JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

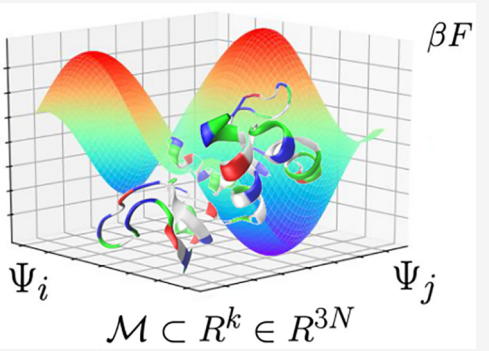
Learned Reconstruction of Protein Folding Trajectories from Noisy Single-Molecule Time Series

Maximilian Topel, Ayesha Ejaz, Allison Squires, and Andrew L. Ferguson*

Cite This: J. Chem. Theory Comput. 2023, 19, 4654−4667 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Single-molecule Förster resonance energy transfer (smFRET) is an experimental methodology to track the real-time dynamics of molecules using fluorescent probes to follow one or more intramolecular distances. These distances provide a low-dimensional representation of the full atomistic dynamics. Under mild technical conditions, Takens' Delay Embedding Theorem guarantees that the full three-dimensional atomistic dynamics of a system are diffeomorphic (i.e., related by a smooth and invertible transformation) to a time-delayed embedding of one or more scalar observables. Appealing to these theoretical guarantees, we employ manifold learning, artificial neural networks, and statistical mechanics to learn from molecular simulation training data the a priori unknown transformation between the atomic coordinates and delay-embedded intramolecular distances accessible to smFRET. This learned transformation may then be used to reconstruct atomistic coordinates from smFRET time series data. We term this approach Single-molecule TAkens Reconstruction (STAR). We have previously applied STAR to reconstruct molecular configurations of a $C_{24}H_{50}$ polymer chain and the mini-protein Chignolin with accuracies better than 0.2 nm from simulated smFRET data under noise free and high time resolution conditions. In the present work, we investigate the role of signal-to-noise ratio, data volume, and time resolution in simulated smFRET data to assess the performance of STAR under conditions more representative of experimental realities. We show that STAR can reconstruct the Chignolin and Villin mini-proteins to accuracies of 0.12 and 0.42 nm, respectively, and place bounds on these conditions for accurate reconstructions. These results demonstrate that it is possible to reconstruct dynamical trajectories of protein folding from time series in noisy, time binned, experimentally measurable observables and lay the foundations for the application of STAR to real experimental data.



INTRODUCTION

Probing single-molecule dynamics is crucial for understanding protein folding and misfolding.[1−5] The behavior of a single protein can be characterized by recording the positions of each of its N constituent atoms over time in a time series of R^{3N} dimensional vectors called a molecular trajectory. Computationally, molecular dynamics (MD) simulations can generate all-atom trajectories of molecules by solving Newton's equations of motion under an interatomic potential defined by an appropriate force field.[6] MD simulations are subject to numerical and finite precision errors[7] and can become prohibitively expensive for the simulation of large or slow folding proteins due to the short time steps required to propagate the simulations accurately. Experimentally, cryo-electron microscopy and X-ray crystallography can provide static reconstructions of protein structures at root-mean-square deviation (RMSD) accuracies of the position of each atom of ∼0.1 nm.[1−4] The structure of the proteins within these crystalline or vitrified states may not correspond to the native functional structure, and the static reconstructions cannot shed light on the nature of the dynamical fluctuations and transitions between metastable states.[8] Study of protein dynamics can be critical for understanding their function or dysfunction, with more than 50 known disorders related to misfolding of functional peptides and proteins.[5] Single-molecule fluorescence resonance energy transfer (smFRET) is a popular technique to experimentally follow the dynamics of biomolecular motions by optically recording energy transfer between fluorescent dyes grafted to particular sites on the molecule.[3,9] This technique permits an observer to track protein dynamics in a coarse-grained sense by following one or more simple geometrical parameters such as the intramolecular distance between the two dyes. No experimental techniques are currently available to track the single-molecule dynamical evolution of a protein with full atomistic resolution.

Takens' Delay Embedding Theorem is a result from dynamical systems theory, which asserts that time series data recording a single scalar observable of a dynamical system can,

Special Issue: Machine Learning for Molecular Simulation

Received: September 8, 2022
Published: January 26, 2023

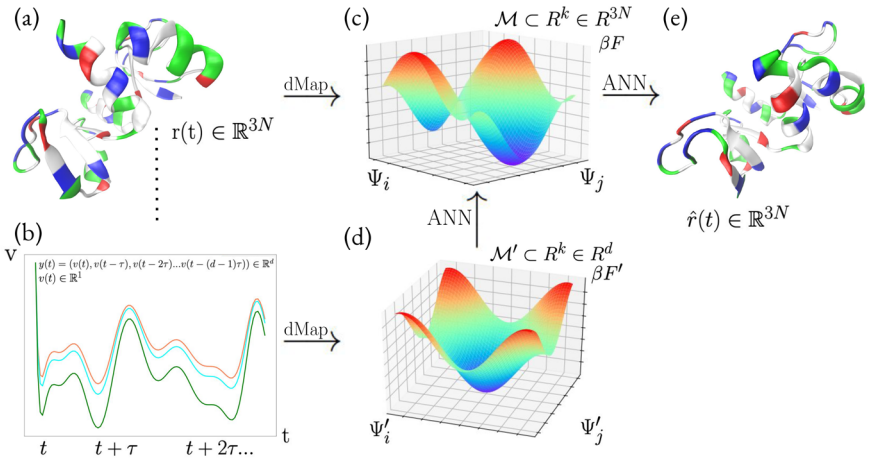


Figure 1. Cartoon of the layout of Single-Molecule TAkens Reconstruction (STAR). STAR takes time series in one or more coarse-grained observables of the molecular system such as an intramolecular distance furnished by smFRET $v(t) \in \mathbb{R}^1$ and reconstructs a molecular trajectory within the Cartesian coordinate space of N atoms $\hat{\mathbf{r}}(t) \in \mathbb{R}^{3N}$ through the three-step pathway b → d → c → e. Each panel corresponds to a different representation of a molecular trajectory, and the arrows between the panels correspond to learning tasks to convert one representation to another. (a,b) For the purposes of training the STAR pipeline, we collect all-atom MD trajectories recording the Cartesian coordinates of the N constituent atoms $\mathbf{r}(t) \in \mathbb{R}^{3N}$ and from these construct synthetic smFRET time series in a single intramolecular distance $v(t) \in \mathbb{R}^1$. Once the STAR pipeline is trained, we no longer need any additional MD simulation data, and the pipeline can operate on new synthetic (or experimental) smFRET data collected under similar conditions. Following the prescription of Takens' Theorem, we construct a delay embedding of the scalar time series as $\mathbf{y}(t) = [v(t), v(t-\tau), v(t-2\tau)...v(t-(d-1)\tau)] \in \mathbb{R}^d$, where τ is the delay time and d is the delay dimensionality. (c) The k-dimensional manifold containing the all-atom simulation trajectory $\mathcal{M} \subset \mathbb{R}^k \in \mathbb{R}^{3N}$ and spanned by the nonlinear collective variables $\{\Psi'_1, \Psi'_2, ..., \Psi'_k\}$. Collective couplings between the molecular degrees of freedom lead to an emergent low-dimensionality wherein $k \ll 3N$. We learn $\mathcal{M}$ from the MD simulation trajectory $\mathbf{r}(t)$ using diffusion maps unsupervised nonlinear manifold learning. (d) The k-dimensional manifold containing the smFRET delay embedding $\mathcal{M}' \subset \mathbb{R}^k \in \mathbb{R}^d$ spanned by the nonlinear collective variables $\{\Psi'_1, \Psi'_2, ..., \Psi'_k\}$. We learn $\mathcal{M}'$ from the delay embedding $\mathbf{y}(t)$ using diffusion maps. Takens' Theorem asserts $\mathcal{M}'$ is an image of $\mathcal{M}$, and, under some technical conditions on symmetries and periodicities, they are related by a diffeomorphism (i.e., a smooth and invertible transformation) Θ: $\mathcal{M}' \to \mathcal{M}$ that we learn by training a simple fully connected feedforward artificial neural network as a flexible and expressive nonlinear function approximator. (e) Reconstruction of the atomistic molecular trajectory within the Cartesian coordinate space of N atoms $\hat{\mathbf{r}} \in \mathbb{R}^{3N}$. We learn the molecular reconstruction from the manifold $\mathcal{M}$ using a second fully connected feedforward artificial neural network. All molecular renderings were made using Visual Molecular Dynamics.[29]

under some weak technical conditions, contain sufficient information to reconstruct the state of the full-dimensional system up to a diffeomorphic (i.e., smooth and invertible) transformation.[10−19] In the context of the present application, the theorem asserts that smFRET measurements can contain sufficient information to reconstruct the full three-dimensional atomic dynamics of the protein via an a priori unknown transformation.

In previous work, we showed that it is possible to learn this transformation from MD simulation training data and then use this learned model to reconstruct molecular trajectories from intramolecular distances accessible to smFRET.[20] We refer to this approach as Single-molecule TAkens' Reconstruction (STAR). Our application of STAR to computer simulations of a $C_{24}H_{50}$ polymer chain and the Chignolin mini-protein demonstrated RMSD reconstruction accuracies from simulated smFRET data of better than 0.2 nm. Although this work served as proof of principle of the technique, it was only validated for long idealized synthetic smFRET time series that were sampled at an extremely high time resolution and free of sampling noise.

The motivation of the present work is to test the ability of STAR to accurately reconstruct protein dynamics under experimentally realistic constraints on the smFRET time series associated with its temporal resolution, trajectory length, and presence of sampling noise. The temporal resolution is limited by the need to accumulate sufficient numbers of photons over a specified time bin to extract a reliable reading of the intramolecular distance between the fluorescent probes.[3] The length of a trajectory is primarily limited by photobleaching or photoblinking of the fluorophores.[21−23] Photobleaching is an irreversible chemical process whereby changes in a fluorophore's electronic structure render it permanently non-emissive.[24] Photoblinking is intermittent emissivity of a fluorophore that may result from temporary changes in electronic structure such as electric charge or isomerization, or from being trapped in a triplet state.[22,24] The Poisson statistics regulating photon emission combined with the ∼1 ns lifetime of the (emitting) excited singlet state[24] define this limitation. Sampling noise within the time series arises due to systematic effects and shot noise (i.e., Poisson statistics).[3,9] Theoretical guarantees on the applicability of Takens' theorem in the presence of limited data, low sampling resolution, and measurement noise are not available. Empirical testing of STAR under physical constraints on these three key variables can assess the degree to which STAR can accurately reconstruct molecular configurations from experimentally realistic smFRET trajectories. In this work, we lay the foundations for the application of STAR to real data by applying it to synthetic smFRET data generated from computational MD trajectories for which the ground truth atomic coordinates are exactly known and for which we can precisely control the length, resolution, and noise of the smFRET trajectories.

