Quantifying multiscale noise sources in single-molecule time series

Christopher P. Calderon[†,*], Nolan C. Harris [¶], Ching-Hwa Kiang [¶], and Dennis D. Cox[§]
[†] Department of Computational & Applied Mathematics,
[§] Department of Statistics,
[¶] Department of Physics and Astronomy,
Rice University, Houston, TX 77005-1892, USA.

corresponding author: calderon@rice.edu September 5, 2008

Abstract:
When analyzing single-molecule data, a low-dimensional set of system observables typically serve as the observational data. We calibrate stochastic dynamical models from time series that record such observables. Numerical techniques for quantifying noise from multiple time-scales in a single trajectory, including experimental instrument and inherent thermal noise, are demonstrated. The techniques are applied to study time series coming from both simulations and experiments associated with the nonequilibrium mechanical unfolding of titin's I27 domain. The estimated models can be used for several purposes: (1) detect dynamical signatures of "rare events" by analyzing the effective diffusion and force as a function of the monitored observable; (2) quantify the influence that conformational degrees of freedom, which are typically difficult to directly monitor experimentally, have on the dynamics of the monitored observable; (3) quantitatively compare the inherent thermal noise to other noise sources, e.g. instrument noise, variation induced by conformational heterogeneity, etc.; (4) simulate random quantities associated with repeated experiments; (5) apply pathwise, i.e. trajectory-wise, hypothesis tests to assess the goodness-of-fit of the models and even detect conformational transitions in noisy signals. These items are all illustrated with several examples.

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

1 Introduction

Recent advances in single-molecule (SM) experimental techniques have allowed researches to explore small scale systems with high spatial and temporal resolution. This has allowed researchers to gain a better understanding of the kinetics and thermodynamics of various complex biological systems, including SM studies of proteins and nucleic acids.[1–10] Often the dynamics are best described by stochastic models due to the inherent thermal noise which is non-negligible at time/length scales associated with current SM experiments.[11]

However, many challenges still exist in SM experiments. The magnitude of the effective diffusive noise associated with a low-dimensional system observable is often not known *a priori* and it tends to be state-dependent in complex systems.[3,12–15] Another complication encountered in SM experiments is that conformational degrees of freedom correlate heavily with the dynamics of the monitored observable, but are usually difficult to measure directly in experiments. In simulations these relevant degrees of freedom may be nonlinear collective coordinates.[16] Such degrees of freedom can often significantly influence the dynamic and static properties of SM observables. For a concrete example, suppose one is doing an experiment of protein unfolding. The root-mean-square displacement (rmsd) partially characterizes the conformational state of the protein. This type of quantity is not usually accessible in dynamic SM experiments, but the rmsd variability can heavily influence the distribution of an end-to-end coordinate and cause heavily skewed histograms of the latter.[9,15,17–20] Furthermore, unobservable conformational transitions can occur on time-scales which are fairly slow relative to the experiment. These considerations substantially complicate using a single low-dimensional model to approximate the stochastic dynamics of the entire population of SM experiments. However, we demonstrate that the variation induced by slowly evolving conformational degrees of freedom sometimes permits one to use a *collection* of low-dimensional surrogate models to describe the rich content of SM experimental time series.[15,21] In addition, atomistic simulation has also advanced considerably in recent years, and this tool can provide information about the behavior of quantities difficult to physically measure in the laboratory and aid in modeling SM systems. As progress is made in both simulation and SM experimental methods, the time and length-scales accessible to both will have appreciable overlap and will greatly assist in our understanding of the factors influencing the dynamics of complex molecules.

In this article, diffusion models are constructed using time series resulting from nonequilibrium mechanical unfolding of macromolecules and are used to help in addressing the issues mentioned above. Simulation data is generated by constant velocity steered molecular dynamics (SMD) modeling the unfolding of the I27 domain of titin. Experimental unfolding data of the same molecule is obtained using atomic force microscope (AFM) experiments to unfold engineered versions of the I27 domain of human cardiac titin.[6] Instrument noise associated with the experimental apparatus is quantified using time domain techniques.[22] Maximum likelihood type approximations are made throughout to estimate the parameters of the stochastic model; instruement/measurement noise is accounted for explicitly by the models assumed.

The estimated time-dependent diffusion models aim at summarizing the wealth of information contained in low-dimensional SM time series. Each time series observed results in the estimation of a new nonlinear time-dependent diffusion model. We refer to the entire collection of estimated models summarizing a batch of time series as the surrogate process approximations (SPAs). We show that the models can be used to:

- Approximate quantities associated with the inherent randomness of SM systems, e.g. the effective force and thermal noise.
- Detect subtle dynamical signatures of slowly evolving conformational degrees of freedom by comparing the differences observed in the measured population of models.[15,23]
- Quantify how quantities, which are difficult to directly measure in the laboratory, influence the dynamics of observable quantities.
- Compare the thermal noise to other noise sources commonly encountered in SM systems such as conformational heterogeneity[4,9,10,24–27] and instrument noise.
- Predict distributions of random quantities, like the nonequilibrium work, associated with repeated experiments/simulations.

These points are illustrated with examples from experiments and/or simulations. In addition, we demonstrate methods for quantitatively assessing the goodness-of-fit of the SPA models using time series testing methods;[28] the tests used also check the validity of the assumptions we make on the instrument noise. The article is organized as follows: Section 2 presents the computational and experimental methods, Section 3 presents the results and discussion. Section 4 contains the conclusion and outlook followed by an appendix summarizing some supplemental statistical results.

2 Materials and Methods

2.1 Local Maximum Likelihood Estimation

Stochastic differential equations (SDEs) are fit to time series coming either from MD output or AFM experiments where an external force is added into the system.[29,30] It is important to stress that each time series results in the estimation of a new SDE, so a given batch of time series results in a corresponding batch of (estimated) SDEs. The global SPA[15,21] representing the dynamics of a single time series trajectory is assumed to be a generic nonlinear diffusion of the form:

$$d\xi_t = \mu(t,\xi_t)dt + \sqrt{2}\sigma(\xi_t)dB_t \tag{1}$$

$$y_{t_i} = \xi_{t_i} + \epsilon_{t_i}, \tag{2}$$

where ξ_t represents the system observable value at time t (throughout this the end-to-end extension of the molecule stretched), B_t represents the standard Brownian motion, $\mu(\cdot,\cdot)$ the time dependent drift function and $\sigma^2(\cdot)$ represents the diffusion coefficient [1]. The drift term is time dependent because we are adding an external force. The contamination process noise (ϵ_{t_i}) does not allow us to directly observe ξ_t, instead we observe $y_{t_i} = \xi_{t_i} + \epsilon_{t_i}$ where the subscript on the time index is used to stress that our observations are discrete.

In SM systems, the complexity of the atomistic system often cannot be ignored and causes problems in developing physically based, accurate parametric SDE models[13,31–33] from *a priori* considerations. Due to this fact, we assume that the global dynamics are completely unknown *a priori*, so appealing to a standard parametric estimation scheme is problematic. To overcome the difficulty of unknown global drift and diffusion functions, we use local models[15,21,33,34] to fit the coefficients of polynomial SDEs whose functional form is motivated by overdamped Langevin equation.[15] The relevant expressions are:

$$\sigma^{\mathrm{loc}}(\xi) := \big(C + D(\xi - \xi^o)\big) \tag{3}$$

$$F^{\mathrm{Ext}}(t,\xi) := k_{\mathrm{pull}}(\lambda(t) - \xi) \tag{4}$$

$$F^{\mathrm{Int}}(\xi) := \big(A + B(\xi - \xi^o)\big) \tag{5}$$

$$\mu^{\mathrm{loc}}(t,\xi) := \frac{\big(\sigma^{\mathrm{loc}}(\xi)\big)^2}{k_B T}\Big(F^{\mathrm{Int}}(\xi) - F^{\mathrm{Ext}}(t,\xi)\Big), \tag{6}$$

where $k_B T$ represents Boltzmann's constant times the system temperature, F^{Ext} the external force applied to the system, $\lambda(t)$ is the pulling protocol (common to all experiments/simulations) we desire the observable monitored to follow, k_{pull} is the spring constant associated with the harmonic constraint used to apply the external force, F^{Int} the force due to internal molecular forces, and $\theta \equiv (A, B, C, D)$ is the local parameter vector estimated by approximate maximum likelihood estimation.[35] ξ^o is a free parameter used only for estimation purposes. Since a constant velocity protocol is used for λ, we set ξ^o to the average (temporal) value in local time series windows. The windows were formed by dividing a single global time series into M windows who all represents an equal temporal fraction of the total time series. A can be interpreted as the effective internal system force associated with the value ξ^o and B the associated linear sensitivity (similarly for C and D). The modeling ideas behind the two-scale realized volatility estimator (TSRV)[22] were used to approximate the variance of the noise process in each local window. We do not assume that the measurement noise magnitude is constant for all state values, however a "white" measurement noise is assumed. Extensions of TSRV can readily accommodate colored noise[36] if the experimental apparatus is believed to produce colored noise. The goodness-of-fit test employed here[28] can detect this type of error [2]. TSRV approximation details are summa-

[1] The thermal noise in the "internal system" has contributions coming from internal molecular fluctuations as well as solvent bombardment on the molecule and the cantilever tip. Methods for decoupling these noise sources are currently under investigation.

[2] An analysis of the autocorrelation of the temporal differences (e.g. $\{\xi_{t_{i+1}} - \xi_{t_i}\}_i^N$) of the time series

rized in the Appendix. One assumption made is that the measurement noise dominates the diffusive noise of the SDE; simple techniques for correcting the bias introduced when the diffusive noise magnitude is commensurate (but still considerably smaller) than measurement noise are also outlined in the Appendix. Recent studies have applied a Bayesian analysis to work distributions at a fixed time.[37] We would like to note that out interest is in characterizing the noise along the entire trajectory. Our methods could in principle be cast into a Bayesian framework, but we prefer a frequentist approach primarily because it facilitates assessing the goodness-of-fit of the models. The information from a TSRV inspired analysis is used in conjunction with likelihood based methods to estimate quantities describing the ξ dynamics. The fitting criterion we use is motivated by the local linear maximum likelihood type method outlined in.[35] The stage location $\lambda(t)$ was denoised using the Daubechies (5 vanishing moments) wavelet family and all measurement noise was assumed to be contained in ξ. The validity of the various assumptions (diffusive noise, local linearity, etc.) are tested in an *a posteriori* fashion using the probability integral transform based Q-test developed in Ref. [28]. It should be noted that if a physics based model is in hand, many of the pathwise testing tools presented here are still applicable and can help in testing theoretical models given nonstationary (or stationary) observations.

