Increasing the accuracy of single-molecule data analysis using tMAVEN

Anjali R. Verma[1,*], Korak Kumar Ray[1,‡,*], Maya Bodick[1], Colin D. Kinz-Thompson[2], and Ruben L. Gonzalez, Jr.[1,†]

[1]*Department of Chemistry, Columbia University, New York, NY 10027 USA.*

[2]*Department of Chemistry, Rutgers University-Newark, Newark, NJ 07102 USA.*

[‡]*Current address: Single Molecule Imaging Group, MRC-London Institute of Medical Sciences, London W12 0HS, UK.*

*These authors contributed equally to this work.

[†]To whom correspondence may be addressed: Ruben L. Gonzalez, Jr, Department of Chemistry, Columbia University, 3000 Broadway, New York, NY 10027, Tel: (212) 854-1096, Email: rlg2118@columbia.edu

Abstract

Time-dependent single-molecule experiments contain rich kinetic information about the functional dynamics of biomolecules. A key step in extracting this information is the application of kinetic models, such as hidden Markov models (HMMs), which characterize the molecular mechanism governing the experimental system. Unfortunately, researchers rarely know the physico-chemical details of this molecular mechanism *a priori,* which raises questions about how to select the most appropriate kinetic model for a given single-molecule dataset and what consequences arise if the wrong model is chosen. To address these questions, we have developed and used time-series Modeling, Analysis, and Visualization ENvironment (tMAVEN), a comprehensive, open-source, and extensible software platform. tMAVEN can perform each step of the single-molecule analysis pipeline, from pre-processing to kinetic modeling to plotting, and has been designed to enable the analysis of a single-molecule dataset with multiple types of kinetic models. Using tMAVEN, we have systematically investigated mismatches between kinetic models and molecular mechanisms by analyzing simulated

examples of prototypical single-molecule datasets exhibiting common experimental complications, such as molecular heterogeneity, with a series of different types of HMMs. Our results show that no single kinetic modeling strategy is mathematically appropriate for all experimental contexts. Indeed, HMMs only correctly capture the underlying molecular mechanism in the simplest of cases. As such, researchers must modify HMMs using physico-chemical principles to avoid the risk of missing the significant biological and biophysical insights into molecular heterogeneity that their experiments provide. By enabling the facile, side-by-side application of multiple types of kinetic models to individual single-molecule datasets, tMAVEN allows researchers to carefully tailor their modeling approach to match the complexity of the underlying biomolecular dynamics and increase the accuracy of their single-molecule data analyses.

Statement of Significance

The power of time-dependent single-molecule biophysical experiments lies in their ability to uncover the molecular mechanisms governing experimental systems by computationally applying kinetic models to the data. While many software solutions have been developed to estimate the optimal parameters of such models, the results reported here show that the models themselves are often inherently mismatched with the molecular mechanisms they are being used to analyze. To investigate these mismatches and demonstrate how to best model the kinetics of a molecular mechanism, we have used time-series Modeling, Analysis, and Visualization ENvironment (tMAVEN), an open-source software platform we have developed that, among other features, enables the analysis of single-molecule datasets using different kinetic models within a single, extensible, and customizable pipeline.

Introduction

Single-molecule kinetics experiments have revolutionized our understanding of the dynamics of biomolecules and, consequently, the mechanisms of biomolecular functionality (1, 2). These techniques provide a uniquely detailed view of molecular mechanisms relative to more traditional

'bulk' techniques where mechanistic details, such as transient intermediates or rare events, are easily masked by macroscopic ensemble averaging. Furthermore, by observing the motions of an individual molecule through time, single-molecule kinetics experiments bypass the difficulties of biochemically synchronizing a stochastic biomolecular process with asynchronous dynamics (3). Altogether, the advantages conferred by single-molecule techniques have expanded both the set of biological systems whose kinetics can be investigated and the complexity of biomolecular dynamics that may be experimentally probed.

Despite the advantages of single-molecule techniques, the challenges of extracting mechanistic information from the associated experimental data (4) have limited their widespread application. Fortunately, many analysis approaches have been developed over the last few decades to address this problem (4, 5). In particular, hidden Markov models (HMMs) (6) have emerged as a popular method to describe the latent dynamics of a biological system by analyzing the time-dependent readout of a chosen experimental single-molecule signal (*e.g.*, fluorescence intensity, spatial position, end-to-end distance, electric current, *etc.*) (7–21). In this context, HMMs are used to mathematically describe the transitions between relatively stable signal values (*i.e.*, signal states) that are often observed in single-molecule experiments, thereby directly extracting the underlying kinetics of the molecular system from the experimental data.

A major consideration when analyzing single-molecule data with models like HMMs is the completeness of the kinetic information present in a signal *vs.* time trajectory. Each trajectory reports on the dynamic behavior of an individual molecule, thus the amount of molecular information that can be extracted from any one trajectory is limited (4)—particularly for techniques that rely on fluorescence, wherein the photophysical processes of the fluorophores severely restrict the length of the trajectory (22). In such cases, researchers often use their knowledge of the underlying physico-chemical properties of the biomolecular system to invoke the ergodic hypothesis (23) and model data from multiple trajectories in aggregate (*e.g.*, as in Refs. (24–27)). Thus, instead of modeling the kinetic behavior of an individual molecule, these analyses infer the behavior of a mesoscopic, homogenous ensemble consisting of hundreds to thousands of molecules that are assumed to be identical.

In reality, however, experimental ensembles of molecules are never identical and always

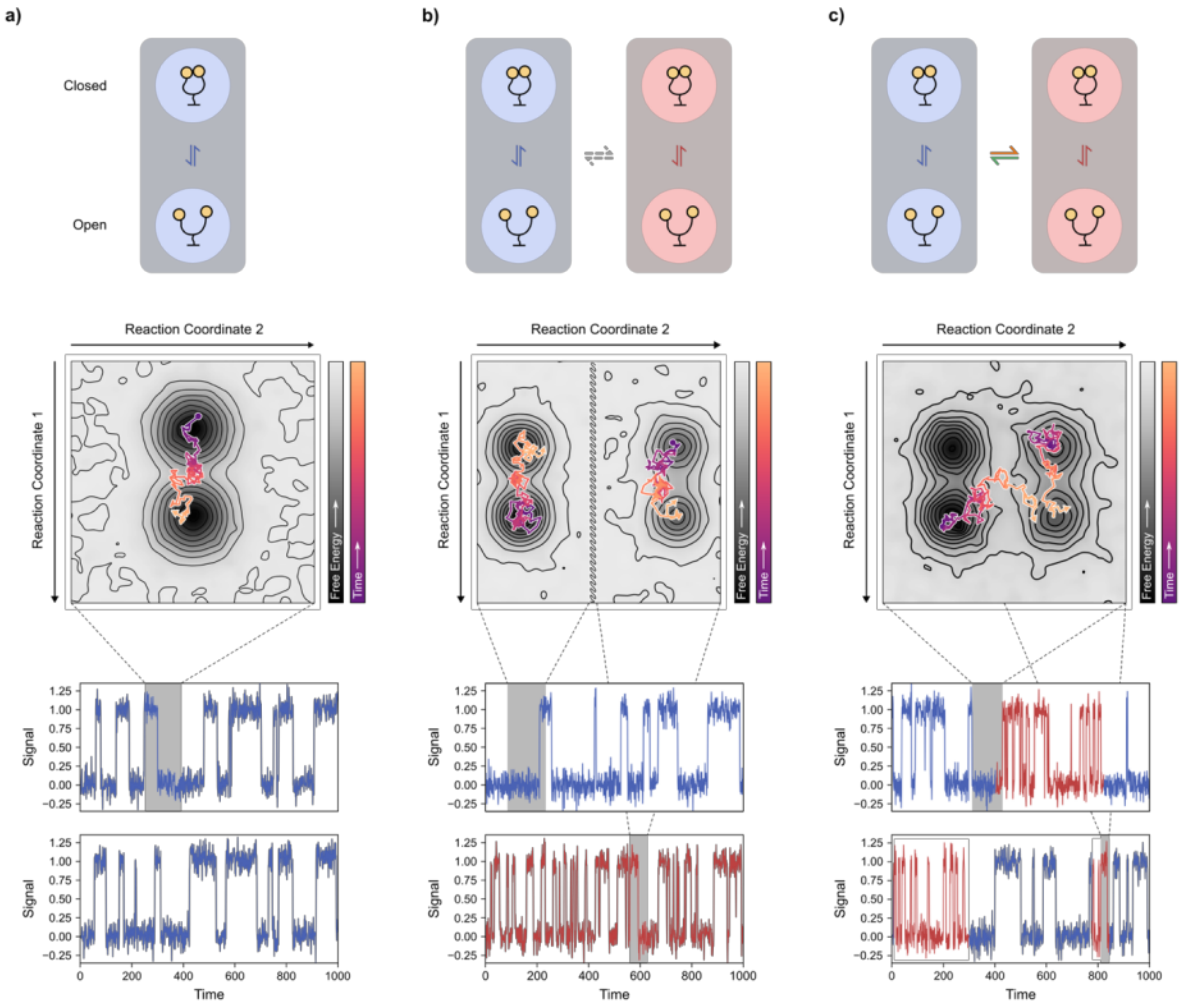

Figure 1. *Molecular mechanisms and their corresponding single-molecule signal vs. time trajectories.* *(top) Schematic of the molecular mechanism, (middle) the corresponding conformational free-energy landscape, and (bottom) single-molecule trajectories that capture changes in signal for Reaction Coordinate 1 for* ***(a)*** *homogeneous,* ***(b)*** *statically heterogeneous, and* ***(c)*** *dynamically heterogeneous biomolecular systems. Simulated random walkers on the conformational free-energy landscape, starting at circles and ending at arrows, show hypothetical individual molecules undergoing transitions that correspond to the grey areas of the single-molecule trajectories. For the heterogeneous cases, blue and red correspond respectively to slow and fast transitioning subpopulations (for static) and phases (for dynamic), which are differentiated along Reaction Coordinate 2. A discontinuity (hatched line) is shown in the landscape for* ***(b)*** *to signify the lack of allowed transition along Reaction Coordinate 2 in this case.*

exhibit some amount of heterogeneity (28). These heterogeneities may arise from the presence of several subpopulations of molecules within the experimental ensemble, or the presence of multiple molecular processes occurring over a range of timescales (Fig. 1). Regardless of the source, such heterogeneities reduce the apparent complexity of the biomolecular system under investigation by collapsing distinct kinetic processes into the same signal state. This creates a mismatch between the 'true' underlying molecular mechanism reported by the experiment (Fig. 1) and the kinetic model that a researcher might choose for analysis based on the observed signal (Fig. 2) (*i.e.*, a model-