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

The primary outcome of this work is to show that STAR is capable of the accurate recovery of molecular configurations to accuracies of 0.12 and 0.42 nm for, respectively, the Chignolin and Villin fast-folding mini-proteins under conditions that bridge computationally tractable simulations to experimentally realistic FRET conditions. These accuracies are achieved using simulated smFRET trajectories of an aggregated length of approximately 0.7−3.3× the characteristic protein folding time, temporal resolutions of 1/120−1/280× the folding time, and sampling noise commensurate with collection of ∼10^5 photons per time bin. For Chignolin (τ_{fold} = 0.6 μs), this corresponds to MD trajectories of approximately 2 μs and time bins shorter than 5 ns, and for Villin (τ_{fold} = 2.8 μs) to trajectories of 2 μs and time bins shorter than 10 ns.[25] Extrapolation of these constraints on trajectory length, time resolution, and signal-to-noise ratio suggests that STAR may currently be deployed upon proteins with characteristic folding times exceeding approximately 100−1000 μs using state-of-the-art photon-by-photon single-molecule instruments with dye pairs like Cy3/Cy5 that have temporal resolutions of approximately 1−10 μs.[26−28]

The structure of this Article is as follows. In the Methods, we summarize the previously reported STAR pipeline and discuss application of noise, data restriction, and time resolution limitations on STAR input data. In the Results and Discussion, we present applications of STAR to MD simulations of the 10-residue artificial mini-protein Chignolin and 35-residue protein Villin under a variety of trajectory lengths, time resolutions, and signal-to-noise ratios. In the Conclusions, we discuss the impact of this work and scope for future development and applications of STAR.

METHODS

Principles of STAR. A schematic illustration of the STAR technique is presented in Figure 1. In this subsection, we provide a brief overview of the approach that we previously reported in ref 20. Details of the mathematical underpinnings and numerical implementations including algorithms, training protocols, and (hyper)parameters are provided in the Supporting Information. Template Jupyter notebooks implementing the STAR pipeline are available at https://github.com/Ferg-Lab/Limits-of-single-molecule-Takens-reconstruction-notebooks.

Takens' Delay Embedding Theorem is a proven result from dynamical systems theory asserting that, under some mild technical conditions, there exists a diffeomorphism (i.e., a smooth and invertible transformation) between the full dimensional state of the system and a so-called delay embedding of one or more coarse-grained observables.[10−19] In the context of protein folding, this theorem implies that smFRET time series may contain sufficient information to reconstruct the all-atom configuration of the molecule. It is the fundamental principle of STAR to learn this a priori unknown transformation from MD simulation training data and then apply the learned transformation to "upgrade" smFRET time series into trajectories of a molecule tracking its Cartesian coordinates. In the illustration of the STAR pipeline in Figure 1, each panel corresponds to a different representation of a molecular trajectory, and the arrows between the panels correspond to learning tasks to convert one representation to another.

The STAR algorithm employs a combination of manifold and nonlinear learning tools to convert a smFRET time series $v(t) \in \mathbb{R}^1$ (Figure 1b) through a three-step pathway (b → d → c → e) to a reconstruction of the atomistic molecular trajectory within the Cartesian coordinate space of N atoms $\hat{\mathbf{r}} \in \mathbb{R}^{3N}$ (Figure 1e). Depending on the application and data quality, one may choose to reconstruct all atoms in the molecule or just a subset, for example, the heavy or backbone atoms. In principle, Takens' Theorem asserts that one could learn this transformation in a single step (i.e., b → e). In practice, we make use of the generically low effective dimensionality of molecular systems[30,31] to recover a $k \ll 3N$ dimensional manifold containing the smFRET trajectory $\mathcal{M}'$ (Figure 1d) and learn the transformation to the analogous k-dimensional manifold containing the all-atom trajectory $\mathcal{M}$ (Figure 1c). This is beneficial in defining a lower-dimensional and better posed mapping that must be learned from the smFRET to atomistic data. It also furnishes an informative and interpretable k-dimensional free energy landscape supported on the manifold $\mathcal{M}$ that provides a wealth of information on the metastable states and transition pathways of the molecular system.

The existence of the transformation (d → c) is guaranteed by Takens' Theorem, but the expression is initially unknown and must be learned from training data. We perform all-atom molecular dynamics simulations to furnish both an all-atom molecular trajectory $r \in \mathbb{R}^{3N}$ (Figure 1a) and a synthetic smFRET time series in a single intramolecular distance $v(t) \in \mathbb{R}^1$ (Figure 1b). In this work, we take v to be the distance between two selected hypothetical fluorophore attachment positions. A more realistic approximation would explicitly model the FRET fluorophores within the MD simulation.[4] These trajectories constitute the training data necessary to construct the k-dimensional embedding of the all-atom simulation trajectory (a → c), construct the k-dimensional embedding of the synthetic smFRET trajectory (b → d), learn the transformation between them (d → c), and learn the reconstruction of the molecular configuration from the low-dimensional all-atom manifold (c → e). Once all steps in the pipeline are learned from the training data, STAR can be applied to reconstruct the molecular trajectories from new synthetic (or real) smFRET trajectories collected under the same conditions without conducting additional molecular simulations via the pathway b → d → c → e.

The primary focus of the present work is to test the application of STAR to smFRET trajectories with constraints on the number of individual smFRET measurements within the trajectory (i.e., data volume, Dv), the temporal resolution of the smFRET trajectory (i.e., bin size, λ), and the presence of sampling noise that is controlled by the brightness of the FRET donor fluorophore (I_D). We train STAR models on computational training data with different values of (Dv, λ, I_D) and then test the models on novel synthetic smFRET data to assess the reconstruction accuracy. Importantly, we intentionally test our approach on synthetic smFRET data for which we can explicitly control Dv, λ, and I_D, and for which we possess the ground truth atomistic molecular simulation trajectories against which we can test the performance of our STAR reconstruction. Having defined the regimes of these three critical parameters within which STAR is determined to perform well for two fast-folding mini-proteins computationally amenable to simulation, we then prospectively identify protein systems and experimentally realistic FRET conditions capable of meeting these constraints and to which future

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

applications of STAR to real experimental smFRET data may be successful.

Training Data: Molecular Dynamics Simulations $\mathbf{r}(t)$ *and Synthetic smFRET Time Series* $\nu(t)$*).* We use all-atom MD simulations to furnish the training data necessary to learn the transformations within the STAR pipeline. The MD trajectory provides a time series of coordinates $\mathbf{r}(t) \in \mathbb{R}^{3N}$ (Figure 1a) from which we generate the synthetic smFRET trajectory corresponding to a scalar time series of an intramolecular distance between two hypothetical FRET fluorophores $\nu(t) \in \mathbb{R}^1$ (Figure 1b). In this work, we assume the fluorophores to be placed at the beginning and end of the linear proteins such that $\nu(t)$ corresponds to the molecular head-to-tail distance. We train STAR models on the synthetic smFRET data extracted from the MD trajectory at particular trajectory lengths $D\nu$, time bin resolutions λ, and signal-to-noise ratios I_D. Models are trained over subsamples of the first 80% of each MD trajectory (vide infra), and the remaining 20% are held out as a test partition. To ensure good configurational diversity in the training data, we bin the training data into 10 equally spaced bins in the molecular head-to-tail distance and randomly sample $D\nu/10$ configurations from each bin. As detailed below, Takens' Theorem requires access to the immediate time history of each point, and so we also collect the configurations preceding each selected point as far back as required, which, for the two protein systems considered in this work, lies in the range 2−220 ns. The training data may therefore be conceived as an ensemble of short contiguous trajectories of the molecule distributed over a variety of head-to-tail distances to ensure good sampling of its full configurational space. Hyperparameters are tuned for each component of the training pipeline using noiseless data at high time resolution and large data volumes and then applied to each ($D\nu$, λ, I_D) triplet. We note that we train a single STAR model at a particular choice of parameters and apply it transferably to all other parameter regimes. An alternative strategy would be to train independent STAR models for each choice of these three parameters to better mimic within the training data the conditions of the testing set during model deployment. In our experience, this did not lead to significant improvements in performance.

Time Resolution, λ*.* The operating principle of smFRET is to label two locations on a molecule with a pair of fluorophores whose absorption and emissions are distinct but overlapping, so that they may nonradiatively exchange energy upon excitation. This exchange of energy is governed by the relative geometry and spectra of the two fluorophores so that energy flows from the higher-energy donor fluorophore to the lower-energy acceptor at distances on the scale of 2−10 nm. The experimentally observed relative fluorescence of the donor and acceptor under donor-only excitation reports the efficiency of energy transfer and can be used to directly estimate the distance separating the two fluorophores as a "molecular ruler".[3,32] Photon emission statistics are Poisson distributed, so FRET efficiencies are computed over finite time bins to mitigate the effects of noise. By recording the number of donor and acceptor photons over the course of a single time bin, a single intramolecular distance is computed corresponding to an estimate of the average donor−acceptor distance over the time bin. Bins of $1-10^5$ μs have been used in practice,[3,27,28] with larger bins possessing improved signal-to-noise ratios but sacrificing temporal resolution. To mimic time binning in our synthetic smFRET trajectories, we bin the $\nu(t)$ time series into a sequence of bins of length λ and report for each bin the mean value of $\nu(t)$ recorded between time t and $(t + \lambda)$ as $\nu(t \rightarrow t + \lambda) = (1/\lambda) \int_t^{t+\lambda} \nu(t)\, \mathrm{d}t$, where we approximate the integral as a discrete sum at the resolution of the synthetic time series. We perform an analogous operation for the corresponding MD training data $\mathbf{r}(t)$ wherein the Cartesian coordinates of each atom are averaged over the time bin λ. Typical values of λ for smFRET are on the order of milliseconds for confocal FRET setups,[3,33] but can be pushed down to 1−10 μs for state-of-the-art photon-by-photon single-molecule instruments.[26−28] Training STAR models for different values of λ enables us to assess how the temporal resolution affects the reconstruction accuracy of the trained pipeline. In this work, we consider $\lambda = \{0.2, 1, 2, 5, 10\}$ ns for Chignolin and $\lambda = \{0.2, 1, 2, 10, 20\}$ ns for Villin as appropriate to capture the dynamics of these ultrafast-folding mini-proteins with characteristic folding times of τ_{fold} = 0.6 and 2.8 μs, respectively.[25] These systems were selected for this work as sufficiently small and fast-folding to be amenable to good sampling with unbiased MD simulations. In the analysis of our results, we focus on the reconstruction accuracy as a function of the ratio λ/τ_{fold}, which enables us to extrapolate our predictions to larger, slower-folding proteins that are much more challenging to sample using MD but are more readily studied by experimental smFRET.

Trajectory Length, $D\nu$*.* Training of the STAR pipeline requires MD training data for the protein of interest to learn the mappings denoted by the arrows in Figure 1. MD simulations are typically limited to time scales of microseconds to milliseconds, even on high performance and bespoke computational hardware.[25,34] To assess the influence of training data volume upon the reconstruction accuracy of the trained STAR model, we consider a variety of training data volumes defined by the number of synthetic smFRET distance measurements $D\nu$ within the training ensemble. In this work, we select $D\nu = \{10^3, 10^4, 2 \times 10^4, 4 \times 10^4\}$ observations using our training data selection criteria. If less than $D\nu/10$ points are available in each head-to-tail decile, the total number of training points selected may be slightly less than $D\nu$. As a point of comparison, experimental smFRET trajectories can vary in length from milliseconds to tens of seconds[3,33,35] and employ temporal resolutions of several microseconds to milliseconds,[26−28] meaning that an experimental trajectory can contain 10^2-10^7 individual distance measurements.