To obtain a global model which can be used to predict random quantities like nonequilibrium work distributions, a penalized spline was used to stitch the piecewise polynomial models together.[38] The full numerical details of this spline procedure are outlined in Ref. [39]. Briefly, a sequence of estimated local θs measured along one trajectory are used to construct both μ and σ. Information about the parameter uncertainty is used to determine a regularized global model from the local θs. The procedure is then repeated for each observed time series.

2.2 Steered Molecular Dynamics Simulations

The NAMD program[40] was used to simulate the unfolding of the I27 domain of titin which was placed in a periodic water box. The water molecules were modeled using the TIP3 model and the CHARMM 27 force field for proteins was employed. All simulations were carried out in the NpT ensemble using 20,705 total atoms. The initial atomic coordinates came from the PDB crystal structure (1TIT). This structure was then solvated in water, using the VMD plugin *Solvate* and then the system was equilibrated. The C_α of the 1^{st} residue was anchored in place using a harmonic constraint k_{pull} 100 $\text{kcal}/(\text{Å}^2\text{mol})$. In the SMD simulations, the C_α of the 89^{th} residue was pulled at constant velocity using a time dependent harmonic potential (spring constant 5 $\text{kcal}/\text{Å}^2\text{mol}$ and velocity=25 Å/ns). The time step used for integrating the SMD simulation was 1 fs and data was discretely sampled every 50 fs for estimation purposes.

suggested that the assumption of white measurement noise was reasonable (not shown) and the goodness-of-fit tests employed (results reported in Appendix) also suggested the white noise proxy was statistically acceptable for our observed time series.

2.3 AFM Experiments

Engineered proteins of eight serially linked repeats of the I27 domain of human cardiac titin (Athena ES) were used. 10 μl of protein solution, 50–100 μg/ml, was incubated on a gold substrate at room temperature for 20 min. A Multimode AFM with Picoforce option (Veeco Instruments) was used for force spectroscopy measurements. Individual protein chains were attached to a silicon nitride cantilever tip with spring constant, $k_{\rm pull} = 50$ pN/nm. The attached molecule was stretched to unfold several domains, allowed to relax back nearly to the substrate surface, and held at a constant position to allow the molecule to refold before repeating the cycle. The stretch and relax portions of the cycle were performed at a constant velocity of 50 nm/s, followed by a rest time of 30 sec. Discrete time series were recorded at the frequency of 20kHz. We present results primarily on analyzing the second force peak observed in each cycle, though the results for the other peaks are similar.

3 Results and Discussion

Figure 1 presents results obtained from SMD simulations of unfolding the I27 domain of titin. The effective force and diffusion coefficient estimated from 20 different SMD time series are displayed. The two different curve types, light-solid and red-dashed lines, denote two different batches of SMD data. The batches are distinguished by the initial position coordinates; within each batch the same positional coordinates are used for each simulation. Each trajectory uses different random initial velocities and a different random number stream to simulate the nonequilibrium unfolding of titin in a Langevin heat bath. These batches were analyzed to determine how long the persistence of the initial configurations can be felt and how this manifests itself in the estimated SPA model approximating this system.[33,41–43] For these trajectories, the diffusion $\sigma^2(\cdot)$ functions are appreciably different for the two batches of curves. An approximation of the statistical uncertainty associated with estimation is quantified in Supp. Mat. Fig. 7 and this curve indicates that the differences cannot be attributed only to estimation uncertainty associated with a finite sample time series.

We demonstrate that the variability observed in these batches of measured SPA curves has physical relevance. Different features of the same titin SMD simulation data displayed in the previous figure are shown in Fig. 2. Here, the color-coding for the curves is the same, but the $\rm rmsd$ of the titin molecule from the crystal structure (PDB:1TIT) is plotted as a function of time in the top panel. The bottom panel plots the nonequilibrium work added into the system. In all of the constant velocity simulations and experiments, when we mention "work", we are referring to the nonequilibrium work definition given in Refs.,[31,44,45] namely $W_T \equiv \int_0^T \left(k_{\rm pull}(\lambda(t) - \xi_t)\right) \frac{d\lambda(t)}{dt} dt$. Close inspection of the $\rm rmsd$ evolution of the two batches reveals that in the initial temporal segment the paths corresponding to two different conformational coordinate initial conditions appear similar, but a distinction between the two batches of curves becomes apparent at later times ($\approx 0.6 - 0.8$ ns). This distinction

occurs up until the well-studied I27 domain rupture event occurring around an extension of $\approx 10 - 15$Å in SMD simulations[2,46–49] [3]. In this application, these coordinates have appreciable "memory" relative to the time-scale of this simulation.[41,50–52] Recall how the diffusion coefficient, $\sigma^2(\cdot)$ in Fig. 1 depended heavily on the initial coordinate conditions used in the simulation.

The differences in dynamical responses can also be attributed in part to "unresolved orthogonal coordinates".[42,50] For example, it is known that the number of hydrogen bonds in the molecule correlate heavily with its mechanical strength.[46,49,53] Other possible "unresolved orthogonal coordinates" can be related to collective conformational degrees of freedom. For example collective motions associated with allosteric motion are known to modulate the dynamical response of simple low-dimensional models.[15,20] These types of collective coordinates are typically associated with relatively slow time-scales. Explicitly including a deterministic memory kernel in a scalar model, as in spirtit of generalized Langevin equation,[52] may not be able capture the effects of these unresolved collective coordinates.

The variation observed in the functions estimated for the SPA description indirectly reflects the variability introduced by both "long-time" memory and conformational heterogeneity. Comparing the information in a collection/population of SPA models, e.g. comparing the different drift and diffusion coefficients, provides one means for quantify this type of variation. The population of SPA diffusion models, which do not explicitly model "memory", provides information about the effective dynamics of the underlying complex system. The variability in the SPA models is due to time-scales slow relative to the experiment or simulation.[15,39,54] It is usually challenging to numerically carry out various statistical inference procedures for generalized Langevin models, even for stationary signals.[52,55] The use of collection of SPA models to quantify the effects due to slow time-scale motion is one alternative to using a generalized Langevin description and/or including additional degrees of freedom in the effective model.

Aside from population differences, the SPA model coefficient can also be used to identify "rare events". A large "outlier" rmsd curve in an unfolding experiment suggests premature mechanically induced denaturation.[56] The dark highlighted curve is used to identify one such "outlier" in the rmsd plot; the curves corresponding to this particular simulation trajectory are also highlighted in Figs. 1-2. The relatively low value of the work path associated with this trajectory, which is measured from the simulation directly, also suggests that significant mechanical denaturation has occurred earlier than usual. Determining the frequency of such rare events has high relevance to free energy computations using nonequilibrium simulation data.[31,57–59] In complex molecules, extracting the frequency of such events is challenging, but works that use both unfolding and refolding data have demonstrated it might be worth pursuing further.[60–63]

[3] In the constant velocity experiments studied, this extension occurs roughly at a time of roughly 0.75 ns. The harmonic constraint used was weak enough to have a readily apparent difference between the target value $\lambda(0.75) \approx 19$Å and the underlying molecular extension at this time. This discrepancy is of no concern to our dynamical modeling.

Our interest in this article is in extracting as much information as possible from a low-dimensional observable time series. Dynamically monitoring a quantity like the rmsd is a luxury we have in simulations, but analogous structural metrics are not usually accessible in experiments. In experiments, some structural metric may be known to be physically important, but is not directly accessible in the laboratory. In these situations, the use of a *collection* of SPA models to summarize system information is also appealing. This is another area where this type of modeling can help in understanding complex SM data.

We have shown how information in the SPA functions can be used to detect conformational differences in the underlying molecule by inspecting a collection of SPA models. Now we move on to show how the *individual* estimated models can be used to make quantitative predictions about variability induced by thermal noise. We demonstrate this first using the titin SMD data where each nonequilibrium simulation was started with a substantially different coordinate initial condition [4]. The top panel of Fig. 3 plots the work added into the system as a function of simulation time. The vertical line corresponds to the time where $\lambda \approx 18.5$Å. Under our simulation conditions, the I27 domain has typically ruptured at this point. We randomly selected 10 curves from the population of 55 to see how well we could approximate features of the work distribution with limited trajectory data. The middle panel plots various estimates (discussed later) of two nonequilibrium work densities obtained by analyzing a subset of the SMD paths.

For the 10 curves selected, we noted the work measured directly from the SMD simulation and also calibrated 10 SPA diffusion models using each of the 10 trajectories. The calibrated SPA models were used to generate 2500 realizations using the same initial condition as the corresponding SMD simulation and the random work introduced to the SPA simulation was recorded. The normalized histogram obtained from the 10×2500 SPA work is plotted as bars in the bottom panel of Fig. 3. The solid curves in the same figure show the contribution coming from each of the 10 different SPA models (summing these curves would result in the histogram represented by bars). Note that many of the work distributions displayed as solid lines in the bottom panel do not appreciably overlap. This means that conformational heterogeneity[4,9,10,24–27] in the simulation data will not allow a *single* scalar diffusion model to accurately capture the factors making significant contributions to the random work process in this system. Given the complexity of the many-body SMD simulation, this is not surprising, but it should be noted that each of the SPA models do pass pathwise goodness-of-fit tests (see Appendix Fig. 1) indicating that each SPA model adequately approximates the SMD time series used to calibrate it. Also note that a *collection* of SPA diffusion models, each calibrated from one SMD trajectory, can approximate the the features of the many-body SMD responses (here an ensemble of work paths). Again we stress that we only steer the end-to-end distance of the molecule with a biasing potential. We force this observable to change at a rate much faster than it typically does in the unperturbed case; this rate is very fast relative to the time-scales associated with other slowly changing conformational degrees of freedom which are relevant to the β-sandwich

[4]We drew configurations from equilibrated samples where the molecular extension was biased by our guiding potential to be at the value observed in the crystal structure.

structure of this molecule.[46,47,65,66] The conformational degrees of freedom modulate the dynamics of ξ, i.e. although we only steer the ξ coordinate the other degrees of freedom in the system are coupled to this observable. On the time-scale of the SMD simulation, these degrees of freedom are effectively "stuck" in one region of phase space. The variation we observe in the SPA curves (see Fig. 1) can be explained by the factors that cause variability in the SPA models being associated with time-scales slow relative to the simulation or experiment. This type of time-scale separation is not unique to titin simulations, it has been observed in various experimental and simulations shown here and elsewhere.[15,18,19,21,23,39,54] Analyzing the diversity in a population of SPA histograms, i.e. the outlined curves in the middle panel, is one way of indirectly quantifying the variation induced by conformational degrees of freedom in this type of setting.