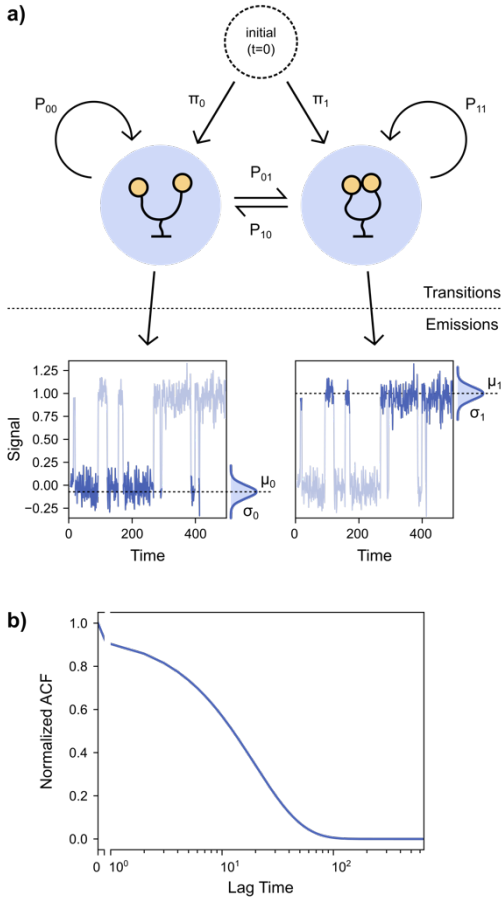

Figure 2. Schematic diagram of a kinetic model. (a) *A schematic diagram of a two-state HMM showing the separation between the transition DoFs comprised of the initial probabilities and the transition probabilities, and the emission DoFs comprised of the emission probability distributions.* ***(b)*** *The normalized ACF corresponding to the HMM in* ***(a)*** *expresses all the dynamics of the kinetic model from both the transitions and the emissions in a single analytical form.*

mechanism mismatch). Furthermore, while prior knowledge of the type or amount of heterogeneity present in an experimental dataset may guide the choice of kinetic model, such knowledge is not always readily available. Faced with a novel single-molecule dataset and a range of different kinetic models, the choice of which kinetic model to use is non-trivial and researchers are likely to incorrectly select a mismatched model. While we naturally expect the application of a mismatched model to affect the accuracy of the subsequent single-molecule analysis, the exact effects of such mismatches on the inferred mechanistic details are unknown.

In this work, we investigate the effects of model-mechanism mismatches using a range of simulated datasets representative of commonly encountered types of heterogeneity in single-molecule experiments (Fig. S1 and Supplementary Information). To perform this analysis, we used a

comprehensive, open-source, and extensible analysis platform we have developed, called time-series Modeling, Analysis, and Visualization ENvironment (tMAVEN). tMAVEN is a software platform that, in addition to enabling the pre-processing of single-molecule time-series data, facilitates the interchangeable application of multiple, distinct kinetic models to the same single-molecule dataset. Beyond the above features, tMAVEN has also been designed to generate reproducible, publication-quality visualizations of experimental data and various modeling outcomes—all within a single computational pipeline.

Utilizing the broad range of modeling approaches implemented in tMAVEN, we show that only by exactly matching the number of free parameters (*i.e.*, the complexity) of the underlying molecular mechanism does a kinetic model accurately infer the often-heterogeneous biomolecular dynamics under investigation. As such, no kinetic modeling strategy is universally appropriate across all experimental contexts. Moreover, this requirement of mechanism matching is separate from, and more fundamental than, questions about the performance of specific algorithmic or software implementations of a model (29), or the strategies used to optimize the parameters of such a model (*e.g.*, maximum likelihood estimation, Bayesian inference, neural networks, *etc.*) (4). In the absence of a universal kinetic model, our study of where and how kinetic models fail in capturing the dynamics of heterogeneous, mesoscopic ensembles of single molecules can aid researchers in determining the optimal approach to quantifying the molecular details of their biomolecular system. Taken together with the capabilities that tMAVEN provides for applying multiple types of kinetic models to individual single-molecule datasets, our investigation facilitates the highly context-dependent kinetic modeling that is required for accurate single-molecule data analysis and, consequently, maximum extraction of biochemical and biophysical insight from single-molecule experiments.

Theory

Molecular mechanisms can be mapped onto a conformational free-energy landscape.

The molecular mechanism through which a biomolecule undergoes a structural rearrangement and/or binding and dissociation process may be explained using the theoretical framework of a conformational free-energy landscape (30). A conformational free-energy landscape is a low-

dimensional projection of the conformational space of a biomolecule onto the relevant reaction coordinates that characterize the specific process of interest, where each point represents the free energy of a specific set of biomolecular conformations. Thus, mapping a biomolecular system onto its corresponding conformational free-energy landscape provides a representation of the molecular mechanism underlying the dynamics of the system. Note that since most single-molecule techniques probe conformational changes, we have restricted our usage to conformational free-energy landscapes, but our discussion generalizes to other types of free-energy landscapes (*e.g.*, chemical free-energy landscapes that represent chemical transformations).

In this work, we have considered a hypothetical biomolecule that exists in two conformational states ('open' and 'closed') with transitions between these states that are governed by first-order kinetics (Fig. 1a). On the corresponding conformational free-energy landscape, the open and closed conformations are represented by two minima or 'wells.' The open and closed wells are separated by a region of higher free energy that serves as a barrier to the transition between these two conformational states along the relevant reaction coordinate (*i.e.*, the transition state). The height of this barrier (*i.e.*, the difference in free energy between the minima of the well and the maxima of the transition state) is a single independent parameter that controls the rate of transitions between the open and closed states (31). For our hypothetical biomolecule, the heights of the barriers for the open→closed and closed→open reactions comprise two independent parameters, representing two degrees of freedom (DoFs), that together define the molecular mechanism involved in the open⇌closed equilibrium. It is worth noting that these two DoFs may be parameterized in different, but equivalent, manners (*e.g.*, two barrier heights, the height of one barrier and the free-energy difference between the states, two rate constants, the equilibrium constant and the relaxation lifetime, *etc.*), but that two independent parameters are always required to describe this particular equilibrium. Together, these two independent parameters comprise the 'mechanistic' DoFs for our two-state system.

A final consideration in the mapping of a molecular mechanism onto a conformational free-energy landscape is how these conformational states are observed in a particular single-molecule experiment. In the ideal case above, we have represented these conformational states as smooth

minima on the conformational free-energy landscape that each generate a distinct signal. Real biomolecules, however, exist in a hierarchy of conformational states (*i.e.*, wells within wells) and consequently traverse 'rugged' conformational free energy landscapes (30, 32, 33). Depending upon the experimental time resolution used to probe the dynamics within this hierarchy, transitions between wells with relatively low barriers will occur so many times during one measurement that they will effectively average into a single state and not be observed in the experiment. In this manner, experimental details effectively determine the level of mechanistic detail that can be inferred from time-dependent, single-molecule data. Thus, the experimental distinguishability of the underlying molecular mechanism (*i.e.*, the number of wells that can be separately observed in the experiment), as scaled by the experimental specifics of the measurement (see below), generates additional 'observational' DoFs for the experimental molecular mechanism.

Hidden Markov models approximate the kinetic behavior of a single molecule.

Having described how a molecular mechanism can be mapped onto a conformational free-energy landscape, we now discuss the process of inferring the details of such mechanisms from single-molecule signal *vs.* time trajectories. This inference requires that we approximate the time-dependent changes of the observed signal using some kinetic model, such as an HMM. HMMs are probabilistic models that describe a dynamic phenomenon that cannot be directly observed and is thus 'hidden' in the data (6). For our purposes, the hidden processes are the conformational dynamics of the biomolecules that are indirectly reported on by single-molecule signal *vs*. time trajectories. In an HMM, every data point in the signal *vs.* time trajectory corresponds to a particular 'hidden' state (*e.g.,* a conformational state). However, the identities of these hidden states are unknown as the noisy signal is simply a proxy for the hidden states. For each of the possible hidden states, the HMM uses an 'emission' probability distribution to describe how that hidden state, as dictated by the experiment (see above), should appear in the noisy, observed signal (Figs. 2a and S2). The signal *vs.* time trajectory can then be approximated as a time-dependent sequence of hidden states (*i.e.,* a Markov chain). In a Markov chain, transitions between the states, or between the same state (*i.e.,* self-transitions), are stochastic and occur at random times. An HMM describes the probabilities of each of

these transitions between the hidden states by assuming that the 'transition probabilities' are time-independent and depend only on the identities of the initial and final states (*i.e.,* they represent a Markovian process exhibiting first-order kinetics (34)). Thus, the general approach of an HMM aligns well with our above understanding of molecular mechanisms. Indeed, by separating transition and emission parameters, HMMs provide a framework wherein both the dynamics of the biomolecule (*i.e.*, the mechanistic DoFs) and the experimental process by which these dynamics are observed (*i.e.*, the observational DoFs) may be independently modeled and mapped onto one another.