Noise, I_D*.* The signal-to-noise ratios in smFRET time series are largely controlled by the intensity of the measured fluorescence: brighter fluorophores produce higher signal-to-noise ratios, whereas dimmer ones suffer more from the effects of noise. There are a number of sources of noise in smFRET emission measurements stemming from thermal fluctuations, biases in dye orientations or spatial distributions relative to the labeling sites, fluctuations in dye photophysical properties such as quantum yield due to local chemical environments, and fast blinking or other sub-time resolution kinetics.[3] Because of the discrete nature of photon counts, the noise in FRET measurements can be modeled as shot or Poisson noise.[3,36] Application of propagation of uncertainties to the relationship between FRET efficiency and donor−acceptor distance ν allows us to derive a closed-form model for the relative uncertainty σ_ν/ν in this distance (see derivation in the Appendix):

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

$$\left(\frac{\sigma_\nu}{\nu}\right)^2 = \frac{1 + \left(\frac{R_0}{\nu}\right)^6}{36 I_D(\lambda)}\left[1 + \left(\frac{\nu}{R_0}\right)^6\left(\frac{\phi_D}{\phi_A}\right)\right] \quad (1)$$

where ϕ_D is the quantum yield of the donor, ϕ_A is the quantum yield of the acceptor, R_0 is the characteristic FRET radius for the donor−acceptor pair, and $I_D(\lambda)$ is the intensity of the donor channel (i.e., number of photons collected over the time bin λ) under direct excitation by the laser without acceptor present in the system. We then sample from this noise distribution to artificially corrupt the idealized bin-averaged distances extracted from our MD simulation trajectories:

$$\nu(t) \leftarrow \nu(t) + \mathcal{N}(0, \sigma_\nu^2) \quad (2)$$

where $\mathcal{N}(0, \sigma_\nu)$ is a random Gaussian variable with mean zero and variance σ_ν^2. A Gaussian noise model is a good continuous approximation for the underlying discrete Poisson statistics for sufficiently large photon counts, which, as described below, is the regime in which we operate.

In this work, we assume the FRET fluorophores to have ideal quantum efficiencies $\phi_A = \phi_D = 1$ and a typical FRET radius of $R_0 = 5$ nm. The uncertainty in the donor−acceptor distance ν is then fully specified by the intrinsic brightness of the donor dye and the time bin λ over which photons are collected, which together define I_D. Brighter donors produce more photons in the donor and acceptor channels for a given distance ν, and larger time bins increase the absolute number of photons collected in the detector. Both of these effects therefore improve the signal-to-noise ratio in the calculated donor−acceptor distance. By deploying our trained STAR models on synthetic smFRET time series with different I_D values, we can quantify how the reconstruction accuracy depends on the signal-to-noise ratio. Experimental setups employing sophisticated burst-search algorithms[36] can track on the order of 1000−10 000 photon bursts per measurement.[33] In this work, we consider $I_D = \{10^3, 10^4, 10^5, 10^6, \infty\}$, where $I_D = \infty$ corresponds to the hypothetical limit of an infinitely bright dye and noiseless conditions.

Molecular Dynamics Simulations. The MD simulations of the mini-protein Chignolin and the actin-binding protein Villin conducted by D.E. Shaw Research were used for STAR training and validation.[25] Simulations were performed on the supercomputer Anton under the CHARMM22* force field[37] with a compatible modified TIP3P water model.[38] Lys, Arg, Asp, and Glu residues and N- and C-termini were simulated in their charged states.[25] Each system was equilibrated in the NPT ensemble using the Desmond software package on a PC cluster, and equilibrium folding simulations were performed on the Anton supercomputer in the NVT ensemble.[25,39] The initial structure for the NVT ensemble simulation was chosen as the frame with the volume closest to the average volume. The system was then coupled to a Nosé-Hover thermostat with a 1 ps relaxation time.[40,41] Equations of motion were integrated at a 2.5 fs time step, and frames were recorded every 200 ps.[25] All simulations were run in the NVT ensemble using a Lennard-Jones potential with a 0.95 nm cutoff distance for short ranged electrostatics and the Gaussian Split Ewald method for long distance electrostatics with a 32 × 32 × 32 cubic grid.[25,42]

Chignolin. The 10-residue 166 atom mini-protein Chignolin peptide (PDB ID: 1UAO)[43] was solvated in a cubic box with 4 nm sides containing approximately 1900 water molecules.[38] The (−2) peptide charge was neutralized with two Na^+ ions. A 107 μs MD simulation in the NVT ensemble was conducted.[25] Training data were subsampled from the first 80% of the trajectory, and the last 20% was used for testing. The characteristic folding time of Chignolin is $\tau_{fold} = 0.6$ μs.[25]

Villin. The 35-residue 577 atom Villin polypeptide with protonated HIS residue (PDB ID: 2F4K)[44] was solvated in 40 mM NaCl in a cube with 5.4 nm sides containing approximately 4400 water molecules. A 125.6 μs MD simulation in the NVT ensemble was conducted.[25] Training data were subsampled from the first 80% of the trajectory, and the last 20% was used for testing. The characteristic folding time of Villin is $\tau_{fold} = 2.8$ μs.[25]

RESULTS AND DISCUSSION

We now report the results of our parametric study of STAR reconstruction accuracy as a function of $(D\nu, \lambda, I_D)$ for each of the two proteins Chignolin and Villin. Our general conclusions are that for data volumes of $D\nu \geq 10^4$ observations, time resolutions of $\lambda \leq 5$ ns for Chignolin and $\lambda \leq 10$ ns for Villin ($\lambda/\tau_{fold} \leq 8.3 \times 10^{-3}$ and 3.5×10^{-3}, respectively), and signal-to-noise ratios corresponding to $I_D \geq 10^5$ photons per bin, we are able to reconstruct Chignolin and Villin structures with heavy atom RMSD accuracies of 0.1−0.4 nm. These reconstruction fidelities lie in the same range as static reconstruction techniques such as X-ray crystallography and cryo-electron microscopy.[1,2]

Chignolin. *STAR Training.* Calibration of the STAR hyperparameters is performed over configurations harvested from the first 80% of the 107 μs simulation trajectory (85.6 μs comprising 427 797 frames at 0.2 ns intervals) employing $D\nu = 40\,000$, $\lambda = 0.2$ ns, and $I_D = \infty$. We choose to reconstruct the $N = 93$ heavy atoms producing a MD trajectory $\mathbf{r}(t) \in \mathbb{R}^{279}$. Diffusion maps were used to extract the manifold $\mathcal{M}$ from $\mathbf{r}(t)$, constructed with a kernel bandwidth $\varepsilon = e^{-3}$ nm. A gap in the eigenvalue spectrum informed a 2D embedding spanned by $\{\Psi_1, \Psi_2\}$ (Figure S1). Delay vectors $\mathbf{y}(t)$ were constructed from the time series in the head-to-tail distance $\nu(t)$ computed between the terminal heavy atoms. A delay dimensionality of $d = 11$ and delay time of $\tau = \lambda$ were employed. Diffusion maps used to extract the manifold $\mathcal{M}'$ from $\mathbf{y}(t)$ were constructed with a kernel bandwidth $\varepsilon = 1$ nm to produce 2D embeddings into $\{\Psi_1', \Psi_2'\}$. The map from $\mathcal{M}'$ to $\mathcal{M}$ was parametrized by a 2−10−10−10−10−2 ANN trained over 100 epochs with a batch size of 500 and a learning rate of 10^{-4} using the Adam algorithm.[45] The map from $\mathcal{M}$ to $\hat{\mathbf{r}}$ was parametrized by a 2−4−189−374−558−279 ANN trained over 120 epochs with a batch size of 400 and a learning rate of 10^{-5} using the Adam algorithm.[45]

STAR Deployment and RMSD Dependence on $(D\nu, \lambda, I_D)$. Using the hyperparameters detailed in the previous subsection, we trained 100 independent STAR models at each combination of $\lambda = \{0.2, 1, 2, 5, 10\}$ ns, $D\nu = \{10^3, 10^4, 2 \times 10^4, 4 \times 10^4\}$, and $I_D = \{10^3, 10^4, 10^5, 10^6, \infty\}$. The heavy atom RMSD reconstruction accuracies of each model on the 20% hold-out test partition (21.4 μs comprising 106 946 frames at 0.2 ns intervals) are illustrated in Figure 2. In general, we observe quite good RMSD reconstruction accuracies between 0.10 and 0.25 nm for all $(D\nu, \lambda, I_D)$ triplets considered with the exception of $I_D = 10^3$ triplets where the signal-to-noise ratio was too low to converge training of an ANN to learn the $\mathcal{M}' \rightarrow \mathcal{M}$ diffeomorphism and the results

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

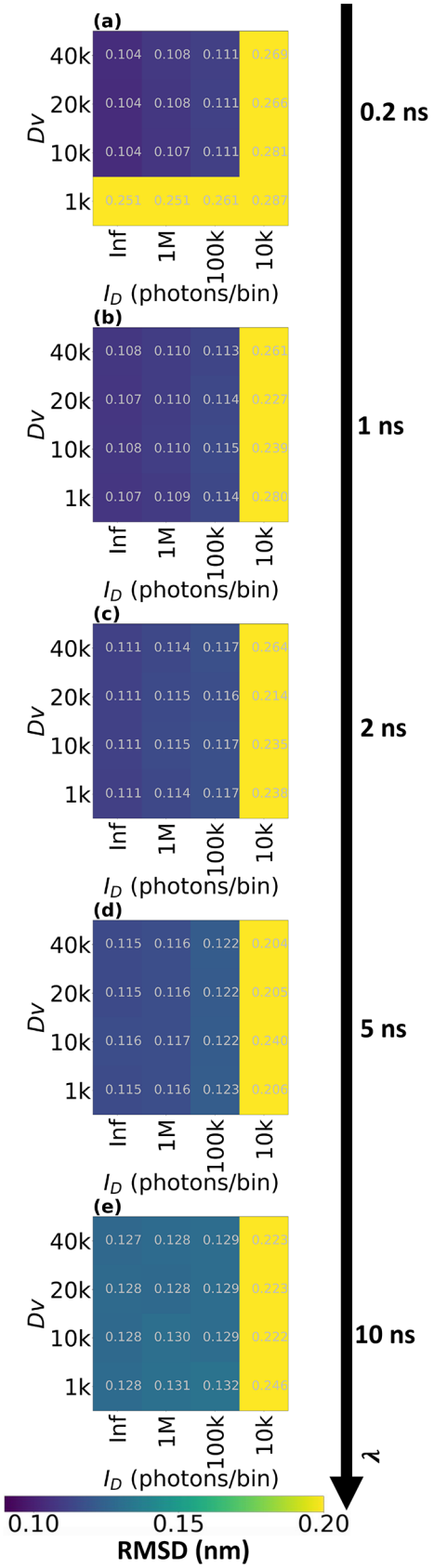


Figure 2. Heavy atom RMSD reconstruction accuracies for Chignolin as a function of training data volumes Dv, time bin resolution λ, and signal-to-noise ratio I_D.

are not reported. We must regard this accuracy with two important caveats. First, Chignolin is a very small protein comprising only 10 residues, so the per residue RMSD lies in the range 0.010−0.025 nm. Second, we can achieve a baseline RMSD accuracy on the test data of 0.155 nm by approximating each configuration by a single configuration from the test trajectory that results in the lowest overall RMSD averaged over all other configurations. The predictive power of the trained STAR models should therefore be viewed in light of improvements beyond this baseline accuracy.