Next, we demonstrate how information in the middle panel of Fig. 3 can be used in a "SPA nonparametric model bootstrapping" type scheme. The details of the scheme used here are provided in the Appendix. The scheme attempts to quantitatively account for variability due to slow time-scale conformational degrees of freedom as well as variability due to thermal noise. In the middle panel of Fig. 3, we plot the normalized work histogram coming from the 55 simulations. We also plot a standard nonparametric density estimate using all 55 work values at the time point studied [5]. We then selected 10 curves randomly from the 55 total curves and fit a Gaussian density to the observations. We also compute the nonparametric density estimate using the smaller data set in addition to plotting the results from applying the "SPA nonparametric model bootstrapping" scheme to the corresponding trajectories. Note in the larger population of nonequilibrium work, that two modes are apparent in both the raw data and the nonparametric density estimate. Empirically determining the number of modes in a histogram is extremely difficult if only a small number of random variables from the distribution are available. This would be the situation we faced if only 10 work curves were available to us. Limited data situations are frequently encountered in simulations due to computational cost limitations and can also be relevant to SM experiments.[49,68] However, by using a small number of trajectories along with the SPA modeling ideas laid out here, we can more easily determine that there are indeed two underlying modes in the data. This is possible because the SPA models have predictive capability. By simulating one SPA model, we can approximate randomness due to inherent thermal noise; the width of the solid densities in the bottom panel reveal this information. Variation induced by conformational heterogeneity can be determined by comparing the output of *multiple* SPA models. The two sources of variation can be quantitatively compared using a relatively small number of SMD simulation trajectories. This type of quantitative tool can possibly help in understanding many different complex SM systems. Particularly systems where " multiple conformational states" cause broadened observable histograms and/or multiple modes.[4,9,10,24–27]

Next we present results where we approximate the effective force and diffusion from experimental AFM time series in the presence of thermal and instrument noise. The AFM

[5]The bandwidth (h) was determined by $h = \hat{\sigma} \times (n^{-\frac{1}{5}})$, where $\hat{\sigma}$ denotes the empirical standard deviation and n the number of observed samples. The Gaussian kernel was employed in the nonparametric estimation.[67]

force extension data consists formed the "sawtooth pattern" associated with titin's I27 domain.[6,65] Some representative trajectories coming from the experimental apparatus are plotted in Appendix Fig. 2. To minimize variation due to the tip attachment point, we captured a titin molecule on the AFM tip and retained the same molecule for a sequence of force extension cycles. Due to the nature of the non-covalent forces binding the titin molecule to the AFM tip, we could only retain the same molecule for a limited number of repeated force extension cycles. For every distinct "force peak" observed, we estimated a separate SPA model. More specifically, each global experimental time series used to estimate a SPA model contains only the increasing portion of a force peak (we do not attempt to model the rapid drop in force after domain rupture). Results from the second and third peaks observed are plotted in Fig. 4. Results from the first peak observed are not shown because they are more likely to be affected by nonspecific binding artifacts.[65,69]

Many of the trends observed in other works involving force-clamp[70] and dynamic force modulation[69] experiments probing the molecular stiffness and internal friction of titin I27 domain appear to also hold in the constant velocity experiments we carried out and analyzed. Namely the gradient of the force shows a generally decreasing magnitude as more domains are unfolded and the internal effective diffusion coefficient demonstrates a roughly decreasing trend as extension increases within one peak. As witnessed in Ref. [69], we also appear to observe that the effective friction, inversely related to the effective diffusion coefficient in our fitted SPA models, also appears to display a slight decrease as the number of unfolded I27 domains increase. However, the estimation uncertainty associated with this effective diffusion coefficient is relatively large. In addition the diffusion term contains noise from multiple sources besides the "regular" thermal noise associated with a molecule in a solvent heat bath (e.g. cantilever solvent bombardment). Currently we are researching methods for using overlapping windows, variance reduction techniques, and methods which account more explicitly for inherent instrument noise to refine the estimation of the effective diffusion coefficient associated with the molecule given its potential relevance to the internal molecular friction of a single macromolecule and the insight it can provide about the effective dynamics on a rough free energy surface.[70]

Figure 5 plots the work distribution predicted by the SPA models calibrated from two different trajectories, each panel corresponding to one trajectory, evolving over time as a sequence of histograms. The value measured directly from the experiment at the corresponding point, determined by $\lambda(\tau)$, is also plotted. The observed experimental work paths are consistent with the SPA work densities for all times observed. Results from the other six curves are similar and this demonstrates that we can, on a pathwise basis, reasonably model the uncertainty due to thermal and instrument noise by simulating the estimated SPA model. We can identify that repeated experiments do show variability which might be attributed to conformational heterogeneity; the details of how the titin molecule refolds at "zero extension" may influence the effective molecular force measured. Large, possibly conformationally induced variation, in dynamical responses has also been recently reported in different experiments probing titin's mechanical properties.[70] Note that we make observations on larger time-scales in the experiment, so making direct analogies to simulation

results is problematic due to the disparate scales involved.[49,53] Again advances in simulation techniques may overcome this problem in the near future.[68,71] The difference between the simulated evolving work distribution are not excessively large, but we believe physically significant conformational differences do exist on a trajectory-wise basis due to the fact the fact that observing a "force-hump" appears to be a random event[2,46–49] .

Finally we subject our titin I27 experimental data to goodness-of-fit tests. Our interest is both in determining the validity of the models assumptions and on attempting to detect subtle phase transitions using our models and the observed time series. Regarding the latter item, it is known that a "force-hump" can be observed in stretching the I27 domain of titin,[2,49] but this force hump can be difficult to observe in constant velocity AFM experiments.[48] Here we analyze data coming from the second force peak obtained by pulling the same titin molecule repeatedly. Fig. 5 displays the Q-test results obtained by analyzing 8 force extension curves. Again we remind the reader that each force extension curve was used to calibrate a new SPA diffusion model. The number of local models in each case was $M = 15$. The estimated SPA models were then used along with the observed data and the Q-test[28] to determine the goodness-of-fit of each SPA model. Out of the 8 AFM trajectories analyzed, 2 were associated with a parameter set that was rejected at the significance level $\alpha = 0.01$ [6]. In the rejected models, the local models broke down at a force ≈ 100 pN where the force-hump "transition" is known to occur in this system.[48,49] The raw data from the AFM is plotted in the inset for the two rejected cases. In the insets the estimated SPA $F^{\mathrm{Int}}(\cdot)$ function is plotted with "o-" lines as well as a wavelet and penalized spline smoothing of the raw AFM data, i.e. $F^{\mathrm{Ext}}(\cdot)$. In the top right plot, the hump is visible using standard smoothing techniques. Our time series methods confirm that something statistically significant/detetable is changing in the dynamics. In the curve on the left, the time series based methods appear to detect a transition that the other smoothing techniques do not. A sudden change in "thermal noise" magnitude occurs around the 100 pN, which is likely an artifact of the transition. This provides another application of our dynamical models: they can be used to determine when a statistically significant change in the stochastic dynamics occurs. Perhaps more importantly, the testing procedure employed gives us a quantitative metric to test our various model assumptions with.

4 Conclusion and Outlook

Methods for analyzing single-molecule data in a pathwise fashion were presented. Each trajectory had the measurement noise quantified and this influence of this random noise sources was included in the model estimate. The individual SPA models passed pathwise goodness-of-fit tests. These tests simultaneously tested various model assumptions, e.g.

[6] It should be noted that we use carry out a test in each local window and determine our critical value using a single hypothesis test. Alternatively we could aggregate the PIT results from the M windows into a vector for each observed time series and carry out simultaneous multiple comparison, adjusting the significance level accordingly (e.g. use Bonferroni's method).[73] In this situation it would make rejecting a model more difficult.

overdamped diffusive dynamics, white measurement noise, etc. The models also demonstrated predictive power, i.e. the probable range of work values predicted by the models was consistent with the simulations/experimental data. More importantly the methods were shown to indirectly quantify the variation induced by conformational degrees of freedom. In experiments this information is typically unobservable. A nonparametric resampling scheme which utilized the work distribution predicted by our surrogate models was demonstrated to qualitatively predict the shape of certain process functionals, e.g. the nonequilibrium work distribution. This is relevant because it allows a researcher studying SM systems to better approximate the basic shape of a nontrivial work distribution using a small number of samples. This can be used for various purposes, e.g. the reliability of a nonequilibrium free energy estimate depending on a non-Gaussian work distribution can more readily be assessed.[3,6,15,45,59] The collection of SPA models could in principle also be used to predict mean first passage times of complex biolmolecules.[74] In addition, "outlier" curves were shown to correlate with physically relevant structural information which is not typically directly accessible in dynamical experiments.[15,23] This type of information cannot be inferred from a single SPA model along, but required one to analyze a population of SPA models calibrated from different trajectories. Also certain transitions were detected apparently using the information contained in our estimated models, e.g. transitions known to exist appeared to be detected by pathwise goodness-of-fit tests.

