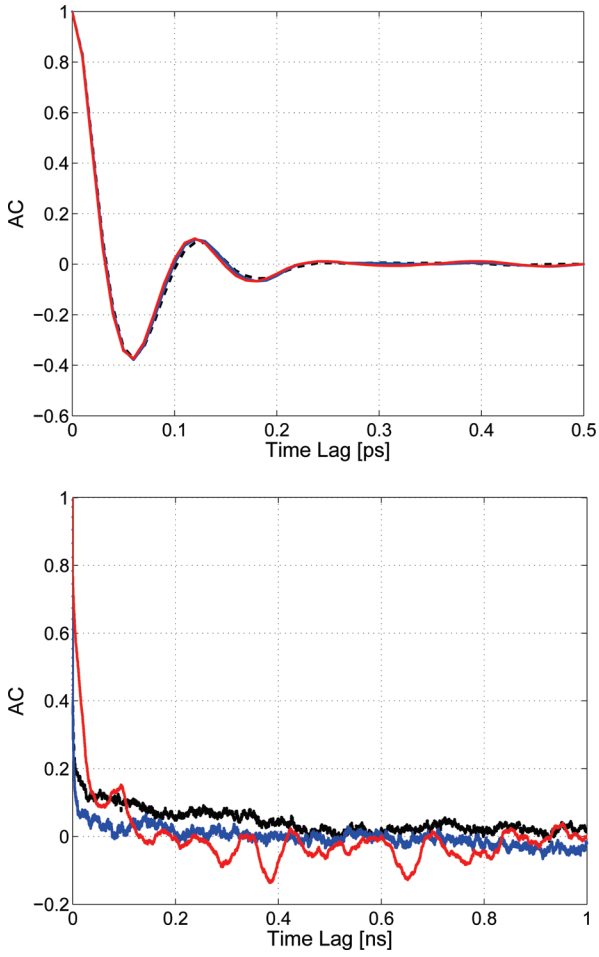

Figure 4. The measured force (top) and position (bottom) autocorrelation (AC) functions. Three different trajectories (one AC corresponding to each) were used to compute the various ACs. Note that the time lag units are different on each *x* axis.

unimportant and a first order Markovian SDE generated the observed noisy time series $\{z_{i\times n}\}_{i=0}^{N\times n}$, then the random force would need to have *all* of the statistical properties of a Brownian motion, namely, iid mean zero increments (not just uncorrelated increments). In this study, we know *a priori* that the z_t's are generated by a dynamical system where many degrees of freedom are not observed. The force ACs all die off relatively quickly and also give rise to very similar ACs, suggesting that the temporal correlation in the force is similar in each case analyzed. However, the channel has other unresolved slow degrees of freedom. Note how in Figure 1 the mean level changes after after 10−50 ps. The ion channel is flexible;[17] undulations of the protein and interactions of the tracked ion with the water chain and other ions in this narrow channel give rise to a more complex noise source. Artifacts of the non-Markovian (in 1-D) slow scale motion are reflected in the position autocorrelation. The three trajectories analyzed gave similar force ACs but very different ACs for the position. The shape of the position AC explains why a simple diffusive SDE is statistically rejected even at fairly large Δt values. Note that subdiffusive processes, e.g., fractional Brownian motion,[42] have been intentionally not considered as surrogates. [The fine structure apparent in the representative trajectory in Figure 1 suggests that a mathematically tractable subdiffusive process would not likely be able pass trajectorywise a goodness of fit test making full use of the implied conditional density of the assumed subdiffusive model. This author prefers the use of physically interpretable Markovian SDEs in part because measurement noise and other relevant physical features can be readily accounted for and powerful hypothesis testing machinery exists in this setting; recall that these tools check both the conditional distribution shape and assumed temporal correlations. If one can develop reliable tests checking various assumptions behind a particular non-Markovian surrogate and/or can demonstrate that such approaches make new useful physical predictions in a given system, these models should certainly be considered as potential surrogates, but model class selection is not the focus of this article.]

