Unsupervised selection of optimal single-molecule time series idealization criterion

Argha Bandyopadhyay[1] and Marcel P. Goldschen-Ohm[1,*]
[1]Department of Neuroscience, University of Texas at Austin, Austin, Texas

ABSTRACT Single-molecule (SM) approaches have provided valuable mechanistic information on many biophysical systems. As technological advances lead to ever-larger data sets, tools for rapid analysis and identification of molecules exhibiting the behavior of interest are increasingly important. In many cases the underlying mechanism is unknown, making unsupervised techniques desirable. The divisive segmentation and clustering (DISC) algorithm is one such unsupervised method that idealizes noisy SM time series much faster than computationally intensive approaches without sacrificing accuracy. However, DISC relies on a user-selected objective criterion (OC) to guide its estimation of the ideal time series. Here, we explore how different OCs affect DISC's performance for data typical of SM fluorescence imaging experiments. We find that OCs differing in their penalty for model complexity each optimize DISC's performance for time series with different properties such as signal/noise and number of sample points. Using a machine learning approach, we generate a decision boundary that allows unsupervised selection of OCs based on the input time series to maximize performance for different types of data. This is particularly relevant for SM fluorescence data sets, which often have signal/noise near the derived decision boundary and include time series of nonuniform length because of stochastic bleaching. Our approach, AutoDISC, allows unsupervised per-molecule optimization of DISC, which will substantially assist in the rapid analysis of high-throughput SM data sets with noisy samples and nonuniform time windows.

SIGNIFICANCE The divisive segmentation and clustering (DISC) algorithm is a computationally efficient and accurate algorithm for idealizing noisy single-molecule time series. Although largely unsupervised, DISC requires user selection of a guiding objective criterion. We show that different criteria are optimal for different types of data. Critically, single-molecule fluorescence data sets typically exhibit stochastic variation in data properties from molecule to molecule. Here, we extend DISC to automate the optimal choice of criterion on a per-molecule basis. This advance, AutoDISC, provides a practical solution for automating analysis of high-throughput single-molecule fluorescence data sets increasingly common because of developments in camera and dye technologies.

INTRODUCTION

Recent advances in detector and imaging technologies have enabled increasingly high-throughput single-molecule (SM) data collection (1–3). For example, faster scientific complementary metal-oxide-semiconductor (sCMOS) cameras with larger chips coupled with photobleach-resistant fluorophores allow simultaneous recording of hundreds of molecules per field of view for extended durations (4–6). Tools for rapid unsupervised analysis of such large data sets are increasingly critical to avoid becoming the bottleneck for experiment progress (7,8).

Hidden Markov models (HMMs) are a widely successful approach for SM time series analysis (9–12). However, global analysis of large high-throughput SM data sets with a specific HMM is challenging for data sets with per-molecule variation in behavior or state emission and noise amplitudes. For example, such variation is typical of camera-based imaging modalities that often contain hundreds to thousands of molecules in which only a subset of molecules exhibits the behavior of interest. Furthermore, spatial nonuniformities in the optical pathways and/or illumination give rise to per-molecule variation in signal/noise, which can be exacerbated by motion within steeply varying excitation fields (e.g., total internal reflection fluorescence microscopy) or fluorophore photodynamics (13–15). Finally, HMMs require postulation of a molecular mechanism (i.e., specification of a model's states and the allowed transitions between them), which may be unknown a priori (9).

Although methods exist for selecting a model from a set of HMMs (16), this poses a heavy computational burden proportional to the size of the test model set. Unsupervised approaches for HMM model selection, e.g., infinite HMMs (17,18) and deep learning neural networks (19–21), automate the process of model identification but remain computationally expensive or require extensive training data sets before their use. This is not to say that one should not use an HMM if the data allow it. However, efficient unsupervised approaches that do not require postulation of a specific mechanism or pretraining on similar known data sets allow rapid screening and/or analysis of high-throughput data sets even in the presence of per-molecule variability as discussed above. Even in cases in which analysis with an HMM is ultimately desired, rapid screening with unsupervised approaches can aid in guiding the choice of experimental conditions or identifying subsets of molecules with particular behaviors of interest.