Data-driven numerical tools for analyzing complex systems with a relatively small number of system observables[75,76] will likely significantly assist researcher in understanding the rich data sets coming from detailed computer simulations and high resolution SM experiments.[1–8,10] Other systems ranging from double and single stranded DNA,[23] ion-channel proteins,[39,54] and small polypeptides[15] have demonstrated that fingerprints of large collective coordinate changes can be detected by analyzing the effective diffusive noise and force as a function of state in both simulations and experiments. In the future, hypothesis testing methods similar to those demonstrated here may be used to help identify and/or more precisely determine the location of "conformational transitions" in SM systems not understood as well as titin. Note that although our focus here was on data-driven methods, many of the estimation and testing procedures could possibly be applied if a "first principles" physically based model is available.[4,53,77]

5 Acknowledgments

CPC thanks NSF DMS 0240058, NSF ACI-0325081, and the computer support provided by NSF CNS-0421109 & a partnership between Rice AMD and Cray. The work of NCH and CPC was supported in part by a training fellowship from the Nanobiology Training Program of the W. M. Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia (NIH T90DK70121-04 and and 5 R90 DK71504-04). CHK thanks NSF DMR-0505814 and Welch Foundation C-1632 for support. DDC thanks NSF DMS-0505584.

References

1. Hegner, M., Smith, S., , and Bustamante, C. *Proc. Natl. Acad. Sci. USA* **96**, 10109 – 10114 (1999).

2. Clausen-Schaumann, H., Rief, M., and Gaub, H. E. *Nat. Struct. Biol.* **6**, 346–349 (1999).

3. Collin, D., Ritort, F., Jarzynski, C., Smith, S., Tinoco, Jr., I., and Bustamante, C. *Nature* **437**, 231–234 (2005).

4. Walther, K., Brujic, J., Li, H., and Fernandez, J. *Biophys. J.* **90**, 3806–3812 (2006).

5. Evans, E. and Calderwood, D. *Science* **316**, 1148–1153 (2007).

6. Harris, N., Song, Y., and Kiang, C. *Phys. Rev. Lett.* **99**, 068101 (2007).

7. Ke, C., Humeniuk, M., S-Gracz, H., and Marszalek, P. *Phys. Rev. Lett.* **99**, 018302 (2007).

8. Fernandez, J. and Li, H. *Science* **303**, 5664 (2003).

9. Liu, S., Bokinsky, G., Walter, N., and Zhuang, X. *Proc. Natl. Acad. Sci. USA* **104**, 12634–12639 (2007).

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

11. Moffitt, J., Chemla, Y., Smith, S., and Bustamante, C. *Annual Review of Biochemistry* **77**, 19.1–19.4 (2008).

12. Sigg, D., Qian, H., and Bezanilla, F. *Biophys. J.* **76**, 782–803 (1999).

13. Hummer, G. *New J. Phys.* **7**, 1–14 (2005).

14. Chahine, J., Oliveira, R., Leite, V., and Wang, J. *Proc. Natl. Acad. Sci. USA* **104**, 14646 (2007).

15. Calderon, C. and Chelli, R. *J. Chem. Phys.* **128**, 145103 (2008).

16. Krishnan, J., Runborg, O., and Kevrekidis, I. *Comp. Chem. Eng.* **28**, 557–574 (2004).

17. Procacci, P., S., M., Barducci, A., Signorini, G., and Chelli, R. *J. Chem. Phys.* **125**, 164101 (2006).

18. Lu, Z., Hu, H., Yang, W., and Marszalek, P. *Biophys. J.: Biophys. Lett.* , L57–L59 (2006).

19. Paramore, S., Ayton, G., and Voth, G. *J. Chem. Phys.* **14**, 105105 (2007).

20. Calderon, C. and Arora, K. *submitted to J. Chemical Theor. and Comp.* (2008).

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

22. Zhang, L., Mykland, P., and Ait-Sahalia, Y. *Journal of the American Statistical Association* **100**, 1394–1411 (2005).

23. Calderon, C., Chen, W., Harris, N., Lin, K., and Kiang, C. *submitted to J. Physics: Condensed Matter* (2008).

24. Zhuang, X., Kim, H., Pereira, M., Babcock, H., Walter, N., and Chu, S. *Science* **296**, 1473–1476 (2002).

25. Min, W., Gopich, I., English, B., Kou, S., Xie, X., and Szabo, A. *J. Phys. Chem. B* **110**, 20093–20097 (2006).

26. Vendruscolo, M. and Dobson, C. *Science* **313**, 1586 (2006).

27. Lange, O., Lakomek, N., Fars, C., Schrder, G., Walter, K., Becker, S., Meiler, J., Grubmller, H., Griesinger, C., and de Groot, B. *Science* **320**, 1471–1475 (2008).

28. Hong, Y. and Li, H. *The Review of Financial Studies* **18**, 37–84 (2005).

29. Balsera, M., Stepaniants, S., Izrailev, S., Oono, Y., and Schulten, K. *Biophys. J.* **73**, 1281 (1997).

30. Li, P. and Makarov, D. *J. Chem. Phys.* **119**, 9260 – 9267 (2003).

31. Park, S. and Schulten, K. *J. Chem. Phys.* **120**, 5946–5961 (2004).

32. Hummer, G. and Kevrekidis, I. *J. Chem. Phys.* **118**, 10762–10773 (2003).

33. Calderon, C. *Mutliscale Modeling and Simulation* **6**, 656–687 (2007).

34. Fan, J., Fan, Y., and Jiang, J. *Journal of American Statistical Association* **102**, 618–631 (2007).

35. Jimenez, J. and Ozaki, T. *J. Time Series Analysis* **27**, 77–97 (2005).

36. Ait-Sahalia, Y., Mykland, P., and Zhang, L. Working Paper 11380, National Bureau of Economic Research, May (2005).

37. Maragakis, P., Ritort, F., Bustamante, C., Karplus, M., and Crooks, G. *J. Chem. Phys.* **129**, 024102 Jul (2008).

38. Ruppert, D., Wand, M., and Carroll, R. *Semiparametric Regression*. Cambridge University Press, New York, (2003).

39. Calderon, C., Martinez, J., Carroll, R., and Sorensen, D. *submitted to PRE* (2008).

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

41. Kupferman, R. and Stuart, A. *Physica D* **199**, 279–316 (2004).

42. Chorin, A., Kast, A., and Kupferman, R. *Proc. Natl. Acad. Sci. USA* **95**, 4094–4098 (1998).

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

44. Jarzynski, C. *Phys. Rev. E.* **56**, 5018–5035 (1997).

45. Hummer, G. and Szabo, A. *Proc. Natl. Acad. Sci. USA* **98**, 3658–3661 (2001).

46. Lu, H., B.Isralewitz, Krammer, A., Vogel, V., and Schulten, K. *Biophys. J.* **75**, 662671 (1998).

47. Marszalek, P., Lu, H., Li, H., Carrion-Vazquez, M., Oberhauser, A., Schulten, K., and Fernandez, J. *Nature* **402**, 100–103 (1999).

48. Higgins, M., Sader, J., and Jarvis, S. *Biophys. J.* **90**, 640–647 (2006).

49. Sotomayor, M. and Schulten, K. *Science* **316**, 1144 – 1148 (2007).

50. Zwanzig, R. *Nonequilibrium Statistical Mechanics*. Oxford Univeristy Press, New York, (2001).

51. Kou, S. and Xie, X. *Phys. Rev. Lett.* **93**, 18 (2004).

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

53. D.E. Makarov and P.K. Hansma and H. Metiu. *J. Chem. Phys.* **114**, 9663 (2001).

54. Calderon, C., Janosi, L., and Kosztin, I. *working paper* `http://www.caam.rice.edu/~cpc1/drafts/cjk_08.pdf` (2008).

55. Horenko, I., Hartmann, C., and Schütte, C. *Phys. Rev. E* **76**, 016706 (2007).

56. Li, M., Hu, C., Klimov, D., , and Thirumalai, D. *Proc. Natl. Acad. Sci. USA* **103**, 93–98 (2006).

57. Jarzynski, C. *Phys. Rev. Lett.* **78**, 2690–2693 (1997).

58. Crooks, G. E. *J. Stat. Phys.* **90**, 1481–1487 (1998).

59. Jarzynski, C. *Phys. Rev. E.* **73**, 046105 (2006).

60. Shirts, M., E., B., Hooker, G., and Pande, V. *Physical Review Letters* **91**, 140601 (2003).

61. Kosztin, I., Barz, B., and Janosi, L. *J. Chem. Phys.* **124**, 064106 (2006).

62. Chelli, R., Marsili, S., and Procacci, P. *Phys. Rev. E* **77**, 031104 (2008).

63. Minh, D. and Adib, A. *Phys. Rev. Lett* , 180602 (2008).

64. Humphrey, W., Dalke, A., and Schulten, K. *Journal of Molecular Graphics* **14**, 33–38 (1996).

65. Carrion-Vazquez, M., Oberhauser, A., Fisher, T., Marszalek, P., Li, H., and Fernandez, J. *Progress in Biophysics and Molecular Biology* **74**, 63–91 (2000).

66. Becker, N., Oroudjev, E., Mutz, S., Cleveland, J., Hansma, P., Hayashi, C., Makarov, D., and Hansma, H. *Nature Materials* **2**, 282 (2003).

67. Scott, D. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York, (1992).

68. Maragakis, P., Lindorff-Larsen, K., Eastwood, M., Dror, R., Klepeis, J., Arkin, I., Jensen, M., Xu, H., Trbovic, N., Friesner, R., Palmer, A., and Shaw, D. *J. Phys. Chem. B* **112**, 6155–6158 (2008).

69. Kawakami, M., Byrne, K., Brockwell, D., Radford, S., and Smith, D. *Biophys. J.* **91**, L16–L18 (2006).

70. Khatri, B. S., Byrne, K., Kawakami, M., Brockwell, D., Smith, D., Radford, S., and McLeish, T. *Faraday Discuss* **DOI: 10.1039/b716418c** (2008).

71. Simms, A., Toofanny, R., Kehl, C., Benson, N., and Daggett, V. *Protein Engineering, Design & Selection* **21**, 369377 (2008).

72. Daubechies, I. *Ten Lectures on Wavelets*. SIAM, Philadelphia, (1992).

73. Lehmann, E. and Romano, J. *Testing Statistical Hypotheses*. Springer-Verlag, (2008).

74. Kopelevich, D., Panagiotopoulos, A., and Kevrekidis, I. *J. Chem. Phys.* **122**, 044908 (2005).

75. Kevrekidis, I., Gear, C., and Hummer, G. *AIChE Jounral* **50**, 474–489 (2004).

76. Dudko, O. K., Mathe, J., Szabo, A., Meller, A., and Hummer, G. *Biophys. J.* **92**, 4188–4195 (2007).

77. Hummer, G. and Szabo, A. *Biophys. J.* **85**, 5–15 (2003).

78. Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL, (1994).