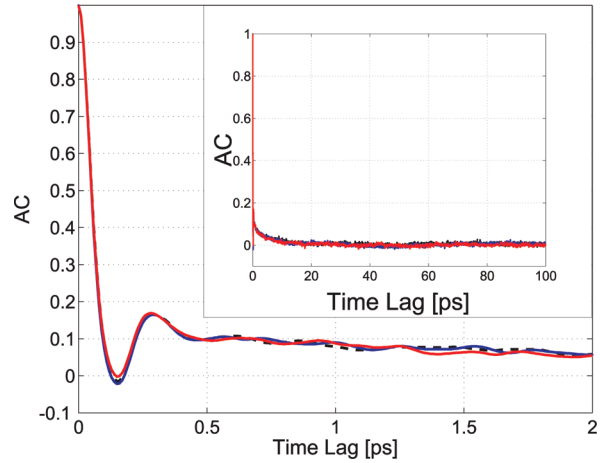

Figure 5. The position AC measured from three different 3 ns runs carried out in the presence of a constraining potential (see text). The inset shows the long time behavior, and the main portion of the figure zooms in only on early times.

One might inspect the position ACs in the unconstrained case and argue that the diffusion constant should be obtained using biased simulations where a harmonic guiding potential is employed in an attempt to constrain the dynamics close to a point of interest.[3,4] A physical motivation for this approach is to make the system's effective drift more closely resemble that associated with a 1-D harmonic potential and focus attention on the resulting effective diffusion.[4] Results using $k_{\text{harmonic}} = 40$ kcal/mol/Å^2 and adding the biasing potential $U_{\text{bias}}(z) = k_{\text{harmonic}}/2(z + 11\ \text{Å})^2$ to the MD evolution equations are reported to demonstrate that many of the previous complications observed before do not go away. The constrained results are qualitatively similar to those observed in the unconstrained case (see the Supporting Information); the one notable difference is in the position ACs. When three separate 3 ns constrained runs are analyzed, the resulting ACs obtained demonstrate "better" ergodic sampling in the sense that now the position ACs overlap substantially (see Figure 5). The global diffusion coefficient estimated using the method in ref 3 (using a 9 ns trajectory) was found to be 13.88 Å^2/ns and that obtained using the OU model was $\tilde{D} = 10.53 \pm 3.64$. Both diffusion coefficient values reported here are in close agreement with those reported in refs 4 and 17, computed using different computational methods also utilizing constrained simulations. Reference 4 used a memory kernel, whereas the values estimated here employed a