The divisive segmentation and clustering (DISC) algorithm is a largely unsupervised top-down approach for rapid idealization of noisy piecewise continuous time series typical of SM imaging experiments (13). For a given noisy time series, DISC estimates the underlying ideal noiseless time series consisting of discrete jumps between a finite number of distinct intensity levels. Both the jumps and the number of distinct intensity levels are determined in an unsupervised fashion and do not require the user to postulate a molecular mechanism before analysis. Compared with HMMs or change point analyses, DISC is orders of magnitude faster while maintaining state-of-the-art accuracy, precision, and recall. Lately, many deep learning techniques reliant on neural networks have been developed for unsupervised SM analysis (19,20). Unlike these approaches, DISC does not require extensive training data sets to guide its idealization, which simplifies its application to multiple different experimental regimes. Rather than relying on training data, DISC utilizes a user-specified objective criterion (OC) that weighs goodness of fit against the complexity of the ideal sequence to guide idealization (22). The OC represents a metric for unsupervised approaches to identify the simplest model that describes the noisy experimental data reasonably well while avoiding complex models that overfit the data. In addition to DISC, OCs have been widely applied in numerous SM analysis approaches, including HMM model selection and change point idealization (10,11,16,23).

Here, we show that different OCs optimize DISC's accuracy, precision, and recall for SM time series that differ in their signal/noise ratio (SNR) or number of sample points. This is crucial for SM fluorescence imaging experiments in which nonuniformity in the optics or illumination and stochastic bleaching of fluorophores result in variable SNRs and observation window durations across molecules even within a single field of view. To maximize the performance of DISC on such a data set, we use a machine learning technique to automate the per-molecule selection of the optimal OC. Critically, this automation makes DISC robust to both the scale and heterogeneity of data sets typical of increasingly common high-throughput SM imaging experiments (24,25).

MATERIALS AND METHODS

Simulations

All simulations and analyses were conducted in MATLAB version R2019a (The MathWorks, Natick, MA). SM time series were simulated as Markov chains of dwells in distinct molecular states governed by the average rate of transitions between states (11). The simulated Markov mechanisms are depicted in Fig. S1. Each state was assigned a mean observable intensity. For models of two-state dynamics at one, two, or four sites, state intensity levels were 0 and 1. For the three-state cyclic and linear models, state intensity levels were 0.2, 0.6, and 0.8 to reflect typical SM fluorescence resonance energy transfer (smFRET) observations. All simulations had a uniform sample frequency f_s, and transition rates between states were specified relative to the sample frequency. For the one-site, two-site, and four-site models, both forward and backward transitions rates were set to the same value, which ranged from 0.001 to 0.1 f_s. For the three-state models, the fast rate k_f ranged from 0.001 to 0.1 f_s, and the slow rate k_s was set to 0.3 k_f. The fast rate always described transitions between the higher-intensity levels. These prescriptions provide simulations that test performance on both equivalent and disparate rates within a given model.

The duration of each distinct dwell in a state was simulated with double precision. For discretized simulated intensities at a uniform sample frequency f_s, we assigned the weighted mean of the intensities for all dwells within a sample duration ($1/f_s$) with weights set to the relative fraction of the sample duration for each dwell. This procedure simulates integration of the signal throughout the sample period analogous to camera-based imaging strategies. This results in some samples having simulated mean intensities that are a weighted mean of the intensities of multiple individual states visited during that sample duration. For example, dwells shorter than a single sample period result in truncated observed intensity values for that sample. However, even at the fastest tested transition rate (0.1 f_s), such truncations were infrequent and had little impact on idealization with DISC. At slower transition rates, these subsample events were relatively rare. We note that at faster rates approaching the sample frequency, subsample events are frequent and result in an overall reduction in the apparent state intensity level separation (data not shown). DISC is likely to be inappropriate for such data, as it requires reasonable resolution of the state intensity levels.

When applicable, we implemented state intensity heterogeneity on a per-event basis by adding a small stochastic offset to the mean intensity of individual events. The offset was drawn from an exponential distribution with a mean set to 4% of the average separation between neighboring state intensity levels, similar to prior observations of such fluctuations in measures of molecular association (13).

To simulate noisy experimental data, Gaussian or Poisson noise was added to the noiseless intensity series described above. For Gaussian noise, the standard deviation (SD) of the added noise (σ) was set to the ratio of the average separation between neighboring state intensity levels (ΔI_{avg}) and the specified per-site SNR (SNR$_S$ ranging from 1 to 8) such that SNR$_S$ = $\Delta I_{avg}/\sigma$. For example, $\Delta I_{avg} = 1$ and $\sigma = 1/$SNR$_S$ for two-state dynamics between intensity levels 0 and 1, whereas $\Delta I_{avg} = 0.3$ and $\sigma = 0.3/$SNR$_s$ for three-state dynamics between intensity levels 0.2, 0.6, and 0.8.