79. Ramsay, J. and Silverman, B. *Functional Data Analysis*. Springer-Verlag, New York, (2005).

80. Ait-Sahalia, Y. and Mancini, L. *J. Econometrics* ((to appear)).

6 Appendix

6.1 Monte Carlo Simulation Diagnostic Figures

The captions describe information which was alluded to in the main text.

6.2 SPA Nonparametric Resampling Procedure

The scheme below was carried out in order to generate the nonparametric densities discussed in the main text. An explanation and physical interpretation of each step is provided below the procedure.

1. Collect N time series from either simulation or experiment. In this paper, these N time series were used to estimate a collection of global drift functions, $\mu_i(\cdot,\cdot)$, and diffusion functions, $\sigma_i(\cdot)$ for $i = 1, \ldots, N$.

2. Simulate K sample paths using the SDE dictated by $\mu_i(\cdot)$ and $\sigma_i(\cdot)$. For the initial condition, use that of the corresponding time series [7]. Do this for all N sets of $\mu_i(\cdot,\cdot)$ and $\sigma_i(\cdot)$.

3. Draw N' different random variables U_1 which are uniformly distributed over the integers $\in [1, N]$.

4. For each N' draw above, draw another random variable and denote this by U_2. A uniform distribution over the integers $\in [1, K]$ is the law associated with U_2. A realization of U_2 is used to resample a work value from the histogram generated in Step 2. For example, if in the step above, we obtain the realizations U_1=2 and $U_2 = 32$, we would look up the 32^{nd} work value generated by $\mu_2(\cdot,\cdot)$ and $\sigma_2(\cdot)$ and then store this value.

5. Refine the estimate the work density of the simulated process by using the empirical data generated in the step above and a nonparametric density estimate.[6][7] Each histogram obtained by recording the results above will consist of N' measurements (for

[7] Alternatively, if one believes the initial condition is "equilibrated" at a value, one could set v_{pull} to zero in the SPA model and obtain the stationary distribution.

our nonparametric density estimated we use a univariate bandwidth suggested in,[67] i.e. $h = \hat{\sigma} N'^{-\frac{1}{5}}$, where $\hat{\sigma}$ is the empirical variation in the sample). Save the density estimate onto disk.

6. Repeat steps 3-5 D times, then average the density estimates.

The first step generates the diffusion models based on observational data. Each time series in 1) gives a different model. The differences are due both to statistical uncertainty and also to different conformational details underlying the system (discussed in text).

In Fig. 3, $M = 25$ local parameter vectors were used to generate 10 drift and diffusion curves from the $N = 10$ observed time series. The second step sets up a collection of work histograms which we will continually resample from. We set $K = 2500$ and $N' = N$. The third step attempts to include the variability of conformational noise. A scheme motivated by bootstrapping ideas[78] is used to accomplish this. The work resamping is similar in spirit to a traditional bootstrapping scheme, however it should be noted that we are doing two types of resampling. When we draw U_1 we are resampling functions (or models), so it should be viewed as a type of functional bootstrapping.[79] N is the same as the observational data in order to approximate finite sampling noise. The fourth step is used to simulate the effects of classic thermal noise assuming a fixed initial positional conformation. The procedure is repeated to average and can also be used to give us an idea of variability caused by finite sample sizes.

6.3 Two-Scale Realized Volatility

We attempt to model noise coming either from fast-scale motion of the dynamics or experimental apparatus noise as a white noise contamination preventing us from directly monitoring the system observable of interest (ξ). Instead of directly observing ξ, we observe the discrete process y_{t_i} (see Eq. 2). To estimate the variance of ϵ_t from our frequently sampled time series, we appeal to modeling ideas falling under the label Two-Scale Realized Volatiliy (TSRV). The basic idea for estimating the variance of ϵ_t within this framework is outlined in Ref. [22], we summarize it here. First compute the following:

$$[Y, Y]_T^{(\text{all})} := \sum_{i=0}^{n} (Y_i - Y_{i-1})^2 \tag{7}$$

in the above, one using all n temporal observations. If the sampling frequency is fairly large, the signal to noise ratio will be such that the signal contained in the ϵ_t process dominates that of the ξ_t process. More formally,

$$[Y,Y]_T^{(\mathrm{all})} \xrightarrow{\mathcal{L}} \langle \xi, \xi \rangle_T + 2n\mathbb{E}[\epsilon^2] + \kappa\eta \tag{8}$$

$$\kappa \equiv \left(4n\mathbb{E}[\epsilon^4] + \frac{2T}{n} \int_0^T \sigma_t^4 dt \right)^{1/2} \tag{9}$$

where η denotes a random variable following the standard normal distribution ($N(0,1)$) and the symbol $\xrightarrow{\mathcal{L}}$ denotes convergence in distribution.[22] The so-called quadratic variation, denoted by $\langle \xi, \xi \rangle_T$, associated with the dynamics of interest is an $\mathcal{O}(1)$ term in the diffusion models we are considering. The second term on the right-hand side is $\mathcal{O}(n)$ and dominates as $n \to \infty$. In mathematical finance, the interest is usually in the object $\langle \xi, \xi \rangle_T$. To get at this object in the TSRV framework, one subsamples the data on several grids. In other words, one skips every k observations and computes $[Y,Y]_T^{\mathrm{k}}$ where the last quantity is an estimator which uses a subsequence $Y_i, Y_i + k, \ldots$ to get a better estimate of the unobserved quadratic variation. Another idea central to the TSRV estimator is to not waist any time series data, i.e. one creates multiple subsequences from the original time series of length n (e.g. $\{Y_1, Y_k, Y_{2k}, \ldots, Y_n\}$, $\{Y_2, Y_{1+k}, Y_{1+2k}, \ldots, Y_{n-1}\}, \ldots$) and then averages the quadratic variation estimates from the resulting batch of subsequences to obtain a refined estimate denoted by $\widehat{\langle \xi, \xi \rangle}^{\mathrm{TSRV}}$. Directly approximating $\mathbb{E}[\epsilon^2]$ by $[Y,Y]_T^{(\mathrm{all})}/(2n)$ is often accurate in typical high-frequency financial data sets (and in experimental situations where experimental noise is fairly large and data is sampled frequently), but we will show that some care needs to be taken when these ideas are applied to frequently sampled molecular dynamics data.

Before continuing on how we deal with the last issue mentioned, we would like to note that we are trying to estimate a signal which has noise contributions coming from many different time-scales. Some noise sources are considered physically interesting (e.g. the magnitude of the thermal noise as a function of ξ) and others are considered uninteresting (e.g. ϵ_t). If these sources can be reasonably be quantitatively approximated, there are techniques which can utilize information contained in the entire time series available. Sometimes experimentalists view the thermal noise magnitude as a fundamental limit regarding experimental resolution, we would like to stress that useful physical information can be possibly be extracted even if the signal of interest is "buried" in noise. For example the influence of conformational degrees of freedom on the ξ dynamics are subtle in the titin system we studied in this article, but the pathwise estimation methods used allowed us to quantitatively measure how these sources of (relatively slow time-scale) variation influence the observed response and the estimation methods also allowed us to approximate thermal noise which might change as a function of ξ. TSRV type methods can be used in various ways to assist parameter estimation. They can provide an initial guess for a likelihood based method which simultaneously estimates the measurement noise along with the system evolution parameters (resulting in a larger optimization problem). Alternatively they can be used to divide the estimation of the measurement noise and the estimation of

parameter's governing the state into separate problems; i.e., the likelihood conditional on the TSRV type noise estimate ca be estimated. This is the viewpoint we take here (though the results shown throughout this article do not change whichever approach is employed).

Now let us return to how we deal with the contamination noise ($2n\mathbb{E}[\epsilon^2]$) being commensurate in magnitude with $\langle \xi, \xi \rangle_T$. First we utilize the assumptions behind the TSRV to estimate $\mathbb{E}[\epsilon^2]$ for each local time series; with this estimate in hand, we then determine the local parameter θ that maximizes the likelihood of the hidden Markov model in Eqn. 1. The local models used allows us to approximate the thermal noise process,σ_t, using $C + D(\xi_t - \xi^o)$. This quantity allows as to *approximate* a TSRV subsampling parameter,k^*, derived in Ref. [22]:

$$\bar{n}^* = \Big(\frac{T(\frac{4}{3} \int_0^T \sigma_t^4 dt)}{8(\mathbb{E}[\epsilon^2])^2} \Big) \tag{10}$$

$$k^* = \frac{\bar{n}^*}{n} \tag{11}$$

It should be mentioned that this rules was determined to balance the standard variance/bias trade-off encountered in nonparametric estimation under a variety of assumptions on the dynamics.[22] Our estimate of σ_t comes from an likelihood estimate which uses an estimate of $\mathbb{E}[\epsilon^2]$ coming from $[Y,Y]_T^{(\mathrm{all})}/(2n)$. This will introduce some bias into the estimated parameter θ (and hence our estimate of ,σ_t), and our k^* estimate is affected by this. However, the subsampling rule above should really just be viewed as a guide for refining our estimates of $\mathbb{E}[\epsilon^2]$. Note that this correction still assumes that the ϵ_t process is white (variants of the TSRV accounting for more complex noise structures are possible[80]). Once k^* is estimated from the data, one can then compute $\widehat{\langle \xi, \xi \rangle}^{\mathrm{TSRV}}$. To remove the bias from finite time series lengths, the following adjusted estimator is recommended in Ref.:[22]

$$\widehat{\langle \xi, \xi \rangle}^{\mathrm{adj}} = (1 - \frac{\bar{n}}{n})^{-1} \widehat{\langle \xi, \xi \rangle}^{\mathrm{TSRV}} \tag{12}$$

We subtract the above quantity from $[Y,Y]_T^{(\mathrm{all})}$ and then divide the result by $2n$ in order to get a revised estimate of $\mathbb{E}[\epsilon^2]$. Finally, with this noise estimate we then find a new parameter vector θ that maximizes the likelihood of our hidden Markov model. In simulation studies meant to mimic our MD data sets (in these controlled simulations ϵ is forced to have the properties assumed by the TSRV), we show that this procedure greatly helps in regards to accuracy (we have a known reference solution). In the empirical MD case studies, we apply this procedure and show that the goodness-of-fit tests are improved. However, surprisingly, in the titin system studied, it turns out a better approximation results when ϵ is set to zero and a diffusion is estimated (but the hidden Markov model was attempted first). This is likely due to the fact that the fast scale motion has a positive correlation (when spaced by only 50 fs) and this is better approximated by a diffusive noise.