To explore this equivalence between molecular mechanisms and HMMs further, we consider the hypothetical biomolecule described above that transitions between an open and closed conformation. In a single-molecule experiment, this system will yield a signal for each state (*e.g.*, 0 for open and 1 for closed) and the resulting signal *vs.* time trajectory generated by such a molecule (Fig. 1a) may be described by an HMM (35, 36) (Fig. 2). Specifically, the emission probability distributions, which are often Gaussian distributions with mean μ and standard deviation σ, describe how the conformational states of the biomolecule manifest in the observed, experimental signal. Similarly, the rates of transition between the two conformations are described using transition probabilities between the two states, P_{01} and P_{10}, where P_{ij} is the transition probability from state i to state j between adjacent measurements in the signal *vs.* time trajectory.

HMMs have more parameters than the ones described above (Fig. 2a), but no more are required to describe a molecular system at equilibrium. The remaining parameters may be derived from constraints based on our prior knowledge of the experiment, and are thus not independent parameters. For our two-state example, the molecule can either remain in its current state (*i.e.*, undergo a self-transition) or transition to the other state. These two options comprise an exhaustive set of mutually exclusive events. Thus, the probabilities of self-transition, P_{00} and P_{11}, are

$$P_{00} = 1 - P_{01} \text{ , and}$$

$$P_{11} = 1 - P_{10}.$$

Similarly, for a system at chemical equilibrium, the probability that the molecule occupies either hidden state at the start of the experiment, π_0 and π_1, should be the steady-state probabilities

$$\pi_0 = \frac{P_{10}}{P_{01}+P_{10}}, \text{ and}$$

$$\pi_1 = \frac{P_{01}}{P_{01}+P_{10}}.$$

Therefore, while an HMM with two hidden states uses six parameters (*i.e.,* the four P_{ij} and two π_i) to model the observed dynamics, there are only two independent parameters if the system is at equilibrium (*i.e.*, P_{01} and P_{10}); we call these the 'transition' DoFs. In general, an HMM with K hidden states will have $K(K-1)$ transition DoFs (see Supplementary Information). Thus, a two-state HMM has two transition DoFs, which correctly matches the two mechanistic DoFs required to describe the conformational free energy landscape of our hypothetical two-state molecule (Fig. 1a).

We can similarly quantify the DoFs associated with the signal emitted from each hidden state in an HMM (*i.e.*, the 'emission' DoFs). While the emission DoFs scale with the number of hidden states, K, the exact number depends on the distribution chosen to represent the hidden states; ideally, this distribution correctly encapsulates the experimental details of signal measurement (*e.g.*, noise introduced by a detector). The number of emission DoFs in a K-state HMM is, thus, mK, where the proportionality factor, m, captures the dependence on the details of measurement. For single-molecule data analysis, the most standard emission distribution for a hidden state is a univariate (one-dimensional) Gaussian distribution. In this case, each hidden state is characterized by two emission DoFs ($m = 2$): one for the state mean, μ, and one for the state standard deviation, σ. In this work, we use univariate Gaussian distributions for the emissions of our hypothetical two-state biomolecule; emissions originating from the open state are modeled by a Gaussian distribution with parameters μ_0 and σ_0, and those from the closed state with μ_1 and σ_1 (Fig. 2a). Thus, our two-state HMM has four total emission DoFs, which correctly matches the four observational DoFs required to describe a typical detector-based measurement of two experimentally distinguishable states (Fig. 1a). However, emissions can be modeled by other sorts of distributions (37). Indeed, previous work on single-molecule fluorescence resonance energy transfer (smFRET) data analysis saw more accurate results by using a multivariate (two-dimensional) Gaussian emissions model for the fluorescence intensities of the donor and acceptor fluorophores (I_D and I_A, respectively) *vs.* a univariate Gaussian emission

model for the normalized FRET efficiency ($E_{FRET} = \frac{I_A}{I_D + I_A}$), because of the flexibility provided by the additional emission DoFs.

In our example above, we have discussed using a two-state HMM with four emission DoFs and two transition DoFs to model a two-state molecular mechanism. However, one can always employ more complex kinetic models with more emission and transition DoFs to explain the dynamics of a biomolecule (Figs. S2 and S3, and Supplementary Information). In the ideal scenario, the transition DoFs of the HMM used to analyze a single-molecule experiment should match the corresponding mechanistic DoFs of the molecular mechanism under investigation, while the emission DoFs should match the observational DoFs of the experiment. As described in the next section, however, experimental complications frequently result in mismatches between the transition DoFs and mechanistic DoFs and/or between the emission DoFs and observational DoFs.

Heterogeneity in a single-molecule ensemble reduces the apparent observational DoFs.

In the previous section, we discussed how the transition and emission DoFs of an HMM may exactly match those of the mechanism underlying the dynamics of an individual molecule. However, this is only true in ideal cases in which each conformational state of the molecule gives rise to a unique, distinguishable signal state. In many experimental scenarios, this is not the case. In particular, heterogeneity within a biomolecular system can significantly complicate this process of kinetic modeling. Molecular heterogeneity is a well-documented phenomenon that affects biomolecular dynamics (30) in ways that are observable using many types of single-molecule techniques (28, 35). While there exist many possible sources of heterogeneity, we define this phenomenon explicitly as any experimental circumstances which cause a change in the free-energy barrier(s) between conformational states and, therefore, in the rates of transitions between them. While this is a somewhat restrictive definition, it covers a large number of situations that are observed in single-molecule studies (38–40) and allows us to quantitatively describe the effects of heterogeneity on the mechanistic and observational DoFs used to describe a biomolecular system. Specifically, we will use this definition to discuss two types of heterogeneity in the section below.

We first consider the scenario where changes to the free-energy barriers are time-independent—a condition referred to as 'static heterogeneity.' For our hypothetical biomolecule undergoing open⇌closed transitions, this type of heterogeneity could occur if a fraction of the biomolecules has undergone an effectively irreversible chemical change, such as a post-transcriptional or post-translational modification, or even some form of chemical damage. As a result of this change, the affected biomolecules may still undergo open⇌closed transitions, but they do so at a different rate than the unaffected molecules. In this case, the entire collection of molecules probed in the experiment (*i.e.*, the experimental ensemble) consists of two subpopulations that undergo the same structural rearrangement, but at different rates—one slow and one fast (Fig. 1b). The conformational free-energy landscape of the molecules in this experimental ensemble can be understood as having split into two regions, each with their corresponding open- and closed-state wells, that are separated by a nearly infinitely high free-energy barrier across which transitions are not allowed on the experimental timescale. Thus, both of these subpopulations are separate two-well systems that are each characterized by two mechanistic DoFs (see above). Since the two subpopulations do not exchange, one additional mechanistic DoF is required to describe the fraction of each of the two subpopulations that comprise the ensemble. To match this mechanism, the corresponding kinetic model therefore requires five transition DoFs.

An additional complication arises when we consider that the chemical change causing the molecular heterogeneity might only alter the dynamics and not be drastic enough to alter the signal for each state in the two subpopulations. An HMM which fully captures this four-state system should, in the case of univariate Gaussian emissions (with $m = 2$), have eight total emission parameters—the four state means (μ_0^s, μ_1^s, μ_0^f, μ_1^f) and the four corresponding state standard deviations (σ_0^s, σ_1^s, σ_0^f, σ_1^f) where the superscript stands for <u>s</u>low- or <u>f</u>ast-transitioning molecules. However, due to the lack of differences between the signals for the states of the two subpopulations, *i.e.*, between μ_0^s and μ_0^f, and μ_1^s and μ_1^f (and σ_0^s and σ_0^f, and σ_1^s and σ_1^f), the number of apparent hidden states is reduced from four to two. The static heterogeneity in this case, therefore, leads to a mismatch between the apparent number of observational DoFs based on the observed signal states in the experimental dataset (four

DoFs) and the expected number of observational DoFs based on the number of states in the underlying mechanism (eight DoFs) (Fig. 1b).

A similar outcome is also seen in the case of 'dynamic heterogeneity,' where the changes in the free-energy barriers are time-dependent. Our hypothetical open⇌closed biomolecular system could exhibit dynamic heterogeneity if the process that creates the slow and fast subpopulations is a reversible change, such as a slow orthogonal conformational rearrangement or the binding of a secondary factor that allosterically modulates the open⇌closed transitions. Such dynamic heterogeneity can cause molecules of the experimental ensemble to transition between a slow and fast phase of the open⇌closed rearrangements (Fig. 1c). The corresponding conformational free-energy landscape has four wells distributed along two reaction coordinates—one for the open⇌closed transitions and the other for the orthogonal slow⇌fast phase transitions. For our hypothetical biomolecule, we have chosen that the free-energy barriers separating the slow and fast phases are higher than those separating the open and closed conformations in each phase. This causes the transition rates between the slow and fast phases to be smaller than those between the open and closed conformations and creates a hierarchical separation of timescale between the two molecular processes (20, 30, 32). The dynamics of each of the two phases are described by two mechanistic DoFs (see above). Unlike the case of static heterogeneity, however, two additional mechanistic DoFs are used to describe the transitions between the two phases. To match this mechanism, the appropriate kinetic model, therefore, requires six transition DoFs.

A commonly occurring complication for the case of dynamic heterogeneity is that many single-molecule experiments are designed to produce signal changes along only one of the reaction coordinates. As a result, orthogonal processes described by the other reaction coordinate(s) may not lead to a change in the observed signal. Just as we described for static heterogeneity, the expected number of emission DoFs for an HMM with univariate Gaussian emissions (with $m = 2$) that describes this four-well system should be eight. However, if the transition between the slow and fast phases is not directly observable in the signal *vs.* time trajectories, just like the case for static heterogeneity, the apparent number of hidden states in the experimental dataset is reduced to two. This, therefore,

causes a mismatch between the apparent number of observational DoFs based on the observed signal states in the experimental dataset (four DoFs) and the expected number of observational DoFs based on the number of states in the underlying mechanism (eight DoFs) (Fig. 1c).