The most obvious trend in Figure 2 is the influence of I_D, for which we see an approximate halving in the RMSD reconstruction error from ∼0.25 to ∼0.12 nm in moving from $I_D = 10^4$ to 10^5. Further increasing I_D to 10^6 or ∞ leads to relatively minor improvements in accuracy on the order of 3%. We were unable to learn a mapping between $\mathcal{M}'$ and $\mathcal{M}$ at $I_D = 10^3$. These trends indicate a clear floor on the signal-to-noise ratio necessary for reliable and accurate training of a STAR pipeline. We find reconstruction accuracy to be quite insensitive to data volume Dv. For a particular choice of I_D and λ, modulating Dv over the full range of 10^3 to 4×10^4 observations leads to only 1% changes in the RMSD accuracy except for $\lambda = 0.2$ ns triplets. One exception to this trend occurs for $\{Dv = 10^3, \lambda = 0.2 \text{ ns}, I_D \geq 10^4\}$ (Figure 2a), where we observe substantially poorer reconstruction accuracies than at larger data volumes Dv (Figure 2a) and lower time resolutions λ (Figure 2b−e). We hypothesize that the high temporal resolution of the training data at $\lambda = 0.2$ ns means that larger data volumes are required for STAR to effectively learn the system dynamics than at larger λ values for two interlinked reasons. First, at fixed Dv, a larger λ corresponds to a sampling of a longer time $Dv \times \lambda = Dv \times \tau$ that samples a longer period of the evolution of the system. Second, a larger λ coarse-grains over the sub-λ temporal fluctuations that may improve the reconstruction accuracy by attenuating the high frequency system dynamics. Finally, the reconstruction error is moderately sensitive to the time resolution λ. For fixed I_D and Dv, increasing λ over the range 0.2−10 ns ($\lambda/\tau_{fold} = 3.3 \times 10^{-4}$ to 1.6×10^{-2}) leads to a progressive 20−30% degradation in the RMSD.

Taken together, our analysis reveals that for training data volumes $Dv \geq 10^4$ samples, photon counts per bin of $I_D \geq 10^5$, and time resolutions of $\lambda \leq 5$ ns ($\lambda/\tau_{fold} \leq 8.3 \times 10^{-3}$), we are able to achieve RMSD reconstruction accuracies of 0.12 nm or better, which is 23% better than the RMSD = 0.155 nm baseline. Conversely, in low signal-to-noise ratio (i.e., $I_D \leq 10^4$) regimes, the reconstruction accuracy is worse than this baseline, or the training data are too noisy to permit convergence of a trained model.

Analysis of $\mathcal{M}$. For signal-to-noise ratios produced by $I_D \geq 10^5$ photons per bin, the dependence of the RMSD reconstruction accuracy upon data volume Dv and donor dye brightness I_D is relatively weak, and the primary determinant of reconstruction accuracy is the temporal time binning λ. To understand how the observed trends in the reconstruction accuracy as a function of λ can be attributed to the all-atom manifold $\mathcal{M}$, we present in Figure 3a the manifold $\mathcal{M}$ recovered from the MD trajectory $\mathbf{r}(t)$ at each λ value for $Dv = 4 \times 10^4$ samples and $I_D = 10^5$ photons per bin. Taking the highest resolution manifold at $\lambda = 0.2$ ns as the ground truth (Figure 3a-A), we observe two metastable free energy minima at ($\Psi_1 \approx 1.6 \times 10^{-6}$, $\Psi_2 \approx 9.5 \times 10^{-5}$) and ($\Psi_1 \approx 1.6 \times 10^{-4}$, $\Psi_2 \approx -8.5 \times 10^{-5}$) corresponding, respectively, to the native

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

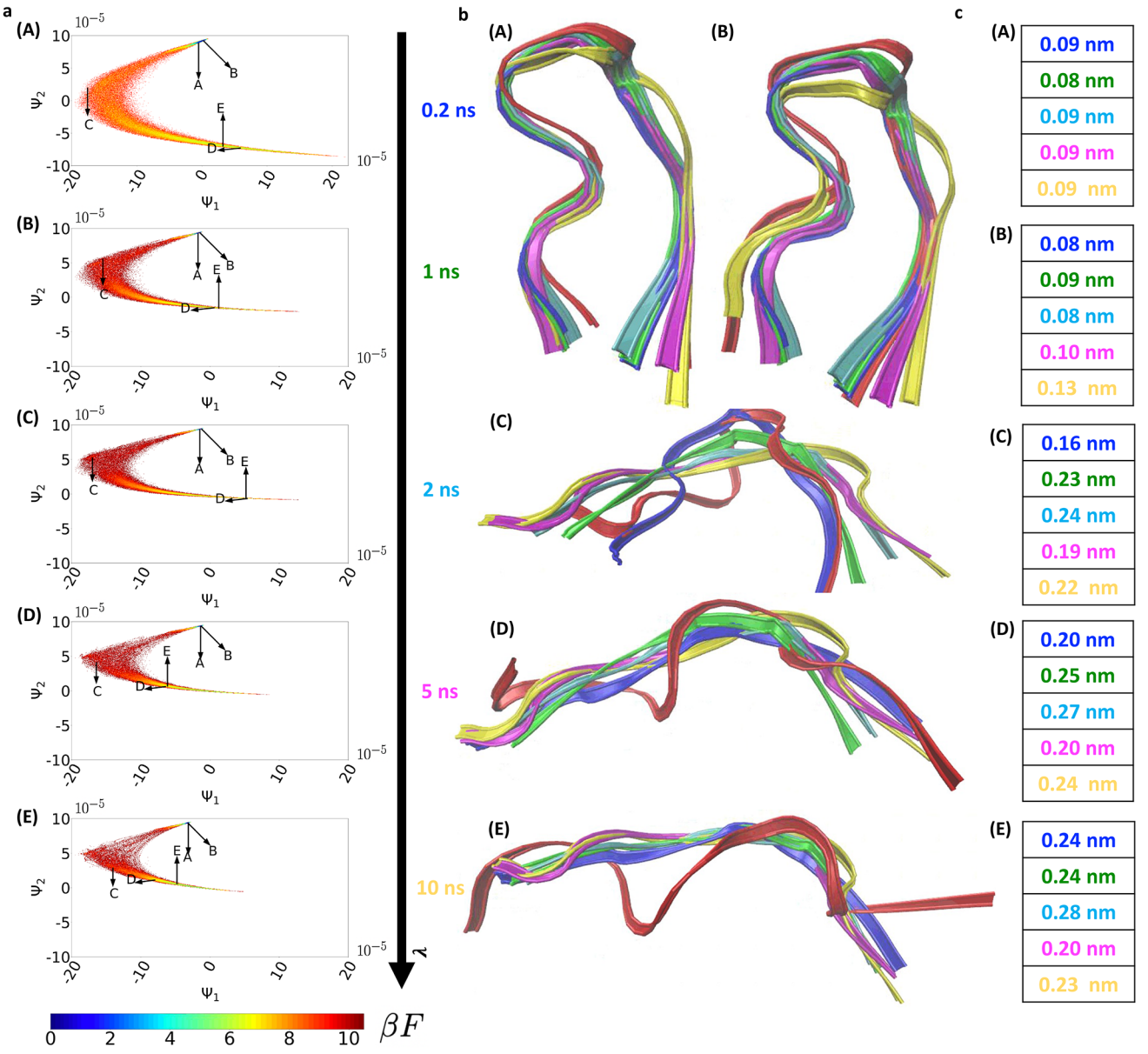


Figure 3. Manifolds $\mathcal{M}$ and position-dependent heavy atom RMSD reconstruction accuracy for Chignolin as a function of time bin resolution λ. (a) All-atom manifolds $\mathcal{M}$ within STAR pipelines trained at $Dv = 4 \times 10^4$ samples and $I_D = 10^5$ for each value of $\lambda = \{0.2, 1, 2, 5, 10\}$ ns. Manifolds are represented as scatter plots showing the embedding of each of the $Dv = 4 \times 10^4$ MD configurations $\mathbf{r}$ into the two diffusion map eigenvectors $\{\Psi_1, \Psi_2\}$ spanning each manifold. Each point is colored by the associated free energy $\beta F(\Psi_1, \Psi_2)$ of that point computed by collecting histograms over the empirical probability distribution at a bin size of $\{\Delta\Psi_1 = 5 \times 10^{-7}, \Delta\Psi_2 = 5 \times 10^{-7}\}$. The arbitrary zero of free energy in each panel is specified by subtracting the computed energy of the minimum free energy point on the landscape from all other points. Five representative configurations A−E are selected from the MD test trajectory and projected onto each manifold: A and B reside in the folded macrostate, C in the transition region, and D and E in the unfolded ensemble. (b) Visualizations of the reconstructions $\hat{\mathbf{r}}$ of each configuration A−E at each time resolution $\lambda = 0.2$ ns (blue), 1 ns (green), 2 ns (cyan), 5 ns (magenta), and 10 ns (yellow). For visual clarity, configurations are represented as ribbons tracing the backbone of the protein and are superposed upon the true configuration extracted from the MD test trajectory $\mathbf{r}$ (red). (c) Heavy atom RMSD reconstruction corresponding to each state is listed next to each image employing the same color-coding as the reconstructions. All molecular visualizations are constructed using VMD.[29]

and unfolded metastable macrostates. The high free energy region connecting them corresponds to the transition paths linking these two states. Prior work has reported two-state or three-state free energy landscapes corresponding to folded, unfolded, and misfolded states,[46,47] with variations attributed to differences in the simulated molecules and simulation protocols. Our results are in good agreement with those of Schaffer et al.,[47] who report a deep free energy well separated from a second metastable extended state by a relatively high free energy transition state. Taking the highest resolution manifold at $\lambda = 0.2$ ns as the ground truth, we observe that the primary impact of increasing bin size is an attenuation in the sampling of the transition region. This results in more sparse sampling of this region and an elevation in the apparent free energy of the transition pathways. This can be understood as a consequence of the enlarged window over which atomic positions are averaged with increasing λ that reduces the influence of transition states that are only fleetingly occupied

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

relative to the comparatively long-lived metastable states. Conversely, the relative location of the metastable free energy basins is insensitive to the value of λ but becomes more smeared out and loses definition with increasing λ due to increased averaging over molecular configurations.