For simulations of two-state dynamics at two or four independent sites, the variance of the added noise σ^2 was scaled for each event by the number of sites in their higher-intensity state. The reason for this procedure is to simulate experimental observations such as binding and unbinding of a fluorescent molecule at multiple sites where the noise is observed to

increase with the number of fluorophores in each diffraction-limited spot (13). The degree to which multiple sites give rise to additive noise depends on the amount of noise arising from the recording system versus the molecular activity, which will vary for each individual experimental setup. For Poisson-distributed noise, the variance naturally increases with increasing photon counts. In the low-photon regime discussed below, the variance approximately doubles upon transitioning from one to two sites adopting their high-intensity states, whereas this increase is ~1.5-fold in the higher-photon regime explored here. As there is no single value that will describe additive noise in all experiments, we chose to scale the noise by the number of "active" or "occupied" sites to provide examples of data that differ substantially from the simulations of all single-site models with uniform noise. Given that our results are qualitatively similar for both uniform and scaled noise, it is likely that our conclusions will remain relevant for scaled noise that falls between these extremes. Finally, we also provide simulations with Poisson noise (see below), which together with uniform and scaled noise models provide a breadth of examples across variable types of simulated noise.

For Poisson-distributed noise, we first converted simulated noiseless intensity traces as described above to number of photons per sample. We assumed a baseline mean photon count p per sample duration which defines a baseline SD of $\sigma_p = \sqrt{p}$. Simulated noiseless intensity traces for all tested models were scaled by the factor $SNR_S\sigma_p$ for a specified per-site SNR_S (ranging from 1 to 8) such that $SNR_S = \Delta I_{avg}/\sigma_p$ after scaling. Finally, the scaled trace was rounded to an integer number of photons, and the baseline photon count p was added. Poisson noise was then applied to each sample in the simulated photon series by drawing from a Poisson distribution with a mean set to the noiseless simulated photon count at that sample. To simulate data from experiments with lower and higher background photon counts, p was chosen to be either 10 or 50, respectively.

A consequence of nonuniform noise in simulations of multiple sites is that SNR is not constant throughout the series but decreases transiently from that specified for each single site during periods in which multiple sites simultaneously adopt their higher-intensity state. SNR is also not constant for the three-state models with uniform noise given that it is based on an average of two different level separations. To obtain a single metric to describe the overall SNR throughout a time series, we define the effective SNR of each series as the ratio of ΔI_{avg} to the SD of the residuals after subtracting the simulated ideal noiseless intensity series from that after adding noise. Note that this effective SNR can vary even for series simulated with the same per-site SNR_S and will, in general, be less than SNR_S for simulations of multiple sites because of additive noise.

For each model and each unique combination of SNR_S and transition rate (k or k_f), we simulated a total of 100,000 sample points split between 10 and 1000 time series depending on series lengths that ranged from 10,000 samples to 100 samples, respectively.

Idealization performance

Noisy simulated time series were idealized with DISC using one of the following OCs: Bayesian information criterion based on either the residual sum of squared errors (BIC_{RSS}) or a Gaussian mixture model (BIC_{GMM}), Akaike information criterion based on either the residual sum of squared errors (AIC_{RSS}) or a Gaussian mixture model (AIC_{GMM}), Hannan-Quinn information criterion (HQC_{GMM}), and minimal description length (MDL) (Eqs. 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11) (23,26–29). In principle, a different OC could be selected for the segmentation and clustering portions of the DISC algorithm. Here, we only explore the impact of selecting a single OC for all portions of the algorithm. The general format for each OC is a summation of two terms that balance goodness of fit with a penalty for overfitting noise with overly complex models (Eq. 1).