Static and dynamic heterogeneities both cause discrepancies between the apparent number of observational DoFs based on the observed signal states and the mechanistically expected number of observational DoFs for a single-molecule signal *vs.* time dataset. These discrepancies obscure the underlying molecular mechanism and interfere with the process of kinetic modeling. For instance, in our examples of heterogeneity given above, a two-state HMM would match the apparent number of observational DoFs in the dataset, but would not have enough transition DoFs to match the molecular mechanism in the case of either static or dynamic heterogeneity. To be explicit, a two-state HMM has two transition DoFs, but the statically and dynamically heterogenous four-state systems that appear as two-state systems have five and six mechanistic DoFs, respectively. One approach to tackling such discrepancies is to select a model which has the correct number of transition DoFs (*i.e.,* a four-state model in this case) and constrain the emission DoFs to match the apparent observational DoFs of the dataset. Hierarchical HMMs, which can be thought of as trees of HMMs (Fig. S3a), are a class of such models that have been developed to tackle these sources of heterogeneity (20, 41, 42). For hierarchical HMMs, the applied constraints reduce the emission DoFs in the kinetic model relative to a standard HMM with the same number of hidden states (see Supplementary Information) (43). Additionally, other physico-chemical constraints such as detailed balance can be applied to standard HMMs to address similar issues (44).

In this work, we investigate how different types of HMMs perform when faced with simulated single-molecule datasets where the states of the underlying molecular mechanisms have been obscured by different types of heterogeneity. In addition to directly comparing the distributions of HMM parameters inferred from simulated datasets to the simulated ‘true’ parameter values, we have also compared the autocorrelation functions (ACFs) of kinetic models using the simulated and inferred HMM parameters (see Supplementary Information). ACFs capture the full dynamics of a model, both emission and transition, in a single analytical form (Fig. 2b). Thus, comparisons between the ACF of the simulated Markov chain and the ACF of an inferred HMM allow us to compare the performance

of kinetic models regardless of whether the models themselves are parametrized the same way. Thus, using ACFs, we can evaluate how well disparate kinetic models can capture the entire range of dynamic information contained within a single-molecule experimental dataset.

Results and Discussion

Development of an analysis platform that facilitates the application of multiple kinetic models.

The analysis of single-molecule kinetics experiments requires several computationally challenging steps, of which, kinetic modeling usually represents the penultimate or ultimate step. Before any kinetic modeling may be performed, however, the raw data must be pre-processed and curated to eliminate spurious signal *vs.* time trajectories and generate a dataset that represents the 'true' mesoscopic ensemble (45–47). Subsequently, the data, the applied models, and the dynamics underlying the data and quantified by the models are visualized and evaluated. At present, there exist multiple computational pipelines that are capable of independently pre-processing, modeling, and visualizing single-molecule datasets (29, 48). These pipelines mostly implement their own, usually HMM-based, kinetic models. However, because these implementations have been developed independently, at different times, and by different research groups, they are not necessarily interoperable. This is also true for implementations of kinetic models that serve as independent platforms and are not part of specific pipelines. Even if they can be used interchangeably, switching between pipelines and platforms still serves as a barrier, since the outputs, in terms of the resulting model parameters and visualizations, are not always comparable. To generally facilitate the analysis of all single-molecule experiments and to specifically enable our investigation of the abilities of different HMMs to accurately infer the kinetics of ensembles exhibiting various types of heterogeneity, we have developed tMAVEN. As a flexible, open-source platform, tMAVEN can be employed for the processing, modeling, analysis, and visualization of single-molecule time-series data from a variety of experimental techniques and contexts.

tMAVEN offers capabilities for pre-processing and curating raw experimental data, in addition to generating multiple plots to visualize both the experimental data and the applied models. In this work, however, we focus on its ability to apply multiple kinetic models, specifically HMMs,

interchangeably from a single platform. Crucially, the outputs for each model are standardized to similar sets of model parameters and associated outputs that describe both the signal and the dynamics of a dataset at the ensemble level. This allows us to construct distributions of model parameters generated from different HMMs when applied to the same set of simulated datasets and calculate their corresponding ACFs (see Supplementary Information). Comparing these parameter distributions and inferred ACFs to the true parameter values and corresponding true ACFs enabled us to not only investigate the abilities of different types of HMMs to infer the underlying kinetics present in various heterogeneous ensembles, but also to highlight the regimes in which certain HMMs failed to accurately capture the kinetics of a particular ensemble.

Global analysis allows accurate estimation of long-timescale dynamics from ensembles of short trajectories.

We first evaluated the practice of combining signal *vs.* time trajectories into a single model (*e.g.*, as in Refs. (24–27)) by investigating the analysis of simulated datasets of homogenous biomolecules (Fig. S1 and Supplementary Information). The simulated datasets were mesoscopic ensembles (*i.e.*, composed of 100-1000s of identical molecules) that exhibited Markovian dynamics and had experimentally optimal signal-to-noise ratios and kinetic rates. Altogether, these properties represent the most ideal situation that one could expect to encounter when using an HMM to extract biomolecular dynamics from a single-molecule experimental dataset. Using these datasets, we compared two common methods by which an ensemble-level HMM is inferred which we call the 'composite HMM' and the 'global HMM' methods. In the composite HMM approach, an HMM is separately inferred for each individual signal *vs.* time trajectory, and the results from these individual HMMs are then composited together to generate an ensemble-level kinetic model (see Supplementary Information for details). In the global HMM approach, all individual trajectories are assumed to describe molecules undergoing dynamics corresponding to the same free-energy landscape and are thus independent and identically distributed according to the same underlying HMM (see Supplementary Information for details).

Surprisingly, we find that the composite HMM approach yields a result that is non-trivially

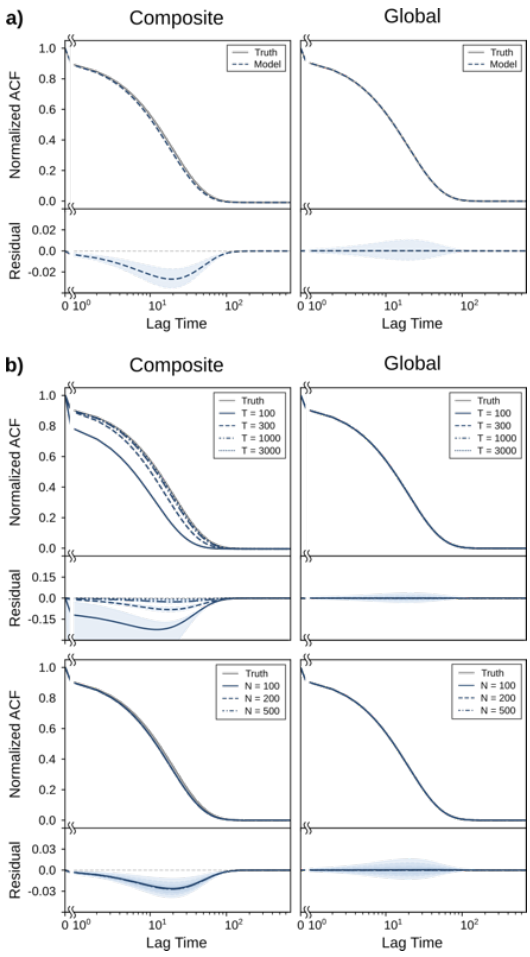

Figure 3. Comparisons of ACFS for homogeneous ensembles. (a) *(top) The true ACF for the homogenous dataset (solid black) along with the mean of the ACFs (dashed blue) calculated using HMMs inferred from 10 ensembles using composite HMMs (left) and global HMMs (right), along with (bottom) the corresponding mean (dashed blue) of the residuals of the inferred ACFs to the true ACF. The blue area denotes the region one standard deviation away from the mean. The grey dashed line corresponds to zero.* ***(b)*** *The true (black) and model (blue) ACFs, along with the means of the residuals (blue), inferred using composite (left) and global (right) HMMs for homogeneous datasets of signal vs. time trajectories of varying lengths (top) and varying numbers (bottom). The blue area denotes the region one standard deviation away from the mean. The grey dashed line corresponds to zero.*

different from the global HMM approach (Fig. 3a). The composite HMM appears to overestimate the transition probabilities for the mesoscopic ensemble, leading to faster decays in the mean inferred ACF when compared to the true ACF. This deviation is absent for the corresponding results from the global HMM analysis. While somewhat surprising, this result recapitulates the findings of previous investigations where a composite model was seen to be less accurate than a similar global model (37). We find that this overestimation of the transition probabilities is strongly correlated with the length of the trajectories comprising the single-molecule dataset (Fig. 3b and Fig. 4a). Interestingly, we also find that the overestimation is independent of the number of trajectories present in the dataset, with datasets containing fewer trajectories showing the same amount of deviation as datasets containing

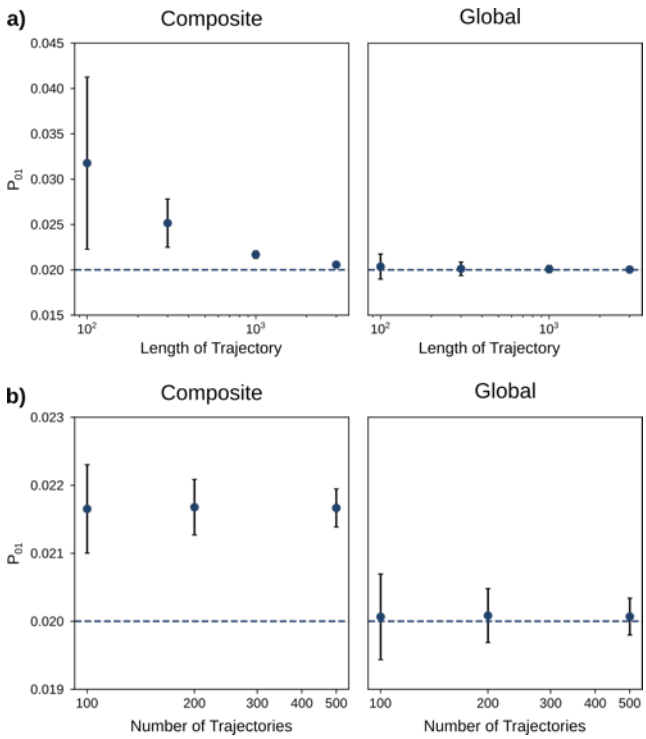

Figure 4. The effects of the lengths and number of trajectories in a mesoscopic ensemble on kinetic modeling. *The transition probabilities from the '0' state to the '1' observed states inferred using (left) composite HMMs and (right) global HMMs from homogenous datasets with* ***(a)*** *varying lengths of trajectories and* ***(b)*** *varying numbers of trajectories. The dashed line represents the true transition probability for the dataset. The transition probabilities from the '1' state to the '0' state follow the same trend (data not shown).*

more trajectories (Fig. 3b and Fig. 4b). In both the composite and global approaches, the precision of the estimation depends on the amount of data, both in terms of lengths and number of trajectories. However, the accuracy of the global approach was notably independent of either length or number of trajectories (Fig. 3b and Fig. 4).