Analysis of changes in the free energy landscape as a function of λ leads us to hypothesize that configurations within the transition region are likely to be more poorly reconstructed than those within the metastable basins due to the relatively poorer sampling of this region that is exacerbated at large λ. To test this hypothesis, we select from our validation trajectory five representative configurations A−E and project them onto each manifold $\mathcal{M}$ in Figure 3a. Configurations A and B reside within the folded macrostate, C within the transition region, and D and E within the unfolded ensemble. In Figure 3b we present the reconstruction of each configuration $\hat{\mathbf{r}}$ at each value of λ superposed together with the true configuration $\mathbf{r}$. The folded configurations A and B possess RMSD reconstruction accuracies of 0.08−0.13 nm, the unfolded configurations D and E possess accuracies of 0.20−0.28 nm, and the transition configuration C accuracies of 0.16−0.24 nm. Infrequent observation of transition states and metastable extended configurations results in poorer reconstruction of these transitory states, and this effect is indeed amplified at larger λ.

Although it is the case that transitory states are more poorly reconstructed than those residing in the folded well, transition states are occupied by only ∼1% of the test trajectory and therefore make only a small contribution to the overall mean RMSD accuracy. The overwhelming determinant of the degradation in the RMSD accuracy with increasing λ is therefore the globally poorer reconstruction of all configurations, even those within the relatively well sampled metastable wells, due to the loss of temporal resolution associated with the more severe degree of temporal averaging that results from larger time bins.

Villin. *STAR Training.* STAR hyperparameters are calibrated over configurations harvested from the first 80% of the 125.6 μs simulation trajectory (100.5 μs with 502 325 frames at 0.2 ns intervals) using Dv = 37 465, λ = 0.2 ns, and $I_D = \infty$. We reconstruct the N = 287 heavy atoms of Villin to produce an MD trajectory $\mathbf{r}(t) \in \mathbb{R}^{861}$. Diffusion maps with a kernel bandwidth $\varepsilon = e^{-2.5}$ nm were used to extract the manifold $\mathcal{M}$ from $\mathbf{r}(t)$, and a gap in the eigenvalue spectrum was used to determine a 3D embedding spanned by $\{\Psi_1, \Psi_2, \Psi_3\}$ (Figure S2). Delay vectors $\mathbf{y}(t)$ were constructed from head-to-tail distance time series data $v(t)$ computed between terminal heavy atoms. A delay dimensionality of d = 11 and delay time of $\tau = \lambda$ were employed. Diffusion maps with a kernel bandwidth ε = 1 nm were used to extract the manifold $\mathcal{M}'$ from $\mathbf{y}(t)$, producing 3D embeddings spanned by $\{\Psi'_1, \Psi'_2, \Psi'_3\}$. The map from $\mathcal{M}'$ to $\mathcal{M}$ was parametrized by a 3−15−15−15−15−3 ANN trained over 200 epochs with a batch size of 750 and a learning rate of 10^{-5} using the Adam algorithm.[45] The map from $\mathcal{M}$ to $\hat{\mathbf{r}}$ was parametrized by a 3−6−578−1150−1722−861 ANN trained over 240 epochs with a batch size of 1000 and a learning rate of 5×10^{-7} using the Adam algorithm.[45]

STAR Deployment and RMSD Dependence on (Dv, λ, I_D). Using the hyperparameters specified in the previous subsection, we trained 100 independent STAR models at each combination of $\lambda = \{0.2, 1, 2, 10, 20\}$ ns, $Dv = \{10^3, 10^4, 2 \times 10^4, 4 \times 10^4\}$, and $I_D = \{10^3, 10^4, 10^5, 10^6, \infty\}$ triplet. The heavy atom RMSD reconstruction accuracies of each model on the 20% hold-out test partition (25.1 μs comprising 125 581 frames at 0.2 ns intervals) are illustrated in Figure 5. We observe RMSD reconstruction accuracies between 0.35 and 0.46 nm for all (Dv, λ, I_D) triplets excluding $I_D = 10^3$ triplets where, as was the case for Chignolin, the low signal-to-noise ratio prevented the ANN from learning a converged $\mathcal{M}' \to \mathcal{M}$ diffeomorphic map. Villin is a medium-sized protein comprising 35 residues, so the per residue RMSD lies in the range 0.010−0.013 nm. We can calculate a baseline RMSD accuracy from test data of 0.51 nm by computing the RMSD of every frame against the configuration from the test trajectory that results in the lowest overall RMSD averaged over all other configurations. STAR reconstruction performance should be judged relative to this baseline.

As with Chignolin, the clearest trend is the variation in reconstruction accuracy with the signal-to-noise ratio I_D, for which we observe in Figure 4 a nearly 10% increase in RMSD reconstruction error from ∼0.39 nm to ∼0.44 nm in moving from $I_D = 10^5$ to 10^4 averaged across the λ and Dv values. Further increases of I_D result in ∼1% improvements averaging across all tested values of Dv and λ upon reaching the $I_D \to \infty$ limit. We note that the degree of improvement varies substantially with time resolution, approaching 10−16% in high time resolution regimes but falling to less than 1% at low time resolutions. We were unable to learn a mapping between $\mathcal{M}'$ and $\mathcal{M}$ at $I_D = 10^3$. These trends reflect a floor on the test data signal-to-noise ratio required to reliably train the STAR pipeline. We find the reconstruction accuracy to be largely insensitive to data volume Dv. For a particular choice of I_D and λ, modulating Dv over the full range of 10^3 to 4×10^4 leads to an average of less than 1% changes in the RMSD accuracy. Finally, for a fixed I_D and Dv, increasing λ over the range 0.2−20 ns ($\lambda/\tau_{fold} = 3.6 \times 10^{-2}$ to 3.6×10^{-3}) leads to up to 14% degradation in the RMSD for $I_D \geq 10^4$, suggesting the reconstruction quality is also quite sensitive to the time resolution λ.

For training data volumes $Dv \geq 10^4$ samples, photon counts per bin of $I_D \geq 10^5$, and time resolutions of $\lambda \leq 10$ ns ($\lambda/\tau_{fold} \leq 3.6 \times 10^{-3}$), we can achieve heavy atom RMSD reconstruction accuracies of 0.42 nm or better, representing an 18% improvement over the RMSD = 0.51 nm baseline. These conditions are largely the same as for Chignolin. However, for Villin, even in the low signal-to-noise ratio regime (i.e., $I_D = 10^4$), reconstruction accuracy is better than baseline value up until the failure to produce a map at $I_D = 10^3$.

Analysis of $\mathcal{M}$. For $I_D \geq 10^5$ photons per bin, the temporal time binning λ determines the majority of variation in reconstruction accuracy, with the dependence of the RMSD reconstruction accuracy on data volume Dv and donor dye brightness I_D being comparatively weak. We present in Figure 5 the all-atom manifold $\mathcal{M}$ at each λ value for $Dv = 4 \times 10^4$ samples and $I_D = 10^5$ photons per bin. For visual clarity, we consider a 2D projection of the 3D manifold into the $\Psi_1 - \Psi_2$ plane spanned by the two leading eigenvectors. This 2D projection is sufficient to illuminate the changes in global structure and free energy landscape over the manifold as a function of λ. Additional projections of the manifold are presented in Figure S3. Considering the highest resolution manifold at λ = 0.2 ns as the ground truth (Figure 5a-A), we observe a single free energy minimum at ($\Psi_1 \approx -3.1 \times 10^{-4}$, $\Psi_2 \approx -9.9 \times 10^{-4}$) corresponding to the folded state. The higher free energy region surrounding this region contains

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

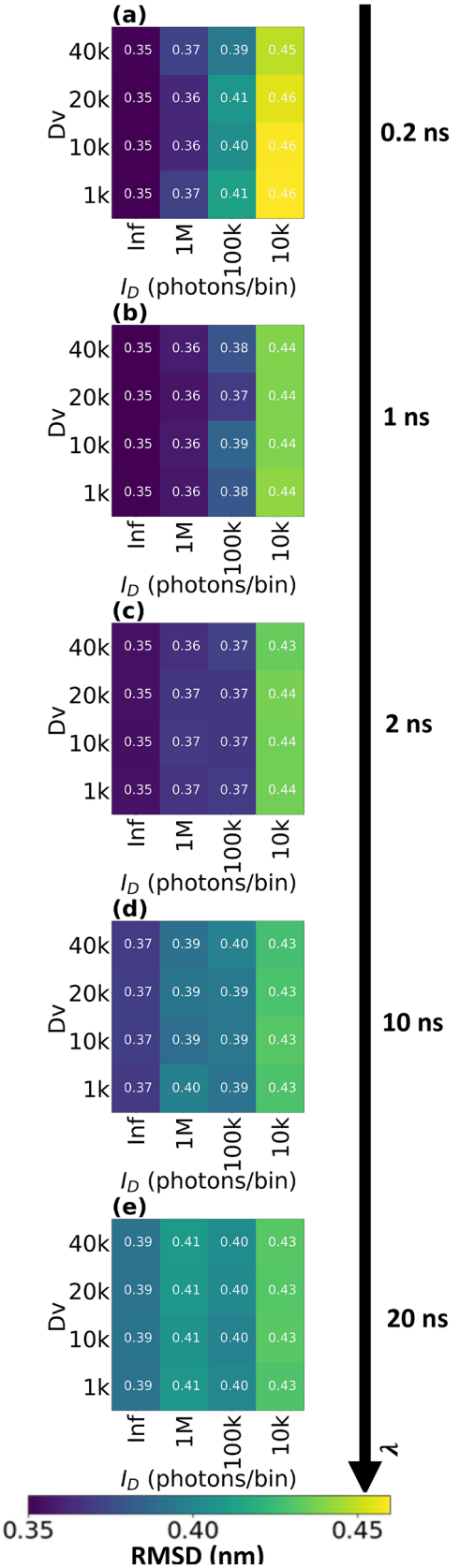


Figure 4. Heavy atom RMSD reconstruction accuracies for Villin as a function of training data volumes *Dv*, time bin resolution λ, and signal-to-noise ratio I_D.

members of the unfolded ensemble. Previous work applying dimensionality reduction to molecular dynamics simulations of Villin similarly reflects this single free energy well.[48] Other works reflect the emergence of multiple energy wells at higher temperatures or due to denaturing.[49] The most obvious effect of increasing λ from 0.2 to 20 ns is the translation and shrinkage of the free energy surface into the lower left corner of the $\Psi_1 - \Psi_2$ projection (Figure 5a). The absolute values and magnitudes of $\Psi_1 - \Psi_2$ between different embeddings are not meaningful, only the relative locations of points within each embedding. As such, it is more informative to compare the zoomed insets of each image Figure 5a-A−E that rescale each manifold onto approximately the same scale. These visualizations reveal that increases in λ result in increased averaging over contiguous molecular configurations and a concomitant mixing of configurations between the folded and unfolded states that smears out of the free energy minimum. At values of λ greater than 2 ns, this results in a large free energy minimum that encapsulates both folded and unfolded states. In all cases, however, the relative locations of the folded and unfolded configurations are insensitive to the value of λ.