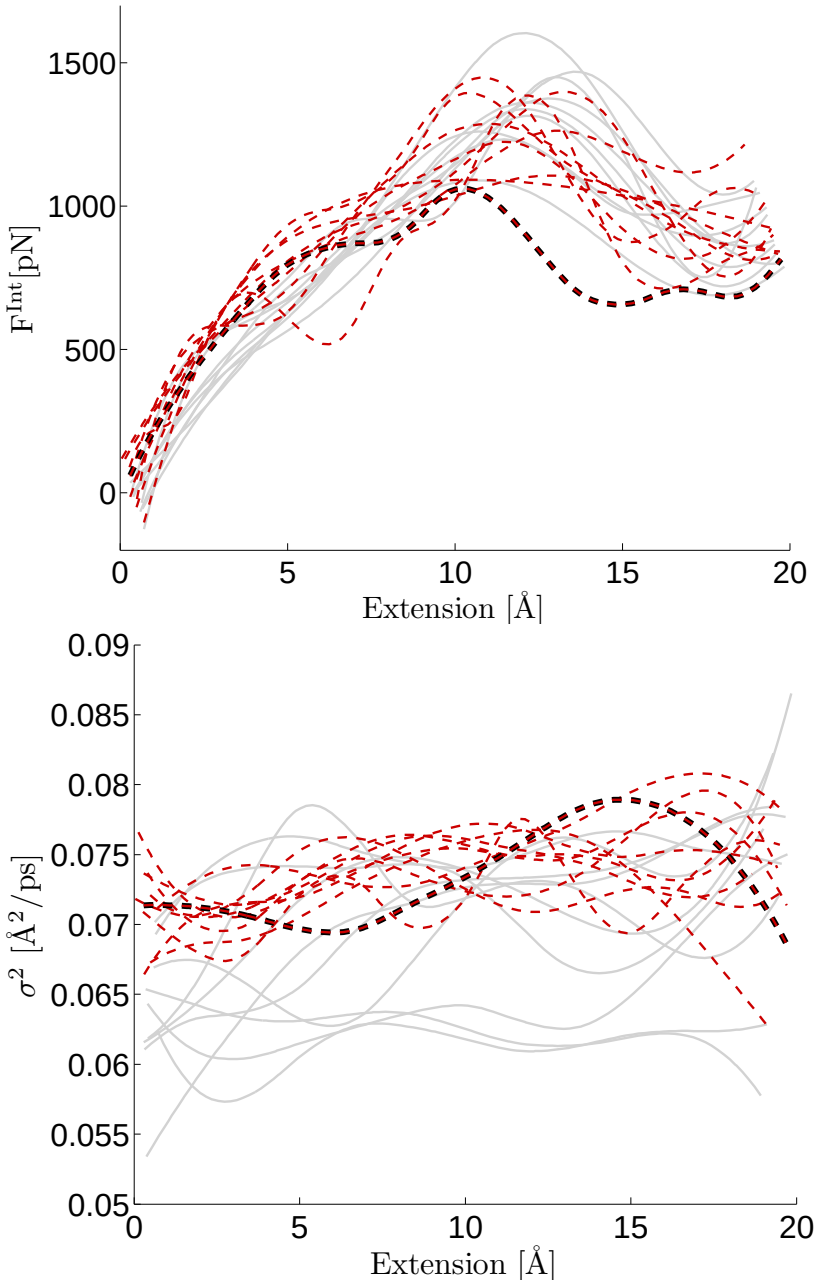


Figure 1: Simulation results from unfolding the I27 domain of titin. Two batches of 10 trajectories were simulated and each trajectory had the effective SPA force and diffusion coefficient estimated using the procedure outlined here. The first batch (solid grey lines) started one common initial coordinate set and the second batch (dotted red lines) used another initial coordinate set. Different random velocities were assigned at time zero in each case. The curve highlighted by a dark thick line denotes a trajectory where protein denaturation occurred unusually early (discussed further in text).

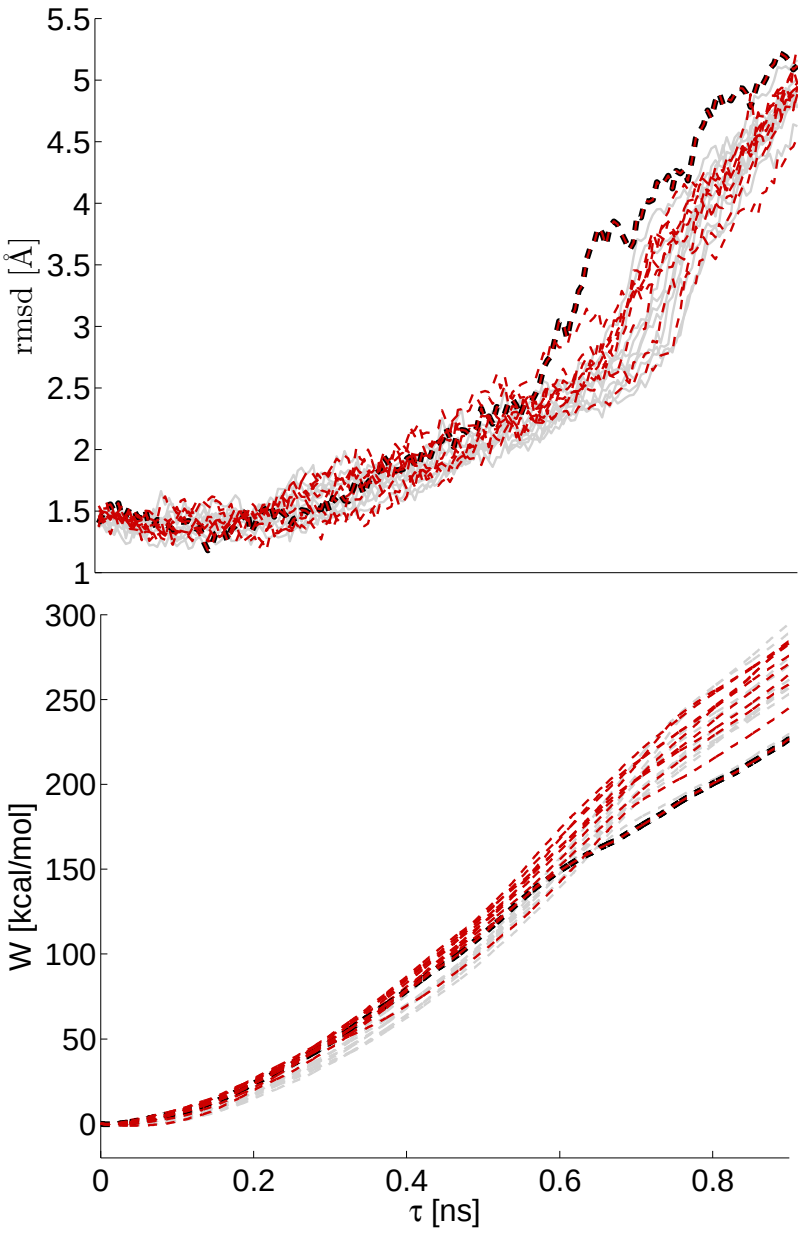


Figure 2: Two batches of 10 trajectories coming from simulations of unfolding the I27 domain of titin. The same trajectories shown in Fig. 1 are analyzed, but this time the temporal evolution of the root-mean-square-displacement (rmsd) from the crystal structure is plotted as well as that of the nonequilibrium work. Both of the aforementioned quantities were taken directly from the SMD simulation using the program VMD.[64] The curve highlighted by a dark thick line denotes a trajectory where protein denaturation occurred unusually early (discussed further in text).

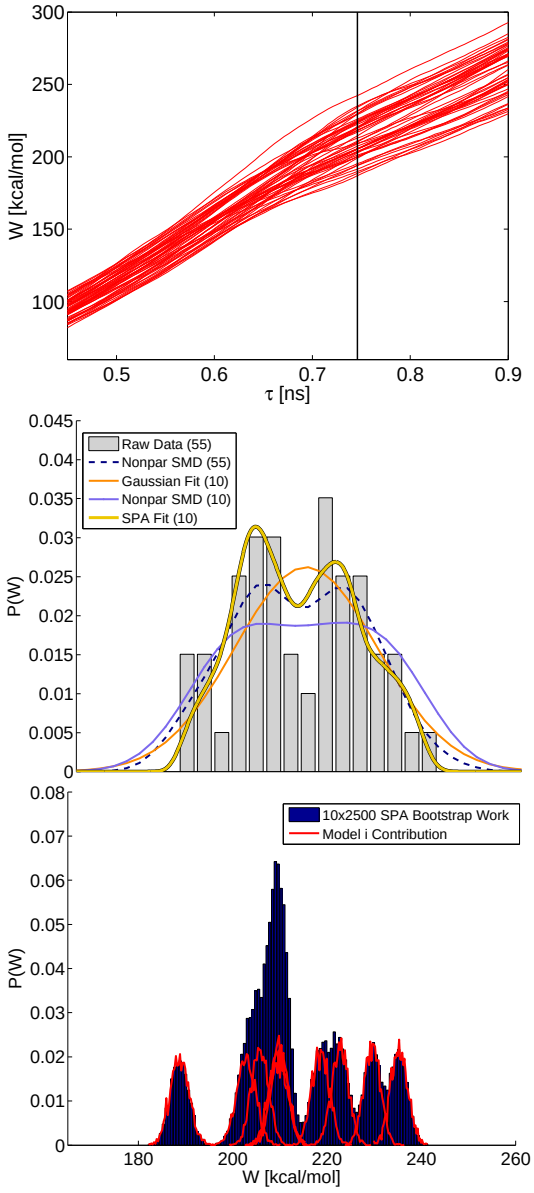


Figure 3: Approximating the nonequilibrium work distribution corresponding to SMD simulation of unfolding the I27 domain of titin. The top panel plots 55 nonequilibrium SMD work paths. The bars in the middle panel plot the normalized histogram obtained by analyzing the SMD data at the time point corresponding the vertical line in the top panel. The solid line curves in the middle panel denote various estimates of the population histogram (see text). The bottom panel displays the simulated SPA work paths used in case labeled "SPA fit" in the middle panel; the bars denote the histogram of all simulated work paths and the solid lines correspond to the contribution of the work density from each individual SPA model. Some SPA models predict effectively disjoint work histograms (all models used passed pathwise goodness-of-fit tests they were subjected to). .