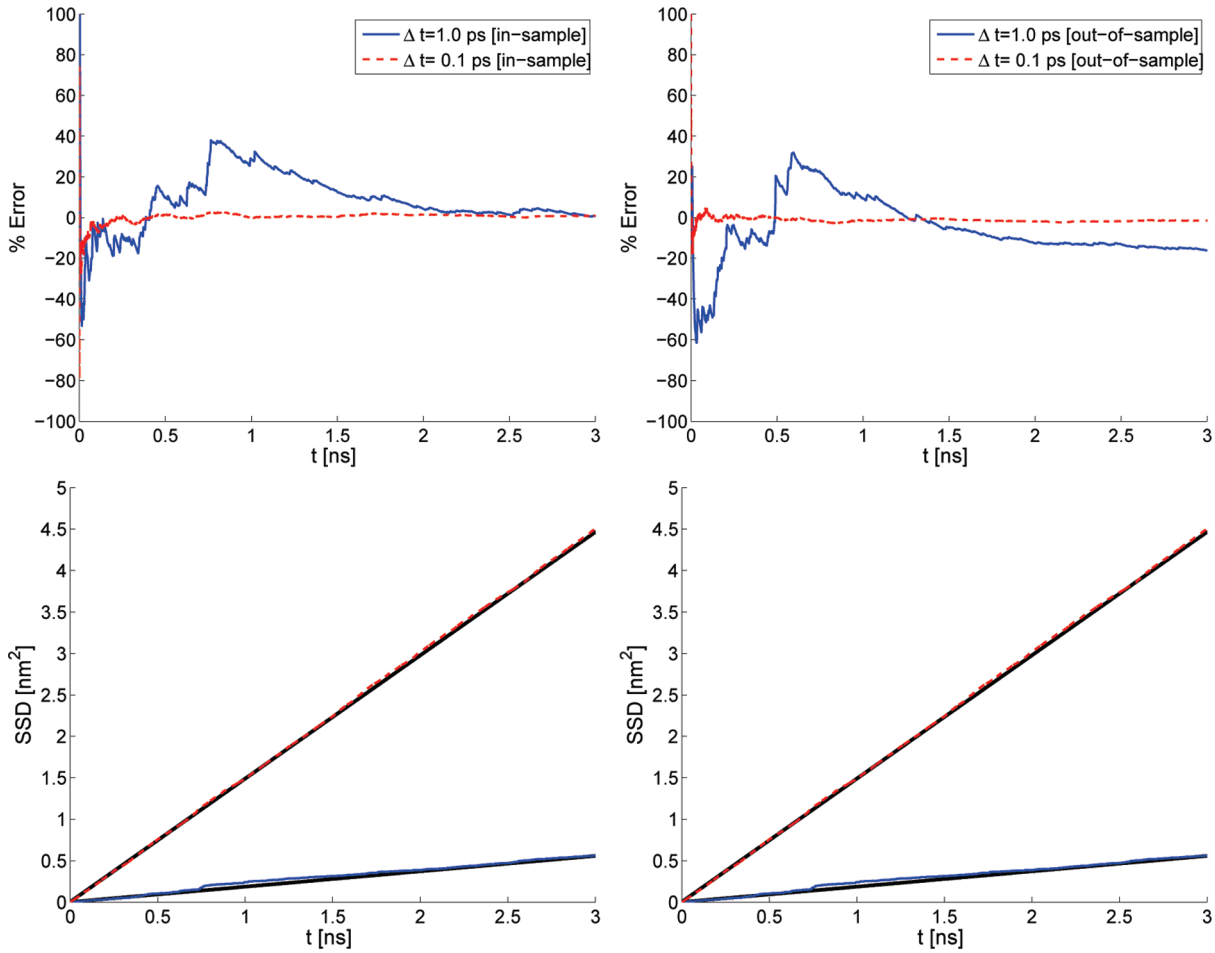

Figure 6. The percent error in the sum of squared displacements (top) and the raw observed and predicted sum of squared displacements (bottom). "In-sample" and "out-of-sample" data were analyzed (see text). The percent error was computed using $(\mathrm{SSD}^{\mathrm{observed}}(t) - \mathrm{SSD}^{\mathrm{predicted}}(t))/\mathrm{SSD}^{\mathrm{predicted}}(t) \times 100$.

Markovian SDE with downsampling (using $n = 100$) in one approach. The other computational method[3] reported in this paragraph utilized information extracted from the "converged" position AC.

Although these formal methods provide diffusion coefficients which are in agreement with one another, this does not imply that the computed quantity is consistent with observed trajectories. The Supporting Information shows that even for fairly large n or Δt even Markovian SDEs with the "correct" diffusion coefficient are still rejected using the M and Q tests in the constrained case. The agreement between formal methods for computing the effective diffusion coefficient does not imply predictive ability either. Furthermore, one is usually interested in computing the diffusion coefficient of the unconstrained system and using this result for predicting various physical quantities; so the artifacts observed in the unconstrained simulations should be dealt with in a surrogate model. That is, nontrivial fluctuations (or their damping rate) can make important contributions to forecasted events. One theme advocated in this article is that formal methods for computing the diffusion coefficient should be consistent with observation in some quantifiable sense and/or be able to make predictions (outside of the fitting criteria) of events occurring over time scales of interest. If neither criteria can be met, one should consider "non-traditional" approaches to computing the diffusion coefficient. For example, recent studies[43,44] demonstrated that the effective friction (which was related explicitly to the diffusion through the fluctuation dissipation theorem) of a coarse-grained model calibrated from observational data coming from high dimensional steered molecular dynamics simulations of a gramicidin channel could be used to predict mean first passage times under zero external force; the success of this approach relied on inferring the effective friction and force from a physically motivated fitting criterion different than formal procedures typically used in classic chemical physics computations.[44] Consequences of the fact that intermediate Δt's (in the 0.1−0.2 ps range) yield the best SDE proxy in the unconstrained case in the gA system studied here, in regards to both goodness of fit to the observed data and the predictive ability of statistics of the complex systems, for the regimes studied is the focus of the next two set of results. These findings also demonstrate that classic formal definitions of the diffusion coefficient should be reconsidered in low dimensional models or summaries of complex dynamical systems.