This deviation in the accuracy of the composite HMM may be rationalized when we consider the effect of the length of the trajectories on the observed dynamics. When a single-molecule trajectory is recorded, the entire trajectory will be highly correlated with the initial data-points. This correlation will continue until enough time has passed for the molecule to reach the steady state and demonstrate that it exhibits stationary behavior (*i.e.*, for accurate kinetic modeling, much more time than the correlation time of the ACF must have passed). Thus, a single, very short trajectory will not appear to be in the correct steady state, which leads to the deviation that we see for the composite HMMs. To understand why this overestimation is not seen in the case of the global HMM, we analyzed the differences in the DoFs of these kinetic models. Both the composite and global HMMs have six total DoFs (two transition DoFs and four emission DoFs). However, for the composite HMM, individual

HMMs are inferred for each trajectory before being composited into a single model. During this first inference step, the composite HMM uses 6N total DoFs, where N corresponds to the number of trajectories. In this step, there may or may not be sufficient information to accurately infer all 6N DoFs, especially when the individual signal *vs.* time trajectories are short. Regardless of how accurately they have been inferred, these 6N DoFs are reduced to six total DoFs in the subsequent compositing step and one should expect that any inaccuracies in the individual HMMs are propagated into the composited kinetic model. On the other hand, the global HMM consists of just six total DoFs throughout, and thus simultaneously integrates information from the entire ensemble to infer the corresponding kinetic model. While a short individual trajectory may not appear to be in the correct steady state, a global HMM accesses the steady state represented within the entire ensemble of molecules, and thus can accurately infer the underlying kinetics.

These results suggest that increasing the number of signal *vs.* time trajectories in a dataset is only advantageous if a global HMM approach is employed to simultaneously incorporate information from multiple trajectories into the analysis. Furthermore, these results also suggest that the correct K-state global model can accurately infer the kinetics of the underlying molecular mechanism even from datasets composed of very short trajectories (Fig. 4). This ability comes from the fact that the constraints used in the global HMM (*i.e.*, all 6N DoFs are the same six DoFs) are applied throughout the entire inference process, unlike the composite HMM, in which the constraints are applied in a secondary step following the initial inference of the kinetic models. Interestingly, these constraints are equivalent to assuming that the ensemble under investigation is ergodic (*i.e.*, all states of the underlying conformational free-energy landscape are accessible in the experimental timescale) (Fig. 1a). Thus, multiple short trajectories can be analyzed together, as though they were all generated from the long-timescale behavior of a single molecule. Such ergodic constraints, while key to the process of kinetic modeling, arise solely from our prior understanding of the molecular mechanism under investigation. The use of ergodicity as a constraint demonstrates how mechanism-informed modeling can improve the accuracy with which the kinetics present in a single-molecule dataset may be inferred.

Constraints on emission distributions are required to accurately characterize static heterogeneity. Having evaluated the abilities of HMM-based modeling strategies to accurately estimate the kinetics of a homogeneous mesoscopic ensemble, we next sought to evaluate the performances of these methods in the presence of varying amounts of static heterogeneity (Fig. 1b). For this purpose, we simulated datasets where each signal *vs.* time trajectory had the same signal characteristics, but a subpopulation of trajectories had 'fast' open⇌closed transitions and the remaining had 'slow' open⇌closed transitions (Fig. S1). In this case, each trajectory is individually Markovian, because the transition probability depends only on the hidden state of the molecule (*i.e.*, open or closed). However, since there are two subpopulations of molecules and thus two variants of each hidden state (*i.e.,* slow or fast), the transition probability is not globally constant; it does not depend solely on the hidden state, rather it also depends on the subpopulation that the molecule belongs to. Furthermore, as is clear from the conformational free-energy landscape (Fig. 1b), all of the states in this mechanism are not accessible due to the presence of the near-infinite barrier separating the two subpopulations (see Theory). Therefore, the ensemble taken as a whole violates both Markovian and ergodic assumptions.

The effects of violating these assumptions can be clearly seen when we use a two-state HMM to analyze the dynamics of these simulated, statically heterogeneous ensembles. Since each individual trajectory within the ensemble is Markovian and has two distinct signal states, a two-state HMM (with four emission DoFs and two transition DoFs) is perfectly parameterized to exactly model each trajectory. Indeed, the individual trajectory-level HMMs that feed into the composite HMM do accurately capture the dynamics of both the slow- and fast-transitioning single molecules (Fig. 5). We further see that the distributions of the transition probabilities for these trajectory-level models clearly separate into two populations that scale linearly with the true proportions of the subpopulations. This signifies that information regarding the statically heterogeneous subpopulations may be captured at the trajectory level. However, we see that the ability to differentiate these subpopulations is completely lost for two-state models at the ensemble level (Fig. 5). Neither the composite HMM constructed from the individual trajectory-level HMMs (Fig. S4) nor the global HMM has enough complexity (*i.e.*, enough transition DoFs) to describe the kinetics of the two subpopulations. However, it is somewhat surprising that the ensemble behaviors of the composite and global HMMs are very similar. In both cases, the

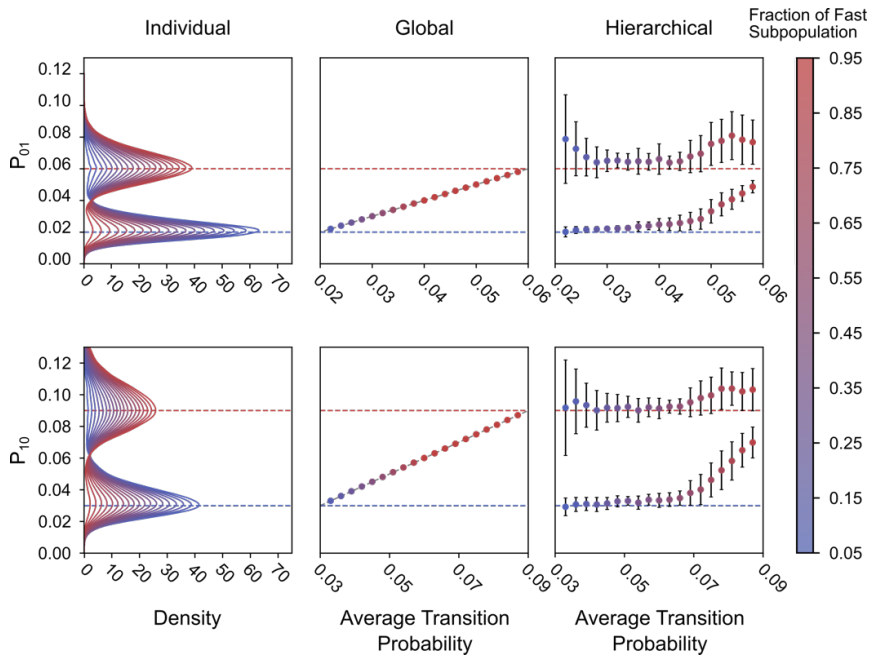

Figure 5. The effects of static heterogeneity on kinetic modeling. *(left) Kernel density estimated distributions of the transition probabilities for the observed 'open' and 'closed' states inferred from the individual trajectory-level HMMs for each molecule in mesoscopic ensembles with varying amounts of static heterogeneity. Dashed red and blue lines denote the transition probabilities from each state for the subpopulation of fast- and slow-transitioning molecules respectively. (middle) The ensemble-level transition probabilities for the observed states inferred using global HMMs as a function of the average transition probability of the observed states (calculated using the proportions of fast- and slow-transitioning molecules). The dashed grey line denotes identity. (right) The two transition probabilities for each observed state as inferred using a hierarchical HMM as a function of the average transition probability of the observed states calculated using the proportions of fast and slow subpopulations.*

inferred transition probabilities linearly scale with the subpopulation averaged transition probability for a specific state. Given that the individual HMMs themselves contain information on the heterogeneous subpopulations, this shows that both the composite and global two-state HMMs lose this information about the underlying heterogeneity due to ensemble averaging.

Interestingly, both the four-state composite and global HMMs, which have the same number of hidden states as observational DOFs of the molecular mechanism (Figs. 1b and S2), are unable to differentiate between the two subpopulations. This failure of the four-state models lies in the inference process, rather than in an intrinsic inability of this kinetic model to capture the ensemble static heterogeneity as in the case of the two-state models. Without constraints to match the apparent four observational DoFs of the ensemble (see Theory), the four-state HMMs (with eight emission DoFs) do not computationally separate the observed emissions into two hidden states with emission means of 0.0 and two hidden states with emission means of 1.0 (Fig. S5). The result is that, despite the flexibility provided by the additional DoFs, the inferred dynamics for the four-state HMMs are nearly

identical to those for the two-state HMMs when characterized by their ACFs (Fig. S6).