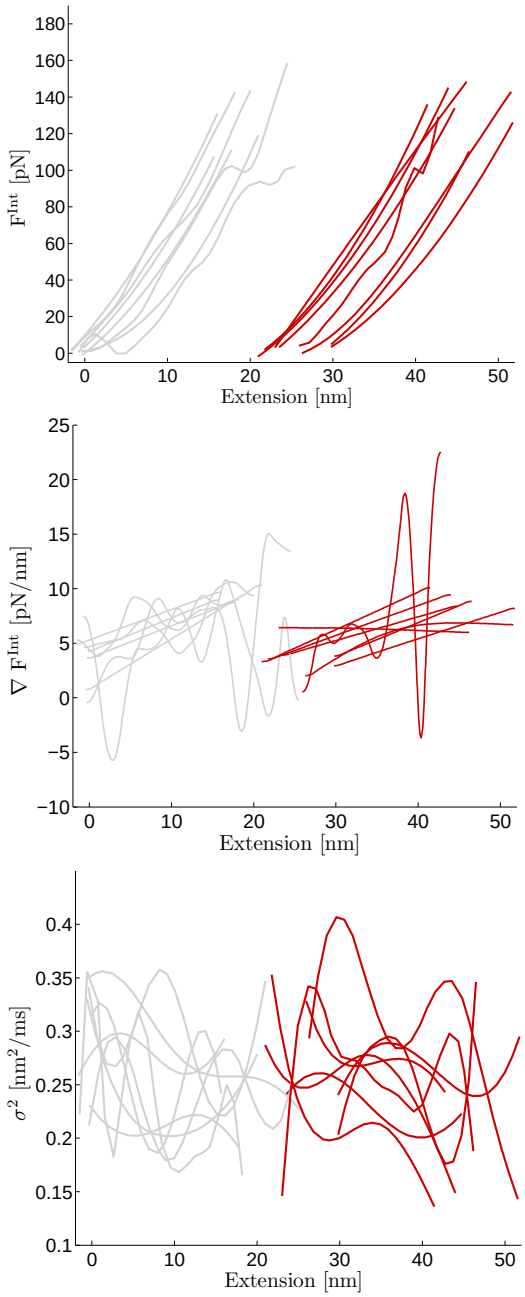


Figure 4: Results from calibrating SPA model using experimental AFM data. The force vs extension AFM data consisted of the typical sawtooth pattern.[6,65] Sample output from the AFM is included in Appendix Fig. 2. We used the second (light grey) and third (dark red) sawtooth in each force extension cycle for estimation purposes. The same I27 molecule remained attached to the tip for a total of 8 force extension cycles. The effective internal force, its gradient with respect to extension, and the effective diffusion coefficient are plotted as a function of extension.

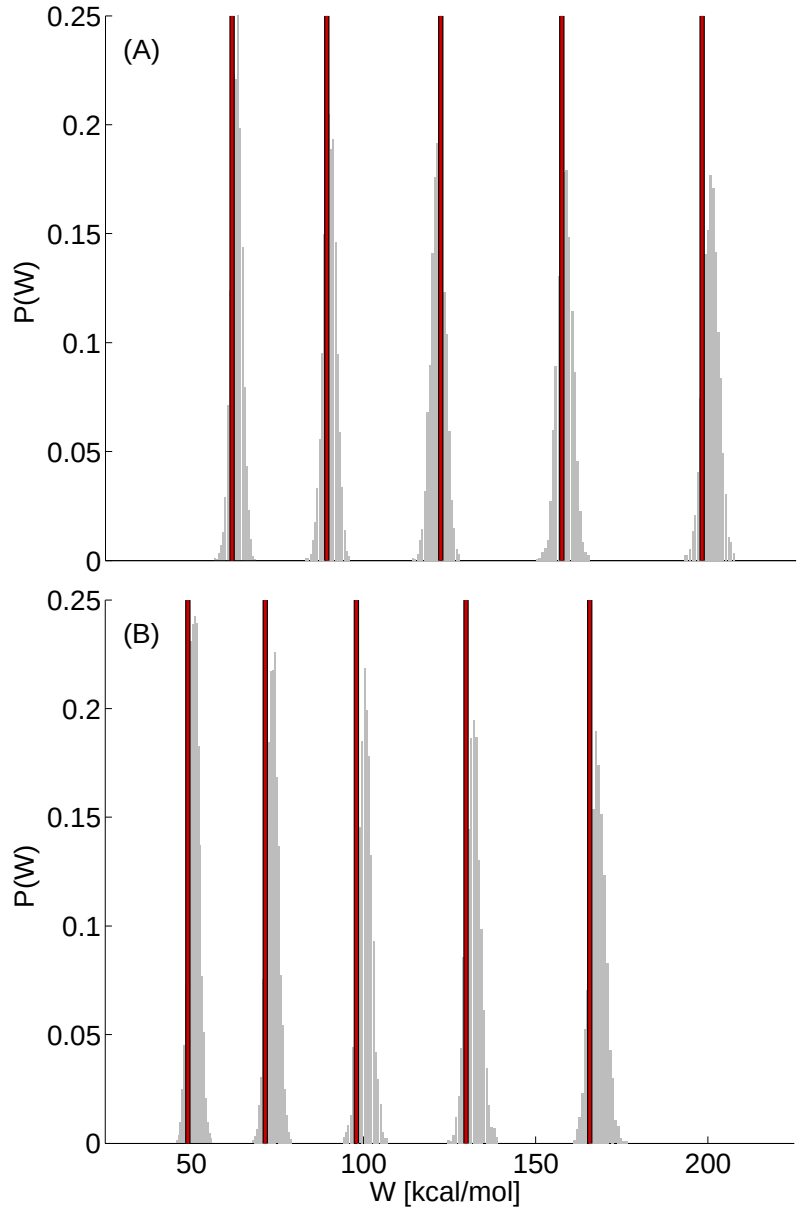


Figure 5: Simulation of SPA work histograms compared to experimental measurements. The solid vertical lines denote the nonequilibrium work value measured from the experimental AFM data corresponding to target extensions of $\lambda, of 10, 12.5, 15, 17.5, 20\mathrm{nm}$; the corresponding work values appear from left to right in each panel. The two panels correspond to two different experimental time series where the same molecule was retained on the AFM tip for multiple force extension cycles. We compare results obtained from two different unfolding cycles where the second sawtooth is used to calibrate two different SPA models. The resulting SPA models were each used to generate 2500 simulated work paths; the time evolving histogram at the λ extensions listed for the experiment are shown. This plot indicates that each individual SPA model can approximate the variation which can be attributed to thermal and instrument noise (see text).

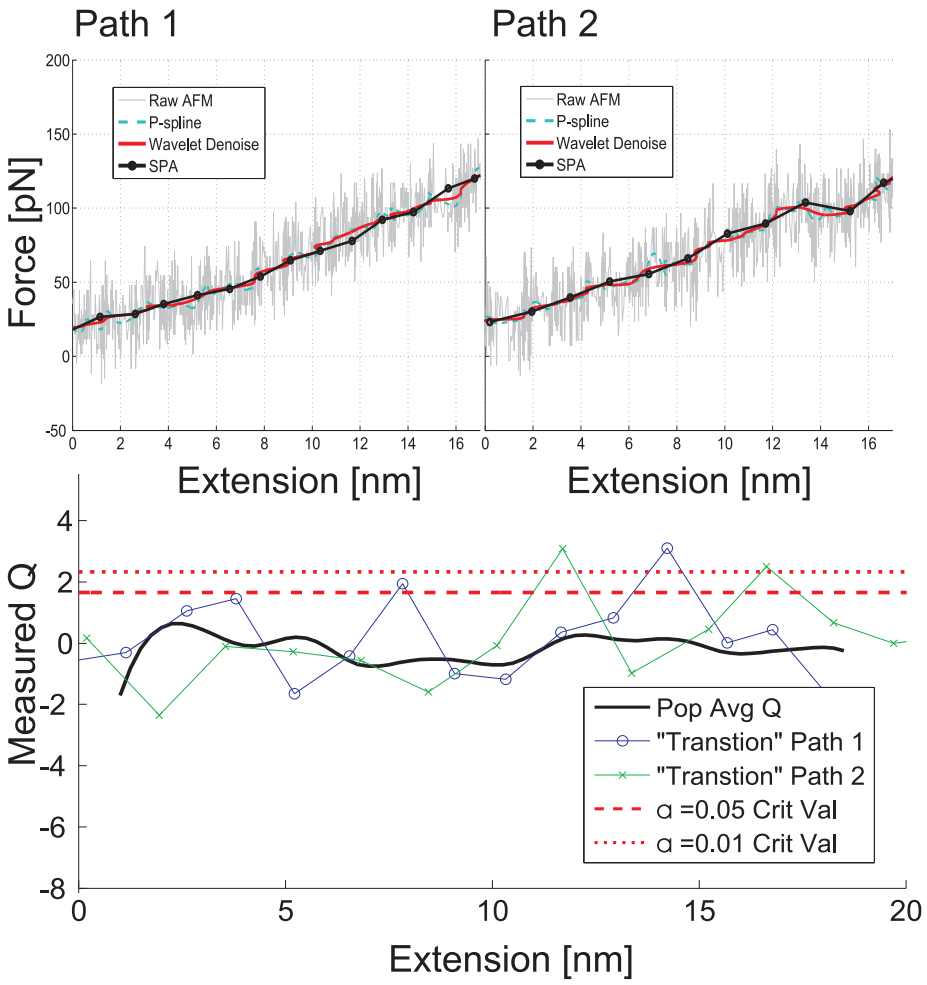


Figure 6: Hypothesis tests based on the fitted models and the observed nonstationary time series describing unfolding the I27 domain of titin via AFM. The resulting SPA models were subjected to the Q-test[28] and the population average (over the 8 SPA models) is plotted along with the critical values corresponding to type I rejection rates (α) of 0.01 and 0.05. Only two models were rejected using $\alpha = 0.01$, those rejected have the corresponding force extension displayed above as an inset. Interestingly both rejection occur near a known "transition".[46–49] Standard smoothing techniques, penalized spline smoothing[38] and wavelet denoising,[72] can readily detect the transition in "Path 2", but have a difficult time detecting the subtle transition associated with "Path 1" whereas the hypothesis test readily identifies this suspicious point (the rejection is caused mainly to a dramatic change in noise magnitude). The noisy curves in the panel correspond to "F^{Ext}" measured directly from the AFM; the standard smoothing techniques were applied to this same quantity. The SPA curves correspond to "F^{Int}".