In order to demonstrate that simple SDE models calibrated at intermediate Δt values possess some predictive ability, Figure 6 plots the sum of squared displacements, $\mathrm{SSD}(t;n)$: $= 1/2\sum_{i=0}^{N_{\mathrm{sim}}(t)} (z_{(i+1)\times n} - z_{i\times n})^2$, coming from the MD simula-

tions and the straight line predicted by using the estimated $\tilde{D}$; $N_{sim}(t)$ corresponds to the number of time ordered observations generated by the MD simulation up to time t. Two regimes are studied, the intermediate regime providing the best fit, as judged by the Q and M statistics, and the larger $\Delta t \approx 1.0$ ps where the observed effective drift is low but the Q and M suggest something is askew. The complex unresolved slow-scale motion prevents a diffusion model from being statistically acceptable even for fairly large n (equivalently Δt). Said differently, the structured exploration of phase space (see the bottom panel of Figure 1) cannot be approximated by a process driven by Brownian motion. In order to demonstrate how these artifacts influence $\tilde{D}$'s ability to predict SSD, a 3 ns trajectory of simulation data was used to calibrate the parameters of the Ornstein−Uhlenbeck model. These calibration data are labeled as "in-sample", and another 3 ns of data (not used for parameter estimation) were labeled "out-of-sample". The Δt corresponding to the the lowest goodness of fit test statistics also provides the best predictive model. It is worth noting that the "best" prediction is judged in terms of percent error in the empirically in and out of sample SSD. In a parametric MLE estimate, drift and diffusion are both explicitly accounted for by the likelihood function (and also by the goodness of fit tests used), but in the SSD, the influence of the drift can adversely affect the SSD for larger Δt. However, the SSD plots suggest that these effects are not too dramatic. The rejection of the Ornstein−Uhlenbeck SDE calibrated using $n = 10$ with a moderately small sample size ($N = 100$) may be due to subjecting the model to an overly stringent hypothesis test; i.e., the errors in the SSD may be acceptable for a practical approximation in the physical sciences. However, it is nonetheless useful to know that the errors observed when using a diffusion approximation to describe increments of a more complex process are systematic and not simply sampling errors. Admittedly, predicting the SSD associated with the Δt yielding the best classic diffusion approximation may not be of interest in chemical applications per se. However, knowledge of the time scale where random forces can be approximated by a diffusion type processes has proven useful in making predictions relevant to nonequilibrium potential of mean force computations.[17]