While the four-state models fail to correctly separate the observed emissions, the four-state hierarchical HMM (Fig. S3) does so by constraining the emission DoFs (Figs. 5 and S5). The four-state hierarchical HMM is the only ensemble kinetic model we used here that is capable of accurately capturing the statically heterogeneous subpopulations in the ensemble of trajectories. In particular, we see that it is most accurate for intermediate amounts of static heterogeneity where the fraction of fast subpopulation molecules in the ensemble is between 20% and 60%. Deviations from the true values of the transition probabilities outside of this regime can be explained by the hierarchical HMM assigning a non-zero probability to the rate of slow-to-fast interconversion at these levels of static heterogeneity. For our simulated ensembles exhibiting static heterogeneity, these probabilities should be zero, because the subpopulations do not interchange. However, since a constraint limiting this exchange is not explicitly applied for this hierarchical HMM, we observe that non-zero transition probabilities are inferred between the slow and fast subpopulations. Notably, this is the case even in the regime where the hierarchical HMM is the most accurate (Figs. 5 and S7). Instead of a near-infinite free-energy barrier separating the two subpopulations (Fig. 1b), the hierarchical HMM thus infers a finite barrier that may be traversed.

Altogether, these results highlight how essential it is to match the correct number of mechanistic and observational DoFs in a molecular mechanism when performing kinetic modeling. Here, static heterogeneity resulted in a reduction of the apparent number of hidden states from four to two (see Theory). A full, four-state HMM with eight emission DoFs was unable to compensate for this reduction. Being over-parameterized for the task, the four-state HMM was unable to correctly identify the two distinct emission states in the dataset. We found that only models with the correct number of emission DoFs were able to identify these states (*i.e.*, both the composite and global two-state HMMs, and the four-state hierarchical HMM). Additionally, none of the HMMs we used had the same number of transition DoFs as the underlying molecular mechanism (*i.e.*, five). The under-parametrized, two-state HMM with two transition DoFs yielded the correct ensemble-averaged transition probabilities but was unable to distinguish between the two subpopulations. On the other hand, the over-parametrized four-state hierarchical HMM with twelve transition DoFs (see

Supplementary Information) was able to distinguish between the two subpopulations. However, the seven additional transition DoFs, which correspond to transitions between the two subpopulations, led to inaccuracies when one of the subpopulations had a significantly greater fraction than the other.

While beyond the scope of this work, our results suggest a simple approach to developing a kinetic model that exactly matches the transition DoFs of the molecular mechanism in the case of static heterogeneity. Since the heterogeneous subpopulations are captured in the individual HMMs with the matching emission DoFs (Fig. 5), a clustering algorithm can be used to classify the transition probabilities of these individual HMMs into a certain number of subpopulations (49), the proportions of each cluster providing the additional transition DoFs; this is similar to, and could be done in conjunction with, the emission means-clustering approach we have used for the composite HMM. These results thus show how existing models like hierarchical HMMs (20) can be used to apply constraints that yield more accurate mathematical descriptions of static heterogeneity, while also serving as a guide for developing newer and better-performing kinetic models.

Kinetic models can accurately describe dynamic heterogeneity when underlying processes are separated across timescales.

Finally, we evaluated the abilities of HMM-based models to infer the kinetics of a mesoscopic ensemble exhibiting different levels of dynamic heterogeneity (Fig. 1c). For this purpose, we simulated datasets of trajectories in which molecules can interchange between a fast-transitioning and a slow-transitioning phase, both of which have the same emission properties (Fig. S1). Unlike in the case of static heterogeneity, each individual trajectory here appears to be non-Markovian. This is because, for each signal state (*i.e.,* the set of hidden states with the same emission distributions), the transition probabilities depend on whether the molecule is in the fast-transitioning phase or the slow-transitioning phase (Fig. 1c). Furthermore, the ensemble, in this case, is ergodic since each molecule can access all of the wells in the underlying conformational free-energy landscape.

The differences between static and dynamic heterogeneity become clear when we consider how an individual, two-state HMM models the trajectories containing dynamic heterogeneity before being used to create the two-state composite HMM (Fig. 6). As the total rate of interchange between

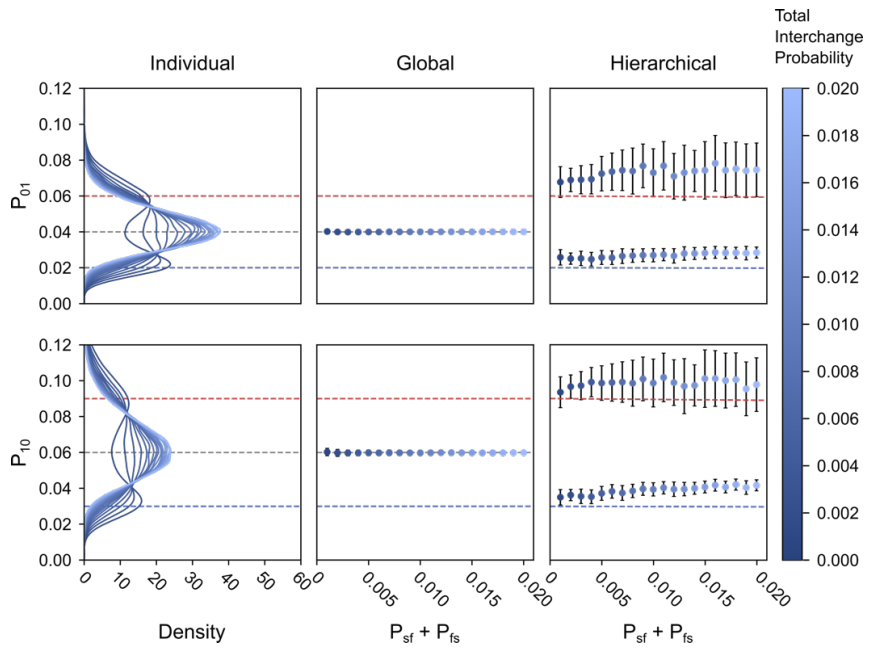

Figure 6. The effects of dynamic heterogeneity on kinetic modeling. *(left) Kernel density estimated distributions of the transition probabilities for the observed 'open' and 'closed' states inferred by the individual trajectory-level HMMs for each molecule in mesoscopic ensembles with varying total probability of transition between slow- and fast-transitioning phases (P_{sf} + P_{fs}). Dashed red and blue lines denote the transition probabilities of each state for the fast- and slow-transitioning phases respectively. The dashed grey line denotes the ensemble average transition probability of each observed state. (middle) The ensemble-level transition probabilities for the observed states inferred using global HMMs as a function of the total probability of transition between slow- and fast-transitioning phases. (right) The two transition probabilities for each observed state inferred using hierarchical HMMs as a function of the total probability of transition between slow- and fast-transitioning phases.*

the two kinetic phases (defined as $P_{sf} + P_{fs}$, where the subscripts stand for the slow and fast-transitioning phases) was increased, we found that the distributions of the individual transition probabilities between the open and closed states approached the ensemble average values for the slow and fast kinetic phases (Fig. 6). This asymptotic behavior highlights how the simulated trajectories more rapidly relax to the equilibrium behavior of the ensemble at higher rates of interchange. At lower rates of interchange, there are insufficient numbers of transitions between the two phases on the timescale of the experiment. As such, the individual trajectories, and thus the individual HMMs, more closely resemble the static heterogeneity case (Fig. 5). Indeed, this observation reflects how static heterogeneity can be thought of as a limiting case of dynamic heterogeneity where the transitions between the different kinetic phases occur much slower than the experimental timescale, making the free-energy barrier between the two effectively infinite within the experimental context (see Theory). Interestingly, we did not see an effect from varying the rates of interchange on the results of either the composite HMM (Fig. S8) or the global HMM (Fig. 6), because

the dynamical effects caused by the underlying dynamic heterogeneity are not able to be captured by either of these kinetic models.

As we saw in the case of static heterogeneity, analyzing simulated datasets exhibiting dynamic heterogeneity using full four-state HMMs that lacked any constraints on the emission distributions yielded the same ACFs as the analysis using two-state HMMs (Fig. S9). This was caused by the same mischaracterization of the emission distributions of the hidden states as was observed for the static case (Fig. S10). In contrast, we saw that analyzing the datasets with a four-state hierarchical HMM was much more successful and could accurately distinguish between the two kinetic phases (Figs. 6 and S10). Although, we note that increasing the rates of interchange did slightly decrease the accuracy of the transition probabilities between the observed signal states (Fig. 6). This inaccuracy appears to correlate with the inaccuracy in estimating the correct values of P_{sf} and P_{fs} themselves at higher rates of interchange (Fig. S11). Indeed, we saw that at lower rates of interchange, the inferred values of P_{sf} and P_{fs} themselves matched the simulated values, albeit with a slight systematic over-estimation, and they plateaued for fast interchange between the kinetic phases. We hypothesize that this is because at higher rates of interchange between the kinetic phases, the underlying molecular mechanism of the dynamic heterogeneity stops reflecting a hierarchical set of dynamics (Figs. 1c and S3a), and more closely resembles a system where transitions between all four states are equally temporally separated on the conformational free-energy landscape. Furthermore, the hierarchical HMM was able to model the statically heterogeneous ensembles with equal proportions of slow- and fast-transitioning molecules (Fig. 5) more accurately than the dynamic heterogeneous ensembles—which also have an equal average proportion of slow- and fast-transitioning molecules. Taking this statically heterogeneous dataset to represent the limit of infinitely slow interchange for the dynamic heterogeneity we have simulated here, our results thus suggest that a large separation of timescales between the directly observed molecular process (*i.e.*, the open⇌closed dynamics) and the ones responsible for causing the dynamic heterogeneity (*i.e.*, the slow⇌fast interchanges) is required for the hierarchical HMM to accurately infer kinetics from these datasets.