$$OC = \textit{fit error} + \textit{overfit penalty}, \qquad (1)$$

$$BIC_{RSS} = n_{pts}ln\left(\frac{RSS}{n_{pts}}\right) + (n_{transitions} + n_{states})ln(n_{pts}), \quad (2)$$

$$BIC_{GMM} = -2\,ln(\mathcal{L}) + (3n_{states} - 1)ln(n_{pts}), \qquad (3)$$

$$AIC_{RSS} = n_{pts}ln\left(\frac{RSS}{n_{pts}}\right) + 2(n_{transitions} + n_{states}), \qquad (4)$$

$$AIC_{GMM} = -2\,ln(\mathcal{L}) + 2(3n_{states} - 1), \qquad (5)$$

and

$$HQC_{GMM} = -2\,ln(\mathcal{L}) + 2(3n_{states} - 1)ln(ln(n_{pts})), \qquad (6)$$

where $\mathcal{L}$ is the likelihood for the estimated model defined as the product of likelihoods for each data point $y(t_i)$, each of which is described by a linear combination of likelihoods for each state's Gaussian emission distribution $\mathcal{N}$ with mean and SD μ_j and σ_j and mixing coefficient w_j (Eqs. 7 and 8) (30).

$$\mathcal{L} = \prod_{i=1}^{n_{pts}} \sum_{j=1}^{n_{states}} w_j\,\mathcal{N}\big(y(t_i)|\mu_j, \sigma_j\big) \qquad (7)$$

$$\mathcal{N}(y|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}exp\left(\frac{-(y - \mu)^2}{\sigma^2}\right) \qquad (8)$$

We define MDL as previously described in (23):

$$MDL = F + G, \qquad (9)$$

$$F = \frac{\sum_{i=1}^{n_{pts}}\big|y(t_i) - y_{fit}(t_i)\big|}{2\sigma}, \qquad (10)$$

and

$$G = \frac{n_{states}}{2}ln\left(\frac{1}{2\pi}\right) + (n_{states})ln\left(\frac{y_{max} - y_{min}}{\sigma}\right)$$
$$+ \frac{n_{transitions}}{2}ln(n_{pts}) + \frac{1}{2}\left[\sum_{i=1}^{n_{states}}ln(n_{pts\ in\ state\ i})\right. \qquad (11)$$
$$\left. + \sum_{j=1}^{n_{transitions}}ln\left(\frac{(\Delta y_{transition\ j})^2}{\sigma^2}\right)\right].$$

To quantify the quality of DISC's idealizations with different OCs, each event returned by DISC was determined to be either a true positive (TP), false positive (FP), or false negative (FN) based on the known simulated noiseless sequence. These same metrics were also used to compare the idealization of AutoDISC to two other unsupervised idealization approaches: step transition and state identification (STaSI) and AutoStepfinder (23,31). Because events were simulated with subsample timing and intensities integrated over the sample period, samples containing transitions between states exhibit intermediate intensities between the true state intensity levels. As these intermediate intensities are an artifact of the discretization process rather than true state intensities, we set such intermediates in the noiseless sequence to the intensity level associated with the state that was occupied for the largest portion of the sample period before

determining TP, FP, and FN events. Furthermore, to prevent slight numerical differences between simulated and idealized intensity levels or subtle offsets in event onset or offset from being construed as errors, TP events for OC comparisons were allowed to have intensities within $\pm 10\%$ of the known SD of the intensity level and onset or offset times within ± 3 samples of the known event timing. For all software comparisons, this intensity envelope was expanded to $\pm 25\%$ of the known SD of the intensity level to limit excessive error attribution to STaSI or AutoStepfinder so long as they are reasonably close to the true intensity levels. Events classified as FPs were either extraneous events or correct events with the wrong intensity. FNs were defined as missed events. For each idealization, accuracy, precision, and recall were calculated as general performance metrics ranging from 0 (worst) to 1 (best) (Eqs. 12, 13, and 14). The F1 score is a widely used overall metric for summarizing performance and ranges from 0 (worst) to 1 (best) (Eq. 15) (32).

$$Accuracy = \frac{TP}{TP + FP + FN} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (15)$$

OC penalty hyperparameter

To test the effects of a variable penalty term, we scaled the penalty terms of AIC_{GMM}, BIC_{RSS}, and MDL using a hyperparameter λ (see Eq. 16). For each unique combination of conditions (SNR_S and transition rate), short (300 samples) or long (3000 samples) simulated traces from the four-site and three-state cyclic models with added state intensity heterogeneity were idealized with DISC using AIC_{GMM}, BIC_{RSS}, or MDL for values of λ ranging from 0.001 to 100.

$$OC = fit\ error + \lambda(overfit\ penalty) \quad (16)$$

Linear decision boundary for AIC$_{GMM}$ vs. BIC$_{RSS}$

To generate a decision boundary for the optimal choice of either AIC_{GMM} or BIC_{RSS} based on the properties of the time series, we used MATLAB's *fitcsvm* function from the Statistics and Machine Learning Toolbox (The MathWorks). For each data series from each tested model, we computed a preference score for AIC_{GMM} vs. BIC_{RSS} based on their individual F1 scores (Eq. 17).