It should be explicitly pointed out that the statistician's tenet of "thou shalt not waste data" was adhered to; i.e., even though subsampling occurred, each observation was eventually used for parameter estimation, e.g., see ref 23. In Figure 6, the slope of the line was predicted using the population average parameter observed using every observation in a single trajectory. The physical intuition behind using the population average implicitly assumes ergodic sampling (i.e., time averages are close to ensemble averages[16,45]). However, given that larger n eventually results in a rejected model, it would be interesting to see if there is enough statistical evidence to suggest that the estimated $\tilde{D}$ depends significantly on the full set of initial conditions for intermediate n. In in attempt to quantify this effect, sometimes referred to as "dynamic disorder",[45] the mixed effect model given in eq 4 was fit to the inferred local diffusion coefficient data. More specifically, an attempt was made to quantify if the variability induced by different initial conditions (drawn from a Boltzmann distribution) can be detected in the presence of sampling uncertainty. $N_{IC} = 20$ common position initial conditions were taken from the equilibrated initial condition, and the position coordinates were recorded every 100 ps in unconstrained simulations. From these multiple IC position files, $N_{Rep} = 10$ different random number streams and velocities' ICs were used to generate N_{Rep} short MD trajectories (i.e., a total of $N_{Rep} \times N_{IC} = 200$ trajectories of size $N = 100$ were analyzed). The subsampling parameter used here was $n = 10$ (corresponding to the "best" models as determined by the goodness of fit analysis presented earlier), and observations were recorded for each IC. The previous set of results demonstrated that quantities depending only on intermediate spacing between adjacent time series entries (recall that this spacing was quantified by $\Delta t = n\delta t$) had predictive ability. For these sampling parameters, there was sufficient evidence indicating the statistical significance of the random effect. The p value obtained when testing a mixed effect model versus a pure fixed effect reference model ($b_i^{\tilde{D}}$ was forced to be zero) was 0.043 (suggesting that the fixed effect model was suspect). The AIC and BIC also suggested that the random effect in the mixed effect model was statistically significant. In terms of statistical mechanics, this translates into the statement that variation induced by the different ICs is statistically significant; the distribution observed in the estimated local diffusion coefficients cannot be attributed to sampling uncertainty alone. A more intuitive demonstration is presented in the box plots of Figure 7 where $N_{IC} = 20$ different box plots are displayed. This lack of ergodic sampling might cause one to claim that this is "the sign of a bad reaction coordinate." However, it is important to attempt to make full use of time series information available; e.g., in single-molecule experiments, the observables available will likely be "imperfect" reaction coordinates.

Only fairly simple Ornstein−Uhlenbeck models were considered in the "practically stationary" regime in this article. This regime was studied mainly to facilitate the statistical analysis and demonstrate that nontrivial features can be detected even in this regime. The transition density, MLE, and associated limiting distribution can be computed in this case for the OU model without having to appeal to additional numerical approximations.[18] However, the tools presented can handle the addition of other features, such as time dependent external forces[13,15,17,29,33,34] and/or position dependent diffusion in a continuous time nonlinear SDE model.[12,16] The basic findings reported here come through even on SDE models with additional features. For example, box plots obtained by estimating the parameters of a continuous (time and state space) SDE taking position dependent "overdamped" noise into account[12] are shown to demonstrate that the mixed model analysis did not contain artifacts of missing position dependence in the local diffusion coefficient (the p value in the corresponding mixed effects analysis was 0.0033, and both the AIC and BIC favored the random effects model).