Again, we hypothesize that configurations within the transition region and extended states outside of the metastable unfolded state are more likely to be poorly reconstructed than those within the metastable basins due to relatively poorer sampling at large λ. To test this hypothesis, we select from our validation trajectory five representative configurations A−E and project them onto each manifold $\mathcal{M}$ in Figure 5a. Configurations A and B belong to the folded macrostate, while C, D, and E are part of the unfolded ensemble. In Figure 5b, we present each reconstructed configuration $\hat{\mathbf{r}}$ at each value of λ superposed together with the true configuration $\mathbf{r}$. The folded configurations A and B possess RMSD reconstruction accuracies of 0.15−0.42 nm, while the unfolded configurations C, D, and E possess accuracies of 0.33−0.55 nm. Although we see that STAR can adequately track the overall features of the protein conformational state over the five selected points, our analysis confirms that less frequently observed (i.e., higher free energy) states tend to have poorer reconstruction accuracies and that this effect is amplified with increasing λ as illustrated in Figure 5c.

CONCLUSIONS

In this work, we have demonstrated the use of an approach based on Takens' Delay Embedding Theorem termed Single-molecule TAkens Reconstruction (STAR) to predict the molecular structure of proteins from low-dimensional time series of intramolecular distances. The fundamental motivation of this approach is to provide a means to "upgrade" experimental measurements of intramolecular distances accessible to experimental techniques such as smFRET to a prediction for the atomistic coordinates of the molecule. In this manner, we can use a trained STAR model to furnish a time-resolved molecular trajectory directly from experimental data. The STAR models are trained over molecular simulation trajectories that provide the molecular configurations and intramolecular distances needed to learn the mapping from the latter to the former. The trained model may then may be applied to novel smFRET data without the need to conduct any additional simulations. Provided the molecular simulation model employed is a good representation of the protein under the conditions of interest, the simulation trajectories are sufficiently long to sample the experimentally relevant

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

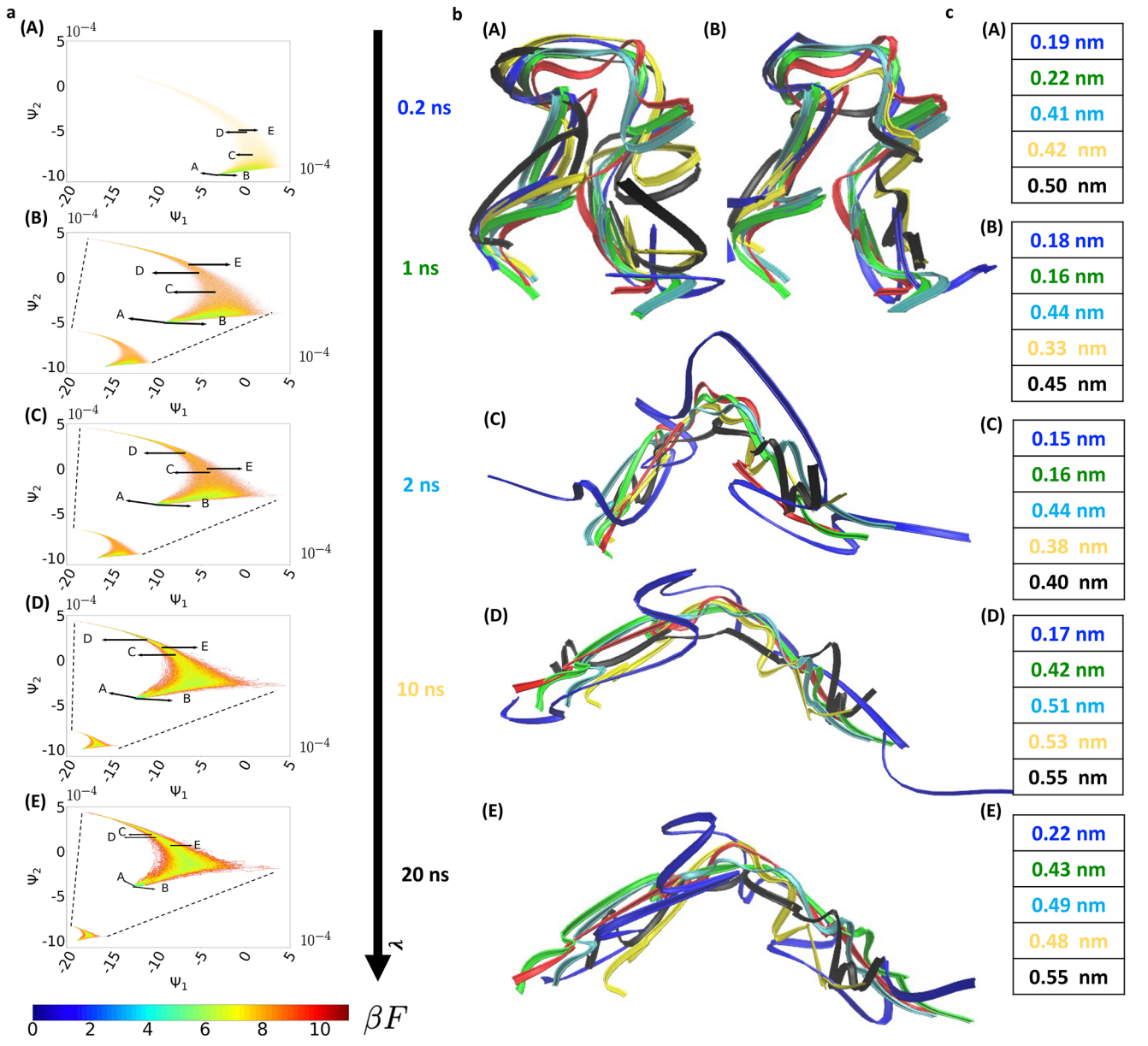


Figure 5. Manifolds $\mathcal{M}$ and heavy atom RMSD reconstruction accuracies for Villin as a function of time bin resolution λ. (a) All-atom manifolds $\mathcal{M}$ within STAR pipelines trained at $Dv = 4 \times 10^4$ samples and $I_D = 10^5$ for each value of $\lambda = \{0.2, 1, 2, 10, 20\}$ ns. Manifolds are represented as scatter plots showing the embedding of each of the $Dv = 4 \times 10^4$ MD configurations $\mathbf{r}$ projected into the two leading diffusion map eigenvectors $\{\Psi_1, \Psi_2\}$ spanning each manifold. Zoomed-in cutouts are provided for the subpanels B−E. Color distributions characterize associated free energy $\beta F(\Psi_1, \Psi_2)$ of each point computed by collecting histograms over the empirical probability distribution at a bin size of $\{\Delta\Psi_1 = 5 \times 10^{-7}, \Delta\Psi_2 = 5 \times 10^{-7}\}$. The arbitrary zero of free energy in each panel is specified by subtracting the computed energy of the minimum free energy point on the landscape from all other points. Five representative configurations A−E are selected from the MD test trajectory and projected onto each manifold. A and B reside in the folded macrostate, while the C, D, and E are in the unfolded ensemble. (b) Visualizations of the reconstructions $\hat{\mathbf{r}}$ of each configuration A−E at each time resolution $\lambda = 0.2$ ns (blue), 1 ns (green), 2 ns (cyan), 10 ns (yellow), and 20 ns (black). For visual clarity, configurations are represented as ribbons tracing the backbone of the protein and are superposed upon the true configuration extracted from the MD test trajectory $\mathbf{r}$ (red). (c) RMSD reconstruction corresponding to each state are listed next to each image employing the same color-coding as the reconstructions. All molecular visualizations are constructed using VMD.[29]

configurational states, and the STAR model is properly trained, we anticipate that the model should be able to accurately predict molecular configurations from new (or hold-out) simulated smFRET trajectories and, ultimately, experimental smFRET data.

The primary contribution of the present work is to demonstrate that we can construct STAR models for two fast-folding mini-proteins, Chignolin ($\tau_{fold} = 0.6$ μs) and Villin ($\tau_{fold} = 2.8$ μs), under conditions of trajectory length, time resolution, and signal-to-noise ratio (i.e., dye intensity) that bridge computationally tractable simulations to experimentally realistic FRET conditions. The trained models achieve heavy atom RMSD reconstruction accuracies over a hold-out molecular dynamics test set of 0.12 and 0.42 nm, respectively. As a point of comparison, these accuracies are commensurate with the ∼0.1 nm accuracies attainable by cryo-electron microscopy and X-ray crystallography.[1,2] In each case, we achieve these results by training over molecular simulation

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

trajectories of 0.7−3.3× the characteristic protein folding time, with a temporal resolution of 1/120−1/280× the folding time, and signal-to-noise ratios commensurate with ∼10^5 photons per time bin.

The present work demonstrates and validates STAR against synthetic smFRET trajectories generated from hold-out molecular simulation trajectories. This is vital for validation of the method because the ground truth atomic coordinates of the testing trajectories are exactly known, but it would be desirable in future work to apply a trained STAR model to real experimental smFRET data. The mini-proteins studied herein are too fast folding to be accessible to existing smFRET technology, but our results lay the foundations and specify the experimental conditions necessary to perform extrapolative identification of putative target systems. State-of-the-art photon-by-photon single-molecule instruments can produce observations at λ = 1−10 μs with a fluorophore pair such as Cy3/Cy5.[26−28] The constraints on data volume, time resolution, and signal-to-noise ratio identified in this work suggest that STAR could be deployed on proteins with characteristic folding times of τ_{fold} = 100−1000 μs. Such a protein system would be simultaneously amenable to sufficiently high temporal resolution smFRET measurements using state-of-the-art probes and sufficiently fast-folding that it would require simulation trajectory training trajectories totaling 100−3000 μs. STAR training only requires temporally continuous blocks of molecular simulation trajectories of length $d\tau$, where d is the delay dimensionality and τ is the delay time, meaning that the training data can comprise a large number of short, discontinuous training trajectories efficiently generated by parallel computation. All-atom molecular dynamics simulations at these time scales are expensive but relatively accessible on high-performance supercomputing hardware.[25,50−52] This analysis suggests as one possible target system the 54-residue engrailed homeodomain protein (PDB: 1ENH), which possesses a characteristic unfolding time of 910 μs at 25 °C that can be modulated to 4.8 μs at 63 °C,[53] has been extensively studied by molecular simulation,[54,55] and is sufficiently large to accommodate FRET fluorophores operating within the preferred range of 2−8 nm.[53,56]

In future work, we plan to increase the robustness of STAR to noise, integrate multichannel smFRET information, and explore the transferability of our models. Reduction in noise effects via kernel choice[57] or integration of hidden Markov models[58] may help reduce photon count requirements. Multichannel smFRET signals simultaneously recording multiple intramolecular distances between multiple pairs of probes can be harnessed in conjunction with multivariate Takens' Theorem[14,59] to improve reconstruction quality through multiplexed dynamical observations. Study of optimal FRET fluorophore placement can help identify preferred grafting positions for the probes to maximize the dynamical information captured by smFRET observables and minimize reconstruction errors.[60,61] Investigation of transferability of latent space manifolds and STAR mappings across temperature, pressure, molecular force field, coarse graining, and solvent viscosity would improve the versatility of trained STAR models while reducing training requirements. Furthermore, transitory and extended state reconstruction can be improved by adaptive sampling of infrequently sampled states[62] and facilitate applications of STAR to larger and slower folding molecules by judicious selection and generation of training data. Beyond protein reconstruction, we would like to study other biomolecules such as DNA and RNA and also consider the incorporation of solvent-based observables.[15,16,63,64] Lastly, we also see opportunities for applications of STAR to other fields where it is of interest to reconstruct the state of a high-dimensional dynamical system that is implicitly observed through an incomplete set of low-dimensional variables, including epidemiology, climatology, and econometrics.