$$Preference\ for\ AIC_{GMM} = \frac{F1_{GMM}}{F1_{GMM} + F1_{RSS}} \quad (17)$$

For each model, these scores were binned in two dimensions according to the effective SNR and $\log_{10}$(number of samples) of their corresponding series. Each bin was labeled as either "AIC_{GMM} optimal" or "BIC_{RSS} optimal" according to its average preference score (AIC_{GMM} optimal for preference scores ≥ 0.5). These bin labels were used to train a support vector machine (SVM) classifier to determine a linear decision boundary in the two-dimensional space of effective SNR and number of samples.

Data availability statement

Software implementing AutoDISC is available at https://github.com/marcel-goldschen-ohm/AutoDISC.

RESULTS

Different OCs optimize DISC's performance under different experimental conditions

The DISC algorithm is a largely unsupervised top-down approach for rapid idealization of noisy piecewise continuous time series typical of SM imaging experiments (13). Although subsequently unsupervised, the algorithm initially requires a user-selected OC to guide its idealization. DISC has three main steps. 1) Divisive segmentation: starting with the intensity values of all data points in the time series assigned to a single cluster (the mean intensity of the time series), each cluster is recursively split into two child clusters until the selected OC is optimized. 2) Hierarchical agglomerative clustering: during segmentation, clusters in separate branches that should belong to the same intensity level may be assigned unique intensity states because of random fluctuations in the data. Agglomerative clustering starting with the segmented clusters from the previous step remerges segments with similar intensity distributions based on the selected OC. 3) Viterbi: the previous steps accurately identify intensity levels in the time series but do not provide a good description of the kinetics of transitions among those levels. To estimate event kinetics, the Viterbi algorithm is applied to determine the most likely sequence of transitions among the identified intensity levels. We have previously shown that DISC provides orders of magnitude faster computational speed while maintaining comparable accuracy, precision, and recall to other commonly used approaches (13). However, the impact of OC choice on the accuracy of DISC's idealization for different kinds of data has yet to be rigorously evaluated.

To explore the impact of OC choice on DISC's performance, we simulated noisy SM time series with known state sequences and evaluated the ability of DISC to identify the correct noiseless sequences using different OCs. We simulated data for several different kinds of mechanisms to evaluate a variety of typical SM data. Simulated mechanisms include two-state dynamics at one, two, or four independent sites similar to binding observations from colocalization SM fluorescence experiments and three-state linear or cyclic models with distinct state emissions typical of smFRET experiments (Fig. S1). Furthermore, for each model we varied simulation parameters such as sample length, SNR, and state transition rates to determine the impact of OC choice on data under different experimental conditions (see Materials and methods). Here, we define SNR as the ratio of the average intensity level separation (ΔI_{avg}) to the SD of the noise fluctuations (σ) (see Materials and methods). In many cases, real experimental SM fluorescence data

additionally contain heterogeneity in state emissions ([13],[33]). The source of this heterogeneity is uncertain but is likely caused by shifts of the molecule in the exponentially decaying excitation field or dye photodynamics ([14],[15]). Changes in observed dye brightness due to dye conformation, polarization orientation, partial quenching via electron transfer, and protein-induced fluorescence enhancement are commonly observed ([34]). To better reflect these real observations, we additionally included heterogeneous state intensities in some of our simulations (see Materials and methods).

Five OCs were initially tested: Bayesian information criterion based on either the residual sum of squared errors (BIC_{RSS}) or a Gaussian mixture model (BIC_{GMM}), Akaike information criterion (AIC_{GMM}) and Hannan-Quinn information criterion (HQC_{GMM}) based on a Gaussian mixture model, and MDL (Eqs. 2, 3, 5, 6, 7, 8, 9, 10, and 11) ([23],[26]–[29]). In each case, DISC's idealization performance on the noisy simulated data was evaluated using standard criteria of accuracy, precision, and recall (Eqs. 11, 12, and 13). An overall measure of this performance is summarized in the F1 score (0–1: worst to perfect), which combines both precision and recall in a single metric (Eq. 15) ([32]). For each OC, DISC's performance was primarily dependent on the SNR and number of samples in the time series (Figs. S2–S14). State transition rates had relatively less of an effect on performance except when rates were extreme (e.g., average rate approaching the sample rate).