Once again, these results demonstrate that the number of transition and emission DoFs

serves as a good rule of thumb to determine how different HMMs will perform at inferring the underlying molecular mechanism of a biological system. When analyzing single-molecule datasets, thorough consideration must be given to how the HMM DoFs correspond to the specific mechanistic and observational DoFs of the molecular mechanism(s) of interest. For example, while ensembles exhibiting static heterogeneity and dynamic heterogeneity differ by a single transition DoF (five and six, respectively), considering the apparent ergodicity of the resulting signals means that we should expect that each individual trajectory is also characterized by two or six transition DoFs, respectively. An individual two-state HMM matches the transition DoFs in the individual trajectories in the static case. However, two-state HMMs are under-parameterized while four-state HMMs, including hierarchical HMMs, are over-parameterized in the dynamic case. Thus, the only general approach to accurately extract heterogeneous kinetics and molecular mechanisms from single-molecule datasets using HMMs is to customize the HMM being used by applying constraints that match the transition DoFs and emission DoFs to the mechanistic and observational DoFs of the underlying molecular mechanism. Such customization requires that a researcher know whether they are expecting static or dynamic heterogeneity *a priori*, so we do not expect a universally applicable method to be easily developed. Regardless of these difficulties, approaching kinetic modeling of single-molecule datasets in this manner will greatly expand the richness of information that can be extracted from single-molecule experiments, as well as the types of biomolecular processes that may be studied using these techniques. Altogether, our results show that developing more accurate kinetic modeling strategies will require thoughtful approaches to applying physico-chemically informed constraints to presently existing models. For instance, the hierarchical HMM does apply constraints to the emission DoFs, but at present is still over-parametrized with twelve transition DoFs compared to the six transition DoFs in the dynamic heterogeneity molecular mechanism. Fully matching all of the mechanistic and observational DoFs of an experimental molecular mechanism will enable kinetic models to better replicate the underlying dynamics they seek to describe.

Conclusions

In this work, we have shown how mismatches in the transition and emission DoFs of kinetic models

and the mechanistic and observational DOFs of molecular mechanisms, respectively, play a role in determining the accuracy of modeling single-molecule dynamics. However, matching mechanistic and observational DoFs is only the first step in mechanism-informed modeling. True mechanism-informed modeling requires an in-depth formulation of how a model captures the underlying physico-chemical properties of both the biomolecular system of interest and the detection process that yields the experimental signal *vs.* time trajectory (50). For example, previous work has shown that accounting for the Poisson-distributed counting statistics of photon emission and the effects of detector noise affects the accuracy of kinetic modeling (37). Similarly, accounting for the integration time of a detector allows for the modeling of sub-temporal-resolution kinetics (51–56).

Taken together, a universal single-molecule kinetic model (29) is unlikely to be appropriate for all experimental contexts. Indeed, in this work, we have focused on analyzing relatively ideal, simulated datasets by using just two hidden states, Gaussian noise, optimal signal-to-noise ratio, a lack of temporally averaged measurements, a large number of long trajectories, and a clear mechanistic separation of static and dynamic heterogeneity. Yet, even under these ideal conditions, we quickly reached the limits of current HMM-based modeling approaches. Real experimental datasets are much more complex, with more hidden states, non-trivial noisy emission distributions, and kinetics that may be slower or faster than is experimentally accessible. Additionally, in many cases, the information in real datasets is limited in terms of the length (*e.g.*, due to photobleaching) or number of trajectories (*e.g.,* due to experimental throughput). In contrast to our simulated datasets, experimental datasets may contain both static and dynamic heterogeneity that simultaneously originate from different sources. In such cases, it is not immediately obvious what the underlying mechanism for a particular biomolecular system should be.

In the absence of a universally optimal kinetic model, we propose that the most appropriate approach to kinetic modeling is to use multiple different kinetic models and compare the relative insight of their results to one another. Our investigation, which demonstrates how different types of models can be expected to perform in several cases of model-mechanism mismatch, can help guide this process. Of course, the seamless application and evaluation of multiple kinetic models to the same dataset is computationally and logistically challenging. This is especially true given that many

kinetic models and their corresponding software implementations were developed at different times, in different research groups, and using different programming languages. The development and use of tMAVEN, which can interchangeably apply and visualize several different types of HMMs and other kinetic models, alleviates many of these challenges. At present, we have packaged tMAVEN with Python implementations of ~15 kinetic models including upgraded versions of those used in previously published work (*e.g.*, HaMMy, vbFRET, ebFRET, hFRET, *etc.*) (6, 11, 12, 16, 20) as well as in the current work. As an open-source, extensible program, we have also equipped it with tools that allow other modeling approaches to be easily added. We envision tMAVEN as a platform that can be used by the entire single-molecule biophysics community for the future development of single-molecule data analysis software, including the further development of some of the modeling approaches we discuss here. By allowing developers to make use of the pre-existing user interface and data processing functions in tMAVEN, we hope to streamline future advances in the kinetic modeling of single-molecule datasets. This accessibility, paired with tMAVEN's user-friendly interface, lowers the barriers for both the use and development of kinetic modeling tools and thereby enables the type of mechanism-informed kinetic modeling that we envision will be so powerful for researchers using single-molecule experimental techniques to study biomolecular dynamics.

Author Contributions

All authors designed the research, analyzed the data, and wrote the manuscript.

Declaration of Interests

The authors declare no competing interests.

Acknowledgements

We thank Chris Wiggins, Jake Hofman, Jonathan Bronson, Jan-Willem van de Meent, Jason Hon, Bridget Huang, and Riley Gentry for many insightful discussions and collaborations regarding the kinetic modeling of single-molecule experimental data and Jason Hon, Bridget Huang, Sukjin Jang, Riley Gentry, Nina Michael, Qiongfang Zhang, Robin Shivnaraine, John Janetzko, Jonathan Deutsch,

and Qianyi Wu for feedback on the software features in tMAVEN. This work was supported by funds to R.L.G. from the National Institutes of Health (NIH) (R01 GM 084288, R01 GM 137608, R01 GM 128239, and R01 GM 136960) and the National Science Foundation (NSF) (CHE 2004016) as well as funds to C.D.K. from the NIH (Training Grant in Molecular Biophysics to Columbia University, T32 GM008281), the Department of Energy (DOE) (Office of Science Graduate Fellowship, DE-AC05-06OR23100), and the NSF (CHE 2137630).

Code and data availability

The open-source code for tMAVEN is implemented in Python and is freely available *via* a Git repository accessible at https://github.com/GonzalezBiophysicsLab/tmaven. The Python code used to generate the simulated ensembles, interface with tMAVEN, and programmatically analyze the simulated ensembles using different HMMs as well as subsequently collect the inferred results and generate the plots published in this article is also freely available *via* a separate Git repository available at https://github.com/GonzalezBiophysicsLab/tmaven_paper.

REFERENCES

1. Bustamante, C., Z. Bryant, and S.B. Smith. 2003. Ten years of tension: single-molecule DNA mechanics. *Nature*. 421:423–427.

2. Tinoco, I., and R.L. Gonzalez. 2011. Biological mechanisms, one molecule at a time. *Genes & Development*. 25:1205–1231.

3. MacDougall, D.D., J. Fei, and R.L. Gonzalez. 2011. Single-Molecule Fluorescence Resonance Energy Transfer Investigations of Ribosome-Catalyzed Protein Synthesis. In: Frank J, editor. Molecular Machines in Biology. Cambridge: Cambridge University Press. pp. 93–116.

4. Kinz-Thompson, C.D., K.K. Ray, and R.L. Gonzalez. 2021. Bayesian Inference: The Comprehensive Approach to Analyzing Single-Molecule Experiments. *Annual Review of Biophysics*. 50:191–208.

5. Du, C., and S.C. Kou. 2020. Statistical Methodology in Single-Molecule Experiments. *Statistical Science*. 35:75–91.

6. Bishop, C.M. 2006. Pattern recognition and machine learning. New York: Springer.

7. Chung, S. -h., J.B. Moore, L. Xia, L.S. Premkumar, and P.W. Gage. 1997. Characterization of single channel currents using digital signal processing techniques based on Hidden Markov Models. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*. 329:265–285.

8. Qin, F., A. Auerbach, and F. Sachs. 1997. Maximum likelihood estimation of aggregated Markov processes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*. 264:375–383.

9. Qin, F., A. Auerbach, and F. Sachs. 2000. A Direct Optimization Approach to Hidden Markov Modeling for Single Channel Kinetics. *Biophysical Journal*. 79:1915–1927.

10. Smith, D.A., and R.M. Simmons. 2001. Models of Motor-Assisted Transport of Intracellular Particles. *Biophysical Journal*. 80:45–68.

11. McKinney, S.A., C. Joo, and T. Ha. 2006. Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling. *Biophysical Journal*. 91:1941–1951.

12. Bronson, J.E., J. Fei, J.M. Hofman, R.L. Gonzalez, and C.H. Wiggins. 2009. Learning Rates and States from Biophysical Time Series: A Bayesian Approach to Model Selection and Single-Molecule FRET Data. *Biophysical Journal*. 97:3196–3205.

13. Kruithof, M., and J. van Noort. 2009. Hidden Markov Analysis of Nucleosome Unwrapping Under Force. *Biophysical Journal*. 96:3708–3715.

14. Okamoto, K., and Y. Sako. 2012. Variational Bayes Analysis of a Photon-Based Hidden Markov Model for Single-Molecule FRET Trajectories. *Biophysical Journal*. 103:1315–1324.

15. Greenfeld, M., D.S. Pavlichin, H. Mabuchi, and D. Herschlag. 2012. Single Molecule Analysis Research Tool (SMART): An Integrated Approach for Analyzing Single Molecule Data. *PLOS ONE*. 7:e30024.