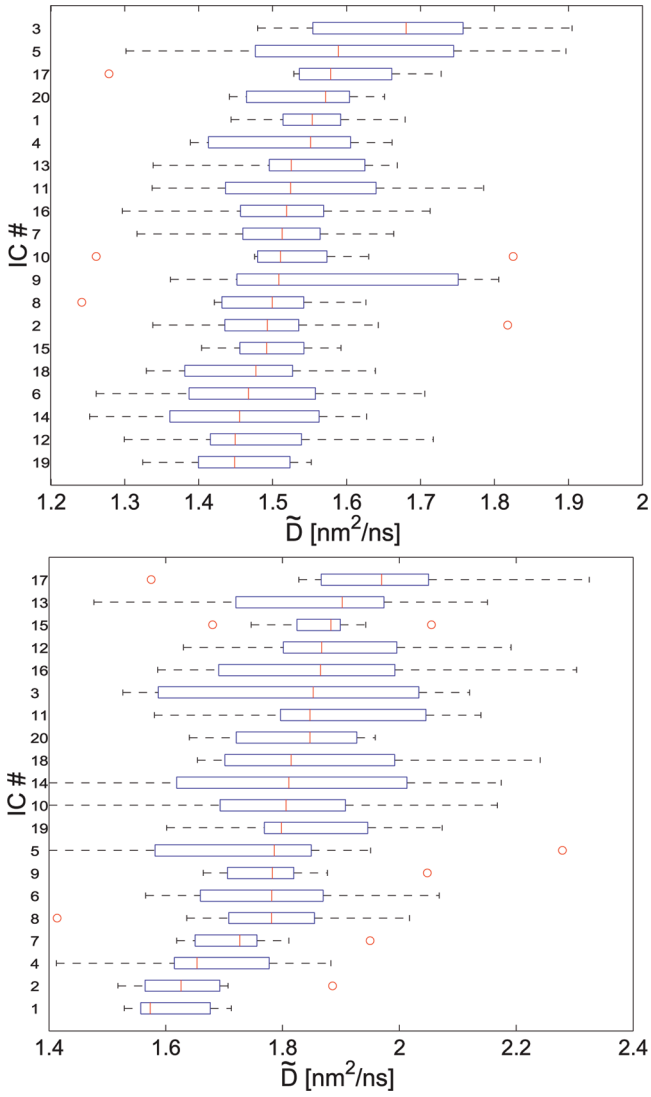

Figure 7. Box plots of $\tilde{D}$. Each box plot corresponds to a different initial condition drawn from an equilibrated MD simulation. The top plot corresponds to the Ornstein Uhlenbeck estimate and the bottom to a position dependent "overdamped" SDE (see text for additional details).

4. Conclusions and Outlook

Time domain maximum likelihood (transition density based) inference methods were demonstrated to be useful for fitting and assessing the statistical validity of a 1-D SDE model approximating the dynamics associated with gramicidin A simulations. Unresolved degrees of freedom were shown to be detectable on a "fast time scale" (time series observations were uniformly separated by time intervals < 0.1ps) where force correlations were found to be statistically significant[4] and also at "slow time scales" studied (time between adjacent observations ranging from 0.8 to 60 ps). At intermediate time scales, a collection of *local* SDEs was shown (1) to be a better statistical summary of the data as measured by goodness of fit tests which made use of the entire assumed conditional distribution and time correlations, (2) to contain enough statistical evidence to indicate that nonergodic sampling was occurring (the mixed effects model approach was also shown useful in quantifying the degree of the initial condition dependence), and (3) to have predictive capability for out-of-sample data; it should be noted that the predicted quantity was not used as a fitting criterion.

It is not too surprising that solely monitoring the axial location of a tagged ion is problematic in defining a 1-D diffusion coefficient in an ion channel where channel undulations, nontrivial solvation effects, and other unresolved factors modulate the dynamics.[2,17,20,45] These factors can significantly complicate estimation of the system's position autocorrelation (as shown here and in ref 16) and prediction of more complex long-term events (such as mean first passage times). However, assessing the validity of various simplifying assumptions, such as the suitability of an assumed Markovian model on a specified time scale[35,36] given time ordered observations, will assist in better understanding/summarizing the rich information coming from all-atom computer simulations and single-molecule experiments.[12,14,28] The utility of frequentist inferential methods employing transition density information in analyzing the effective diffusive noise as opposed to fitting an autocorrelation function or only focusing on some other low order stationary moments was also demonstrated (aspects of this issue are discussed more extensively in ref 32). In regard to some traditional chemical physics computations, such as mean first passage time computations,[3,6,46] often computations of both the effective diffusion coefficient and free energy differences are required. In such cases, it is possible that systematic errors in both the classic diffusion coefficient and free energy estimates can cancel to provide the same mean first passage time ("rate") prediction. It is useful to have reliable criteria for checking various model assumptions with hypothesis testing machinery.[28] The methods presented here can be used to determine if the implicit assumptions behind a given coarse system description[3,6,46] are appropriate given observational data. Careful statistical analysis of the diffusion coefficient can potentially help in assessing the accuracy of estimated free energy or PMF differences (indirectly) if one is only given experimental flux measurements. Such analyses can potentially identify factors that may be confounded in traditional mean first passage time analyses. If formal definitions for the diffusion coefficient are not consistent with data and/or unable to make useful predictions, one should consider alternate approaches for quantifying "thermal noise", as demonstrated here and in refs 44 and 45.