APPENDIX

Synthetic smFRET Noise Model

The distance between a smFRET donor and acceptor pair is computed from the measured FRET efficiency, the fraction of donor excitons that is transferred to the acceptor, by measuring the emission intensities ("brightnesses") of the donor and acceptor fluorophores, assuming isotropic dye orientations.[3,33] Mathematically, the distance r is related to the emission intensities as[3,33]

$$r = R_0\left(\frac{\phi_A I_{DA}}{\phi_D(I_{AD} - I_A)}\right)^{1/6} \tag{3}$$

where ϕ_D is the quantum yield of the donor, ϕ_A is the quantum yield of the acceptor, R_0 is the characteristic FRET radius for the donor−acceptor pair, I_A is the intensity of the acceptor channel under direct excitation by the laser without donor present in the system, I_{DA} is the intensity of the donor channel when acceptor is present, and I_{AD} is the intensity of the acceptor channel when donor is present. The quantum yields and FRET radius are determined by the particular choice of donor and acceptor fluorophores. Ideal quantum yields correspond to $\phi_D = \phi_A = 1$. Experiments typically use dyes with quantum yields on the order of 0.1−1, with dyes like Rhodamine 6G having a yield of 0.95.[3,65] Typical FRET radii are on the order of 1−10 nm.[3]

Assuming no correlations between independent variables, no detector noise, isotropic dye orientations, and no direct excitation of the acceptor fluorophore by the excitation laser ($I_A = 0$), propagation of uncertainties yields

$$\sigma_r^2 = \left(\frac{\partial r}{\partial I_{DA}}\right)^2 \sigma_{I_{DA}}^2 + \left(\frac{\partial r}{\partial I_{AD}}\right)^2 \sigma_{I_{AD}}^2 \tag{4}$$

where σ_r is the standard deviation in r, $\sigma_{I_{DA}}$ is the standard deviation in I_{DA}, and $\sigma_{I_{AD}}$ is the standard deviation in I_{AD}. The partial derivatives follow straightforwardly from eq 3 as

$$\frac{\partial r}{\partial I_{DA}} = \frac{r}{6I_{DA}} \tag{5}$$

$$\frac{\partial r}{\partial I_{AD}} = -\frac{r}{6I_{AD}} \tag{6}$$

Assuming Poisson statistics in photon counting, $\sigma_{I_{DA}}^2 = I_{DA}$ and $\sigma_{I_{AD}}^2 = I_{AD}$, and it immediately follows that

$$\sigma_r^2 = \frac{I_{AD}r^2 + I_{DA}r^2}{36I_{AD}I_{DA}} \tag{7}$$

$$\Rightarrow\left(\frac{\sigma_r}{r}\right)^2 = \frac{1}{36I_{DA}}\left(1 + \frac{I_{DA}}{I_{AD}}\right) \tag{8}$$

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

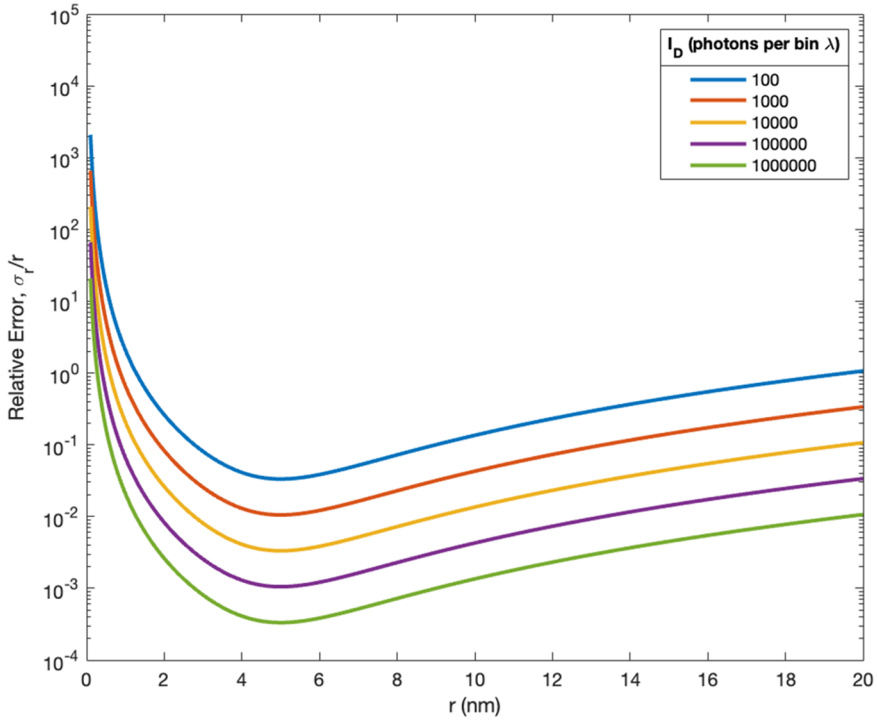


Figure 6. Variation of σ_r/r defining the relative uncertainty in r as a function of the donor−acceptor distance r at various values of donor intensity I_D (eq 13). For the purposes of this calculation, we adopt prototypical values of $\phi_D = \phi_A = 1$ for the donor and acceptor quantum yields and $R_0 = 5$ nm for the FRET radius of the donor−acceptor pair.

Because I_{DA} and I_{AD} are not directly available from our MD simulation trajectory, it is convenient to re-express the right-hand side as a function of r and any fluorophore-specific constants. To do so, we first rearrange eq 3 to eliminate I_{DA}/I_{AD}:

$$\frac{I_{DA}}{I_{AD}} = \left(\frac{r}{R_0}\right)^6 \left(\frac{\phi_D}{\phi_A}\right) \tag{9}$$

and substitute into eq 8 to yield

$$\left(\frac{\sigma_r}{r}\right)^2 = \frac{1}{36 I_{DA}}\left[1 + \left(\frac{r}{R_0}\right)^6 \left(\frac{\phi_D}{\phi_A}\right)\right] \tag{10}$$

To eliminate I_{DA}, we require an additional equation. Equation 3 computes r from experimental measurements of I_{AD} and I_{DA} that include both the donor and the acceptor channels. It is also possible to estimate r using the donor channel alone:[33,66]

$$r = R_0 \left(\frac{I_{DA}}{I_D - I_{DA}}\right)^{1/6} \tag{11}$$

where I_D is the intensity of the donor channel under direct excitation by the laser without acceptor present in the system, which is dictated by the choice of donor fluorophore and laser power and is not a function of r. Rearranging this expression for I_{DA} yields

$$I_{DA} = \frac{I_D}{1 + \left(\frac{R_0}{r}\right)^6} \tag{12}$$

Inserting eq 12 into eq 10 results in our noise model:

$$\left(\frac{\sigma_r}{r}\right)^2 = \frac{1 + \left(\frac{R_0}{r}\right)^6}{36 I_D}\left[1 + \left(\frac{r}{R_0}\right)^6 \left(\frac{\phi_D}{\phi_A}\right)\right] \tag{13}$$

In Fig. 6, we present a plot of eq 13 to illustrate the variation of σ_r/r as a function of r parameterized by I_D. We observe the lowest uncertainties in the vicinity of the FRET radius R_0 and an asymmetric increase as the donor−acceptor pair moves to smaller or larger separations. Physically, this is due to increased shot noise in I_{DA} at shorter distances as the FRET efficiency increases and the donor becomes less bright (i.e., fewer emitted donor photons), and increased shot noise in I_{AD} at longer distances as the FRET efficiency decreases and the acceptor becomes less bright. As anticipated, the relative error in r is mitigated by the use of brighter donors with larger values of I_D.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00920.

Supplementary methods − technical details of the application of the STAR methodology: learning the atomistic manifold $\mathcal{M}$ from the MD trajectory $\mathbf{r}(t)$; learning the Takens' manifold $\mathcal{M}'$ from the smFRET time series $v(t)$; learning the diffeomorphism from $\mathcal{M}'$ to $\mathcal{M}$; learning the reconstruction $\hat{\mathbf{r}}(t)$ from the manifold $\mathcal{M}$; and deploying the trained STAR model; and supplementary figures − diffusion map eigenvalue spectra for Chignolin and Villin, and additional projections of the Villin all-atom manifold $\mathcal{M}$(PDF)

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667

AUTHOR INFORMATION

Corresponding Author

Andrew L. Ferguson − *Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0002-8829-9726; Email: andrewferguson@uchicago.edu

Authors

Maximilian Topel − *Department of Physics, University of Chicago, Chicago, Illinois 60637, United States*

Ayesha Ejaz − *Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States*

Allison Squires − *Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0002-2417-1432

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c00920

Notes

The authors declare the following competing financial interest(s): A.L.F. is a co-founder and consultant of Evozyne, Inc. and a co-author of U.S. Patent Applications 16/887710 and 17/642582, U.S. Provisional Patent Applications 62/853919, 62/900420, and 63/314898, and International Patent Applications PCT/US2020/035206 and PCT/US2020/050466.

ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under grant no. DMS-1841810. We are grateful to D.E. Shaw Research for sharing the simulation trajectories used as the basis of our model training. This work was completed in part with resources provided by the University of Chicago Research Computing Center. We gratefully acknowledge computing time on the University of Chicago high-performance GPU-based cyberinfrastructure supported by the National Science Foundation under grant no. DMR-1828629.

REFERENCES

(1) Schüttelkopf, A. W.; van Aalten, D. M. F. PRODRG: A tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica Section D* **2004**, *60*, 1355−1363.

(2) Chang, J. C.; Rosenthal, S. J. In *Biomedical Nanotechnology: Methods and Protocols*; Hurst, S. J., Ed.; Humana Press: Totowa, NJ, 2011; pp 51−62.

(3) Roy, R.; Hohng, S.; Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **2008**, *5*, 507−516.

(4) Zerze, G. H.; Best, R. B.; Mittal, J. Modest influence of FRET chromophores on the properties of unfolded proteins. *Biophys. J.* **2014**, *107*, 1654−1660.

(5) Knowles, T. P. J.; Vendruscolo, M.; Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **2014**, *15*, 384−396.

(6) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic Press: New York, 2002.

(7) Wong-ekkabut, J.; Karttunen, M. The good, the bad and the user in soft matter simulations. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2016**, *1858*, 2529−2538.

(8) Childers, M. C.; Daggett, V. Insights from molecular dynamics simulations for computational protein design. *Molecular Systems Design & Engineering* **2017**, *2*, 9−33.

(9) Lankiewicz, L.; Malicka, J.; Wiczk, W. Fluorescence resonance energy transfer in studies of inter-chromophoric distances in biomolecules. *Acta Biochimica Polonica* **1997**, *44*, 477−489.

(10) Takens, F. Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence* **1981**, *898*, 366−381.