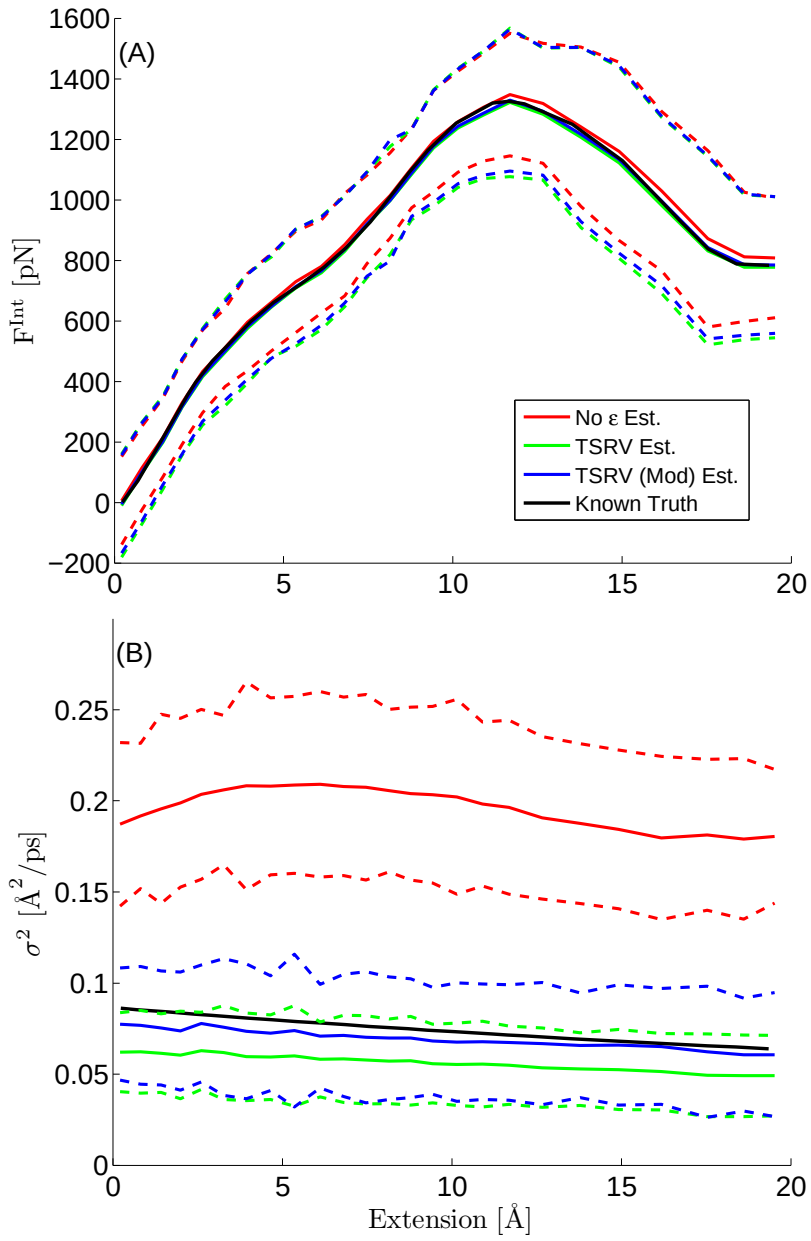
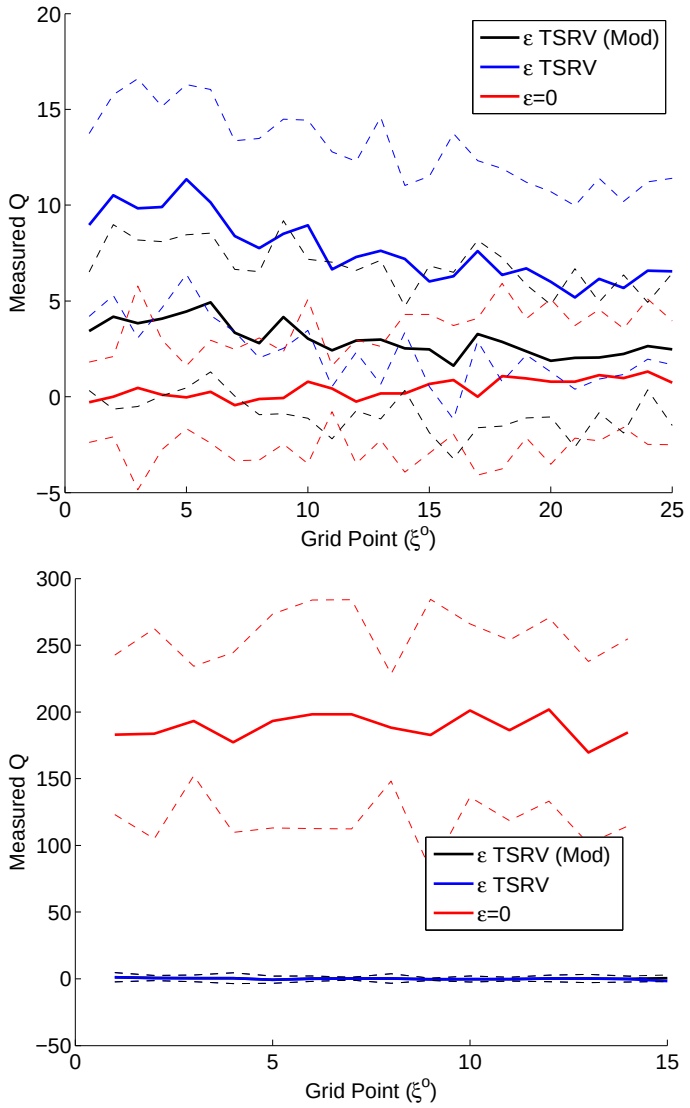
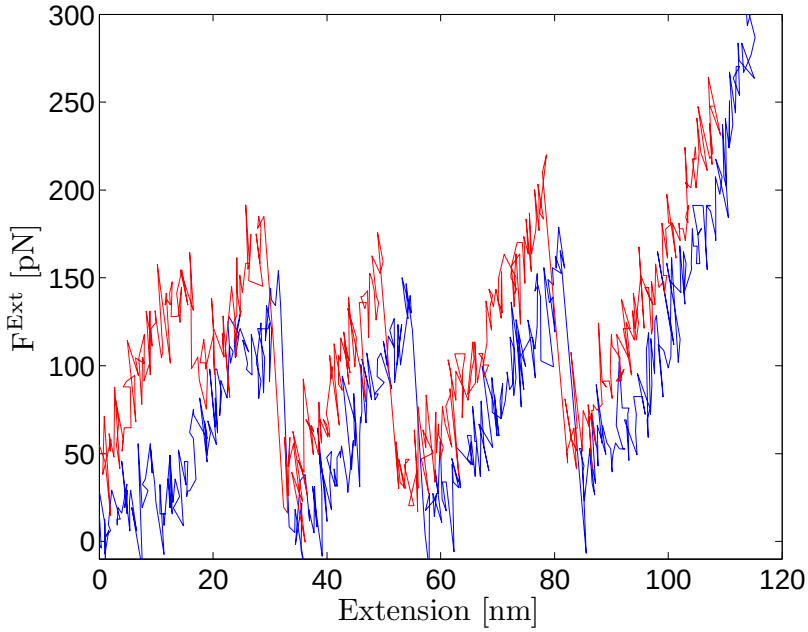


Figure 7: Top panel: Confidence bands for idealized model (genuine SDE). 500 Monte Carlo paths were generated using a *single* diffusion (with known drift and diffusion). Local parameter estimates were obtained and the average and standard deviation in each local windows were computed. A spline fit to the average represents the mean SPA function obtained. The standard deviation of the 500 paths in each window was also computed and a spline was fit to that data (the dotted curves represent $\mu(\xi) \pm 2\sigma(\xi)$ where the functions represent the spline fits to the corresponding quantities). Normally distributed measurement noise (of known variance) was added and various techniques for estimating the noise were attempted. Under these controlled conditions, it is demonstrated that the estimator we employed[35] does consistently estimate the local parameters. Note how ignoring the noise strongly influences the diffusion coefficient estimate.



Appendix. Fig. 1: Top panel: The Q-test statistic average using three noise estimation schemes applied to SMD simulation data. 1) Modified TSRV, 2) TSRV, 3) Ignoring the Measurement noise. Surprisingly the best model using this criterion is the $\epsilon = 0$ case. The fast-scale motions (the time series were sampled uniformly with 50 fs between observations) cannot be adequately be represented by "white measurement noise" in this system. This tests dictated the model we used for subsequent approximations of the SMD data. Bottom panel: The Q-test statistic average (over the 8 experimental curves) using three noise estimation schemes. Note that the test has no problem in rejecting (with virtually no uncertainty) the case where measurement noise is not taken into account. The test statistics under the asymptotic null is a standard normal distribution. Fig. 6 focuses on the region near zero.



Appendix. Fig. 2: Sample experimental force extension curve obtained when the AFM is used to unfold the I27 domain of titin.