With the advent of single-molecule experiments and ever increasing MD simulation power, it is important to consider physically interpretable data summaries that possess predictive ability if we hope to fully utilize the wealth of information coming from these new data sources. The need for models that have testable criteria which can be applied to both experimental and simulation time series poses new and exciting challenges in describing biological systems.[12,14,44,47] For simplicity, the focus here was on a roughly stationary signal [that is, the moments of the time series were roughly time independent[4]] approximated by an SDE with a constant diffusion coefficient, variants of the MLE type of approach are applicable to nonergodic cases where time dependent external forces

are added into the system and the local diffusion coefficient depends on the value of the order parameter.[33,34] Similar approaches have been demonstrated to be useful in understanding experimental data where measurement noise (on top of thermal noise) is also present.[13−15] The magnitude of the thermal and measurement noise can both be fit from observational data (these quantities do not need to be guessed or assumed *a priori*), and the goodness of fit tests can be used to determine if a proposed model is appropriate given the data. For example, one can explicitly test if thermal noise dominates measurement noise without requiring an implicit stationarity assumption.[14] Fourier transform based methods, popular in statistical physics, often require such stationary assumptions, but these can be hard to satisfy in single-molecule data.[13,15] With time domain likelihood based approaches, dynamic signatures of unresolved degrees of freedom have been suggested by the estimated position dependent diffusion coefficient in studies analyzing experimental data where time-dependent forces are added, e.g., see refs 13 and 14. The type of data summary presented here, where the entire distribution implied by an assumed surrogate model is used to assess the fit and a mixed effects model is used to respect the variability induced by a lack of ergodic sampling shows promise in understanding complex data sets arising from future simulations and experiments. Attempts were made to avoid appealing to "memory kernels"[4] or long memory processes in order to facilitate the physical interpretation of the surrogate SDEs and utilize information that is experimentally accessible (e.g., force or position). Regardless of the type of surrogate model used (i.e., one with or without memory), mixed effect modeling techniques show promise as tools for quantitatively summarizing the dynamics observed when unresolved degrees of freedom are believed to be important but not directly measurable.[34]

Acknowledgment. The author 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 Available: The Ornstein−Uhlenbeck model, goodness of fit tests, rough guidelines to more general SDE modeling, and mixed effects models. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Roux, B.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 4856–4868.

(2) Burykin, A.; Kato, M.; Warshel, A. *Proteins* **2003**, *52*, 412–426.

(3) Hummer, G. *New J. Phys.* **2005**, *7*, 34.

(4) Mamonov, A.; Kurnikova, M.; Coalson, R. *Biophys. Chem.* **2006**, *124*, 268–278.

(5) Forney, M. W.; Janosi, L.; Kosztin, I. *Phys. Rev. E* **2008**, *78*, 051913.

(6) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13841–13846.

(7) Moffitt, J.; Chemla, Y.; Smith, S.; Bustamante, C. *Annu. Rev. Biochem.* **2008**, *77*, 19.119.4.

(8) Nollmann, M.; Stone, M. D.; Bryant, Z.; Gore, J.; Crisona, N. J.; Hong, S. C.; Mitelheiser, S.; Maxwell, A.; Bustamante, C.; Cozzarelli, N. R. *Nat. Struct. Mol. Biol.* **2007**, *14*, 264–271.

(9) Walther, K.; Gräter, F.; Dougan, L.; Badilla, C.; Berne, B.; Fernandez, J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7916–21.

(10) Greenleaf, W.; Frieda, K.; Foster, D.; Woodside, M.; Block, S. *Science* **2008**, *319*, 630–633.

(11) Hodges, C.; Bintu, L.; Lubkowska, L.; Kashlev, M.; Bustamante, C. *Science* **2009**, *325*, 626–628.

(12) Calderon, C.; Arora, K. *J. Chem. Theory Comput.* **2009**, *5*, 47.

(13) Calderon, C.; Harris, N.; Kiang, C.; Cox, D. *J. Phys. Chem. B* **2009**, *113*, 138.

(14) Calderon, C.; Chen, W.; Harris, N.; Lin, K.; Kiang, C. *J. Phys.: Condens. Matter* **2009**, *21*, 034114.

(15) Calderon, C.; Harris, N.; Kiang, C.; Cox, D. *J. Mol. Recognit.* **2009**, *22*, 356.