16. van de Meent, J.-W., J.E. Bronson, C.H. Wiggins, and R.L. Gonzalez. 2014. Empirical Bayes Methods Enable Advanced Population-Level Analyses of Single-Molecule FRET Experiments. *Biophysical Journal*. 106:1327–1337.

17. Schmid, S., M. Götz, and T. Hugel. 2016. Single-Molecule Analysis beyond Dwell Times: Demonstration and Assessment in and out of Equilibrium. *Biophysical Journal*. 111:1375–1384.

18. Sgouralis, I., and S. Pressé. 2017. An Introduction to Infinite HMMs for Single-Molecule Data Analysis. *Biophysical Journal*. 112:2021–2029.

19. Lindén, M., and J. Elf. 2018. Variational Algorithms for Analyzing Noisy Multistate Diffusion Trajectories. *Biophysical Journal*. 115:276–282.

20. Hon, J., and R.L. Gonzalez. 2019. Bayesian-Estimated Hierarchical HMMs Enable Robust Analysis of Single-Molecule Kinetic Heterogeneity. *Biophysical Journal*. 116:1790–1802.

21. Karslake, J.D., E.D. Donarski, S.A. Shelby, L.M. Demey, V.J. DiRita, S.L. Veatch, and J.S. Biteen. 2020. SMAUG: Analyzing single-molecule tracks with nonparametric Bayesian statistics. *Methods*.

22. Kong, X., E. Nir, K. Hamadani, and S. Weiss. 2007. Photobleaching Pathways in Single-Molecule FRET Experiments. *J. Am. Chem. Soc.* 129:4643–4654.

23. Chandler, D. 1987. Introduction to modern statistical mechanics. New York: Oxford University Press.

24. Fei, J., P. Kosuri, D.D. MacDougall, and R.L. Gonzalez. 2008. Coupling of Ribosomal L1 Stalk and tRNA Dynamics during Translation Elongation. *Molecular Cell*. 30:348–359.

25. Cornish, P.V., D.N. Ermolenko, H.F. Noller, and T. Ha. 2008. Spontaneous Intersubunit Rotation in Single Ribosomes. *Molecular Cell*. 30:578–588.

26. Munro, J.B., R.B. Altman, C.-S. Tung, K.Y. Sanbonmatsu, and S.C. Blanchard. 2010. A fast dynamic mode of the EF-G-bound ribosome. *The EMBO Journal*. 29:770–781.

27. Chen, C., B. Stevens, J. Kaur, D. Cabral, H. Liu, Y. Wang, H. Zhang, G. Rosenblum, Z. Smilansky, Y.E. Goldman, and B.S. Cooperman. 2011. Single-Molecule Fluorescence Measurements of Ribosomal Translocation Dynamics. *Molecular Cell*. 42:367–377.

28. Kaufman, L.J. 2013. Heterogeneity in Single-Molecule Observables in the Study of Supercooled Liquids. *Annual Review of Physical Chemistry*. 64:177–200.

29. Götz, M., A. Barth, S.S.-R. Bohr, R. Börner, J. Chen, T. Cordes, D.A. Erie, C. Gebhardt, M.C.A.S. Hadzic, G.L. Hamilton, N.S. Hatzakis, T. Hugel, L. Kisley, D.C. Lamb, C. de Lannoy, C. Mahn, D. Dunukara, D. de Ridder, H. Sanabria, J. Schimpf, C.A.M. Seidel, R.K.O. Sigel, M.B. Sletfjerding, J. Thomsen, L. Vollmar, S. Wanninger, K.R. Weninger, P. Xu, and S. Schmid. 2022. A blind benchmark of analysis tools to infer kinetic rate constants from single-molecule FRET trajectories. *Nat Commun*. 13:5402.

30. Frauenfelder, H., S. Sligar, and P. Wolynes. 1991. The energy landscapes and motions of proteins. *Science*. 254:1598–1603.

31. Fersht, A. 2017. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. WORLD SCIENTIFIC.

32. Mustoe, A.M., C.L. Brooks, and H.M. Al-Hashimi. 2014. Hierarchy of RNA Functional Dynamics. *Annu. Rev. Biochem.* 83:441–466.

33. Herschlag, D., S. Bonilla, and N. Bisaria. 2018. The Story of RNA Folding, as Told in Epochs. *Cold Spring Harb Perspect Biol*. 10:a032433.

34. Kampen, N.G.V. 2007. Stochastic Processes in Physics and Chemistry. 3rd edition. Amsterdam ; Boston: North Holland.

35. Colquhoun, D., and A.G. Hawkes. 1982. On the Stochastic Properties of Bursts of Single Ion Channel Openings and of Clusters of Bursts. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*. 300:1–59.

36. Kinz-Thompson, C.D., N.A. Bailey, and R.L. Gonzalez. 2016. Chapter Seven - Precisely and Accurately Inferring Single-Molecule Rate Constants. In: Spies M, YR Chemla, editors. Methods in Enzymology. Academic Press. pp. 187–225.

37. Liu, Y., J. Park, K.A. Dahmen, Y.R. Chemla, and T. Ha. 2010. A Comparative Study of Multivariate and Univariate Hidden Markov Modelings in Time-Binned Single-Molecule FRET Data Analysis. *J. Phys. Chem. B*. 114:5386–5403.

38. Zhuang, X., L.E. Bartley, H.P. Babcock, R. Russell, T. Ha, D. Herschlag, and S. Chu. 2000. A Single-Molecule Study of RNA Catalysis and Folding. *Science*. 288:2048–2051.

39. English, B.P., W. Min, A.M. van Oijen, K.T. Lee, G. Luo, H. Sun, B.J. Cherayil, S.C. Kou, and X.S. Xie. 2006. Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nat Chem Biol*. 2:87–94.

40. Solomatin, S.V., M. Greenfeld, S. Chu, and D. Herschlag. 2010. Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature*. 463:681–684.

41. Fine, S., Y. Singer, and N. Tishby. 1998. The hierarchical hidden Markov model: Analysis and applications. *Machine Learning*. 32:41–62.

42. Wakabayashi, K., and T. Miura. 2012. Forward-backward activation algorithm for Hierarchical Hidden Markov Models. . pp. 1493–1501.

43. Hon, J., and R.L. Gonzalez. 2019. Bayesian-Estimated Hierarchical HMMs Enable Robust Analysis of Single-Molecule Kinetic Heterogeneity. *Biophysical Journal*. 116:1790–1802.

44. Zhang, Y., J. Jiao, and A.A. Rebane. 2016. Hidden Markov Modeling with Detailed Balance and Its Application to Single Protein Folding. *Biophysical Journal*. 111:2110–2124.

45. Thomsen, J., M.B. Sletfjerding, S.B. Jensen, S. Stella, B. Paul, M.G. Malle, G. Montoya, T.C. Petersen, and N.S. Hatzakis. 2020. DeepFRET, a software for rapid and automated single-molecule FRET data classification using deep learning. *eLife*. 9:e60404.

46. Li, J., L. Zhang, A. Johnson-Buck, and N.G. Walter. 2020. Automatic classification and segmentation of single-molecule fluorescence time traces with deep learning. *Nat Commun*. 11:5833.

47. de Lannoy, C.V., M. Filius, S.H. Kim, C. Joo, and D. de Ridder. 2021. FRETboard: Semisupervised classification of FRET traces. *Biophysical Journal*. 120:3253–3260.

48. Lerner, E., A. Barth, J. Hendrix, B. Ambrose, V. Birkedal, S.C. Blanchard, R. Börner, H. Sung Chung, T. Cordes, T.D. Craggs, A.A. Deniz, J. Diao, J. Fei, R.L. Gonzalez, I.V. Gopich, T. Ha, C.A. Hanke, G. Haran, N.S. Hatzakis, S. Hohng, S.-C. Hong, T. Hugel, A. Ingargiola, C. Joo, A.N. Kapanidis, H.D. Kim, T. Laurence, N.K. Lee, T.-H. Lee, E.A. Lemke, E. Margeat, J. Michaelis, X. Michalet, S. Myong, D. Nettels, T.-O. Peulen, E. Ploetz, Y. Razvag, N.C. Robb, B. Schuler, H. Soleimaninejad, C. Tang, R. Vafabakhsh, D.C. Lamb, C.A. Seidel, and S. Weiss. 2021. FRET-based dynamic structural biology: Challenges, perspectives and an appeal for open-science practices. *eLife*. 10:e60416.

49. Ghassempour, S., F. Girosi, and A. Maeder. 2014. Clustering Multivariate Time Series Using Hidden Markov Models. *International Journal of Environmental Research and Public Health*. 11:2741–2763.

50. Flomenbom, O., J. Klafter, and A. Szabo. 2005. What Can One Learn from Two-State Single-Molecule Trajectories? *Biophysical Journal*. 88:3780–3783.

51. Berezhkovskii, A.M., A. Szabo, and G.H. Weiss. 1999. Theory of single-molecule fluorescence spectroscopy of two-state systems. *The Journal of Chemical Physics*. 110:9145–9150.

52. Berezhkovskii, A.M., A. Szabo, and G.H. Weiss. 2000. Theory of the Fluorescence of Single Molecules Undergoing Multistate Conformational Dynamics. *J. Phys. Chem. B*. 104:3776–3780.

53. Gopich, I.V., and A. Szabo. 2010. FRET Efficiency Distributions of Multistate Single Molecules. *J. Phys. Chem. B*. 114:15221–15226.

54. Gopich, I.V., and A. Szabo. 2012. Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule FRET. *Proceedings of the National Academy of Sciences*. 109:7747–7752.

55. Kinz-Thompson, C.D., and R.L. Gonzalez. 2018. Increasing the Time Resolution of Single-Molecule Experiments with Bayesian Inference. *Biophysical Journal*. 114:289–300.

56. Kilic, Z., I. Sgouralis, and S. Pressé. 2021. Generalizing HMMs to Continuous Time for Fast Kinetics: Hidden Markov Jump Processes. *Biophysical Journal*. 120:409–423.