(11) Sauer, T.; Yorke, J. A.; Casdagli, M. Embedology. *J. Stat. Phys.* **1991**, *65*, 579−616.

(12) Packard, N.; Crutchfield, J.; Farmer, J.; Shaw, R. Geometry from a time series. *Phys. Rev. Lett.* **1980**, *45*, 712−716.

(13) Broomhead, D. S.; King, G. P. Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena* **1986**, *20*, 217−236.

(14) Cao, L.; Mees, A.; Judd, K. Dynamics from multivariate time series. *Physica D: Nonlinear Phenomena* **1998**, *121*, 75−88.

(15) Stark, J. Delay embeddings for forced systems. I. Deterministic forcing. *J. Nonlinear Sci.* **1999**, *9*, 255−332.

(16) Stark, J.; Broomhead, D. S.; Davies, M.; Huke, J. Delay embeddings for forced systems. II. Stochastic forcing. *J. Nonlinear Sci.* **2003**, *13*, 519−577.

(17) Vialar, T. *Complex and Chaotic Nonlinear Dynamics: Advances in Economics and Finance, Mathematics and Statistics*; Springer: New York, 2009.

(18) Kantz, H.; Schreiber, T. *Nonlinear Time Series Analysis*, 2nd ed.; Cambridge, 2005.

(19) Ye, H.; Beamish, R. J.; Glaser, S. M.; Grant, S. C. H.; Hsieh, C.-H.; Richards, L. J.; Schnute, J. T.; Sugihara, G. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E1569−76.

(20) Topel, M.; Ferguson, A. L. Reconstruction of protein structures from single-molecule time series. *J. Chem. Phys.* **2020**, *153*, 194102.

(21) Salem, C.-B.; Ploetz, E.; Lamb, D. C. In *Spectroscopy and Dynamics of Single Molecules*; Johnson, C. K., Ed.; Developments in Physical & Theoretical Chemistry; Elsevier: New York, 2019; pp 71−115.

(22) Demchenko, A. P. Photobleaching of organic fluorophores: Quantitative characterization, mechanisms, protection. *Methods and Applications in Fluorescence* **2020**, *8*, 022001.

(23) Ha, T.; Tinnefeld, P. Photophysics of fluorescence probes for single molecule biophysics and super-resolution imaging. *Annu. Rev. Phys. Chem.* **2012**, *63*, 595−617.

(24) Gensch, T.; Böhmer, M.; Aramendía, P. F. Single molecule blinking and photobleaching separated by wide-field fluorescence microscopy. *J. Phys. Chem. A* **2005**, *109*, 6652−6658.

(25) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517−520.

(26) Pirchi, M.; Tsukanov, R.; Khamis, R.; Tomov, T. E.; Berger, Y.; Khara, D. C.; Volkov, H.; Haran, G.; Nir, E. Photon-by-photon hidden Markov model analysis for microsecond single-molecule FRET kinetics. *J. Phys. Chem. B* **2016**, *120* (51), 13065−13075.

(27) Phelps, C.; Israels, B.; Jose, D.; Marsh, M. C.; von Hippel, P. H.; Marcus, A. H. Using microsecond single-molecule FRET to determine the assembly pathways of T4 ssDNA binding protein onto model DNA replication forks. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E3612−E3621.

(28) Oikawa, H.; Takahashi, T.; Kamonprasertsuk, S.; Takahashi, S. Microsecond resolved single-molecule FRET time series measurements based on the line confocal optical system combined with hybrid photodetectors. *Phys. Chem. Chem. Phys.* **2018**, *20*, 3277−3285.

(29) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.

(30) Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 13597−13602.

(31) Ferguson, A. L.; Panagiotopoulos, A. Z.; Kevrekidis, I. G.; Debenedetti, P. G. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chem. Phys. Lett.* **2011**, *509*, 1−11.

(32) Stryer, L.; Haugland, R. P. Energy transfer: A spectroscopic ruler. *Proc. Natl. Acad. Sci. U.S.A.* **1967**, *58*, 719−726.

(33) Hellenkamp, B.; Schmid, S.; Doroshenko, O.; Opanasyuk, O.; Kühnemuth, R.; Rezaei Adariani, S.; Ambrose, B.; Aznauryan, M.;

Barth, A.; Birkedal, V.; Bowen, M. E.; Chen, H.; Cordes, T.; Eilert, T.; Fijen, C.; Gebhardt, C.; Götz, M.; Gouridis, G.; Gratton, E.; Ha, T.; Hao, P.; Hanke, C. A.; Hartmann, A.; Hendrix, J.; Hildebrandt, L. L.; Hirschfeld, V.; Hohlbein, J.; Hua, B.; Hübner, C. G.; Kallis, E.; Kapanidis, A. N.; Kim, J.-Y.; Krainer, G.; Lamb, D. C.; Lee, N. K.; Lemke, E. A.; Levesque, B.; Levitus, M.; McCann, J. J.; Naredi-Rainer, N.; Nettels, D.; Ngo, T.; Qiu, R.; Robb, N. C.; Röcker, C.; Sanabria, H.; Schlierf, M.; Schröder, T.; Schuler, B.; Seidel, H.; Streit, L.; Thurn, J.; Tinnefeld, P.; Tyagi, S.; Vandenberk, N.; Vera, A. M.; Weninger, K. R.; Wünsch, B.; Yanez-Orozco, I. S.; Michaelis, J.; Seidel, C. A. M.; Craggs, T. D.; Hugel, T. Precision and accuracy of single-molecule FRET measurements—a multi-laboratory benchmark study. *Nat. Methods* **2018**, *15*, 669−676.

(34) Wolf, S.; Lickert, B.; Bray, S.; Stock, G. Multisecond ligand dissociation dynamics from atomistic simulations. *Nat. Commun.* **2020**, *11*, 2918.

(35) Asher, W. B.; Geggier, P.; Holsey, M. D.; Gilmore, G. T.; Pati, A. K.; Meszaros, J.; Terry, D. S.; Mathiasen, S.; Kaliszewski, M. J.; McCauley, M. D.; Govindaraju, A.; Zhou, Z.; Harikumar, K. G.; Jaqaman, K.; Miller, L. J.; Smith, A. W.; Blanchard, S. C.; Javitch, J. A. Single-molecule FRET imaging of GPCR dimers in living cells. *Nat. Methods* **2021**, *18*, 397−405.

(36) Nir, E.; Michalet, X.; Hamadani, K. M.; Laurence, T. A.; Neuhauser, D.; Kovchegov, Y.; Weiss, S. Shot-noise limited single-molecule FRET histograms: Comparison between theory and experiments. *J. Phys. Chem. B* **2006**, *110*, 22103−22124.

(37) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **2011**, *100*, L47−9.

(38) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.

(39) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341−346.

(40) Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511−519.

(41) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695−1697.

(42) Shan, Y.; Klepeis, J. L.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J. Chem. Phys.* **2005**, *122*, 054101.

(43) Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. 10 residue folded peptide designed by segment statistics. *Structure* **2004**, *12*, 1507−1518.

(44) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. Sub-microsecond protein folding. *J. Mol. Biol.* **2006**, *359*, 546−553.

(45) Kingma, D. P.; Ba, J. *Adam: A Method for Stochastic Optimization*, 2014.

(46) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; de Fabritiis, G.; Noé, F.; Clementi, C. Machine learning of coarse-grained molecular dynamics force fields. *ACS Central Science* **2019**, *5*, 755−767.

(47) Shaffer, P.; Valsson, O.; Parrinello, M. Enhanced, targeted sampling of high-dimensional free-energy landscapes using variationally enhanced sampling, with an application to chignolin. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 1150−1155.

(48) Sittel, F.; Jain, A.; Stock, G. Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates. *J. Chem. Phys.* **2014**, *141*, 014111.

(49) Lei, H.; Wu, C.; Liu, H.; Duan, Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4925−4930.

(50) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J. P.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J.; Chow, E.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. Millisecond-scale molecular dynamics simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, 2009.

(51) Han, W.; Schulten, K. Fibril elongation by AB17−42: Kinetic network analysis of hybrid-resolution molecular dynamics simulations. *J. Am. Chem. Soc.* **2014**, *136*, 12450−12460.

(52) Yin, Y.; Arkhipov, A.; Schulten, K. Simulations of membrane tubulation by lattices of amphiphysin N-BAR domains. *Structure* **2009**, *17*, 882−892.

(53) Mayor, U.; Johnson, C. M.; Daggett, V.; Fersht, A. R. Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13518−13522.

(54) Zhao, X.; Huang, X.; Sun, C. Molecular dynamics analysis of the engrailed homeodomain−DNA recognition. *J. Struct. Biol.* **2006**, *155*, 426−437.

(55) Koulgi, S.; Sonavane, U.; Joshi, R. Insights into the folding pathway of the Engrailed Homeodomain protein using replica exchange molecular dynamics simulations. *Journal of Molecular Graphics and Modelling* **2010**, *29*, 481−491.

(56) Huang, F.; Settanni, G.; Fersht, A. R. Fluorescence resonance energy transfer analysis of the folding pathway of Engrailed Homeodomain. *Protein Engineering, Design and Selection* **2008**, *21*, 131−146.

(57) Takeda, H.; Farsiu, S.; Milanfar, P. Robust kernel regression for restoration and reconstruction of images from sparse noisy data. *2006 International Conference on Image Processing*; 2006; pp 1257−1260.

(58) Liu, Y.; Park, J.; Dahmen, K. A.; Chemla, Y. R.; Ha, T. A comparative Study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis. *J. Phys. Chem. B* **2010**, *114*, 5386−5403.

(59) Gambin, Y.; Deniz, A. A. Multicolor single-molecule FRET to explore protein folding and binding. *Molecular BioSystems* **2010**, *6*, 1540−1547.

(60) Mittal, S.; Shukla, D. Maximizing kinetic information gain of Markov state models for optimal design of spectroscopy experiments. *J. Phys. Chem. B* **2018**, *122*, 10793−10805.

(61) Dimura, M.; Peulen, T.-O.; Sanabria, H.; Rodnin, D.; Hemmen, K.; Hanke, C. A.; Seidel, C. A.; Gohlke, H. Automated and optimally FRET-assisted structural modeling. *Nat. Commun.* **2020**, *11*, 5394.

(62) Chen, W.; Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* **2018**, *39*, 2079−2102.

(63) Pietrucci, F. Novel enhanced sampling strategies for transitions between ordered and disordered structures. *Handbook of Materials Modeling: Methods: Theory and Modeling* **2020**, 597−619.

(64) Han, J.; Zhang, L.; Car, R.; E, W. Deep potential: A general representation of a many-body potential energy surface. *Commun. Comput. Phys.* **2018**, *23*, 629−639.

(65) Magde, D.; Rojas, G. E.; Seybold, P. G. Solvent dependence of the fluorescence lifetimes of xanthene dyes. *Photochem. Photobiol.* **1999**, *70*, 737−744.

(66) Schuler, B. Single-molecule FRET of protein structure and dynamics - a primer. *J. Nanobiotechnol.* **2013**, *11*, S2.

https://doi.org/10.1021/acs.jctc.2c00920
J. Chem. Theory Comput. 2023, 19, 4654−4667