(16) Calderon, C. P. *Phys. Rev. E* **2009**, *80*, 061118.

(17) Calderon, C.; Janosi, L.; Kosztin, I. *J. Chem. Phys.* **2009**, *130*, 144908.

(18) Tang, C.; Chen, S. *J. Econometrics* **2009**, *149*, 65–81.

(19) Allen, T. W.; Bastug, T.; Kuyucak, S.; Chung, S. H. *Biophys. J.* **2003**, *84*, 2159–2168.

(20) Miloshevsky, G. V.; Jordan, P. C. *Biophys. J.* **2004**, *86*, 92–104.

(21) Allen, T. W.; Andersen, O. S.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 117–122.

(22) Braun-Sand, S.; Burykin, A.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **2005**, *109*, 583–592.

(23) Zhang, L.; Mykland, P.; Ait-Sahalia, Y. *J. Am. Stat. Assoc.* **2005**, *100*, 1394–1411.

(24) Pavliotis, G. A.; Stuart, A. M. *J. Stat. Phys.* **2007**, *127*, 741–781.

(25) Calderon, C. *Multiscale Model. Simul.* **2007**, *6*, 656–687.

(26) Pinheiro, J.; Bates, D.; DebRoy, S.; Sarkar, D. *nlme: Linear and Nonlinear Mixed Effects Models*, R package version 3.1−96; R Core team: 2009.

(27) Ruppert, D.; Wand, M.; Carroll, R. *Semiparametric Regression*; Cambridge University Press: New York, 2003; pp 91−110.

(28) Calderon, C. P. *J. Phys. Chem. B* **2010**, *114*, 3242–3253.

(29) Calderon, C.; Chelli, R. *J. Chem. Phys.* **2008**, *128*, 145103.

(30) Burykin, A.; Kato, M.; Warshel, A. *Proteins* **2003**, *52*, 412–426.

(31) Smith, G. R.; Sansom, M. S. *Biophys. Chem.* **1999**, *79*, 129–151.

(32) Pokern, Y.; Stuart, A.; Vanden-Eijnden, E. *Multiscale Model. Simul.* **2009**, *8*, 69–95.

(33) Calderon, C. *J. Chem. Phys.* **2007**, *126*, 084106.

(34) Calderon, C.; Martinez, J.; Carroll, R.; Sorensen, D. *Multiscale Model. Simul.* **2010**, *8*, 1562–1580.

(35) Hong, Y.; Li, H. *Rev. Fin. Studies* **2005**, *18*, 37–84.

(36) At-Sahalia, Y.; Fan, J.; Jiang, J. *Annals of Statistics* **2010**, *38*, 3129−3163.

(37) Protter, P. E. *Stochastic Integration and Differential Equations*, 2nd ed.; Springer: New York, 2003; pp 12−84.

(38) Chen, S.; Gao, J.; Tang, C. *Ann. Stat.* **2008**, *36*, 167–198.

(39) Johnson, V. E. *Ann. Stat.* **2004**, *32*, 2361.

(40) Ait-Sahalia, Y.; Fan, J.; Peng, H. *JASA* **2009**, *104*, 1102–1116.

(41) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(42) Kou, S.; Xie, X. *Phys. Rev. Lett.* **2004**, *93*, 180603.

(43) Kamerlin, S.; Vicatos, S.; Dryga, A.; Warshel, A. *Annu. Rev. Phys. Chem.* **2011**; doi: 10.1146/annurev-physchem-032210-103335.

(44) Dryga, A.; Warshel, A. *J. Phys. Chem. B* **2010**, *114*, 12720−12728.

(45) Kuo, T. L.; Garcia-Manyes, S.; Li, J.; Barel, I.; Lu, H.; Berne, B. J.; Urbakh, M.; Klafter, J.; Fernandez, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 11336–11340.

(46) Pisliakov, A. V.; Cao, J.; Kamerlin, S. C.; Warshel, A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, 17359–17364.

(47) Stock, G.; Ghosh, K.; Dill, K. *J. Chem. Phys.* **2008**, *128*, 194102.

(48) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

CT1004966