For data with heterogeneous state intensities, there was no single OC that was optimal for all simulated conditions (Figs. 1 and S2–S6). Thus, the conditions of each time series dictate the optimal choice of OC. Across tested models, BIC_{RSS} is almost always the best choice for data with few sample points and low SNR. In some cases, MDL performs as well as BIC_{RSS} in this regime. However, MDL's relative performance decreases substantially for faster transition rates, especially as the underlying model complexity increases or the SNR decreases. In contrast, BIC_{RSS} exhibits either a smaller or negligible drop in relative performance at these faster rates. At the lowest SNRs, there is a trend for performance to decrease with increasing trace length and transition rates. In extreme cases, this occurs because of the idealization converging on a constant value that is the mean of the data. For slower rates and shorter traces, this is more likely to be true than for longer traces with faster transition rates, at which increasing trace length increases the odds of missing an actual transition because of poor SNR. Generally, the GMM-based OCs including BIC_{GMM}, AIC_{GMM}, and HQC_{GMM} tend to outperform BIC_{RSS} as the number of sample points and/or the SNR increases. For shorter time series and lower SNRs, AIC_{GMM} outperforms both BIC_{GMM} and HQC_{GMM}. However, BIC_{RSS} is better yet for these series. At higher SNRs at which BIC_{RSS} performs relatively poorly, all the GMM-based OCs perform similarly well. Thus, a simple choice between

BIC_{RSS} and one of the GMM-based OCs would provide an optimal solution in nearly all tested cases. Given that AIC_{GMM} has the overall best performance across tested data conditions, the remainder of our analysis focuses on AIC_{GMM} and BIC_{RSS}.

In general, BIC_{RSS} outperforms AIC_{GMM} for short traces with less than ~1000 samples and low SNRs of ~3 or less (Fig. 1), whereas AIC_{GMM} outperforms BIC_{RSS} for longer traces with more samples and/or higher SNR (Fig. 1). This is largely because AIC_{GMM} tends to underfit shorter series with low SNR, whereas BIC_{RSS} overfits longer series with high SNR. Each OC balances a goodness of fit term and a penalty term that attempts to prevent overfitting for overly complex models or sequences (Eq. 1). The observed behavior can be understood by examination of the penalty terms for AIC_{GMM} and BIC_{RSS} (Eqs. 2 and 5). Because the number of level transitions or change points is a stochastic function of the length of the time series, BIC_{RSS} will generally have a lower penalty than AIC_{GMM} for shorter traces with few transitions. For short traces with a low SNR, AIC_{GMM} is likely to underfit the number of distinct intensity levels because of the small amount of data and relatively large variation around each level, whereas the relatively smaller penalty for adding a level with BIC_{RSS} means that such underfitting is less likely (Fig. 1, *top*). However, the lower penalty for additional levels also means that BIC_{RSS} tends to split levels with heterogeneous event intensities into multiple sublevels because of a marginal increase in transitions per sublevel compared to a significant reduction in the sum of squared residuals. The amount of heterogeneity in each level will increase with the number of transitions into each level, which increases with the length of the series. Also, as SNR increases, such heterogeneity becomes more distinct. Thus, for longer traces with a high SNR, heterogeneous event intensities are likely to be overfitted by BIC_{RSS} in comparison to AIC_{GMM} (Fig. 1, *middle*). In the absence of heterogeneous state intensities, BIC_{RSS} either matches or outperforms AIC_{GMM} for most experimental conditions (Figs. S7 and S8). However, AIC_{GMM} continues to outperform BIC_{RSS} in a few cases for longer traces with high SNR and rapid transition rates (Fig. S7). This, again, can be understood by examination of the penalty terms for each OC (Eqs. 2 and 5). At faster rates, long traces will have large numbers of change points that result in a larger penalty term for BIC_{RSS} than for AIC_{GMM}. Because of this larger penalty, BIC_{RSS} is more prone to underfit these traces than AIC_{GMM}.

These trends in OC performance were consistent for both Gaussian and photon-based Poisson noise, suggesting that DISC can be applied to data with both types of noise despite the GMM-based OCs assuming Gaussian distributed noise (see Materials and methods) (Figs. S9–S16). As compared to simulations with Gaussian noise, the upper limit of performance for simulations with Poisson noise was a bit lower because of larger noise envelopes in high-intensity states.
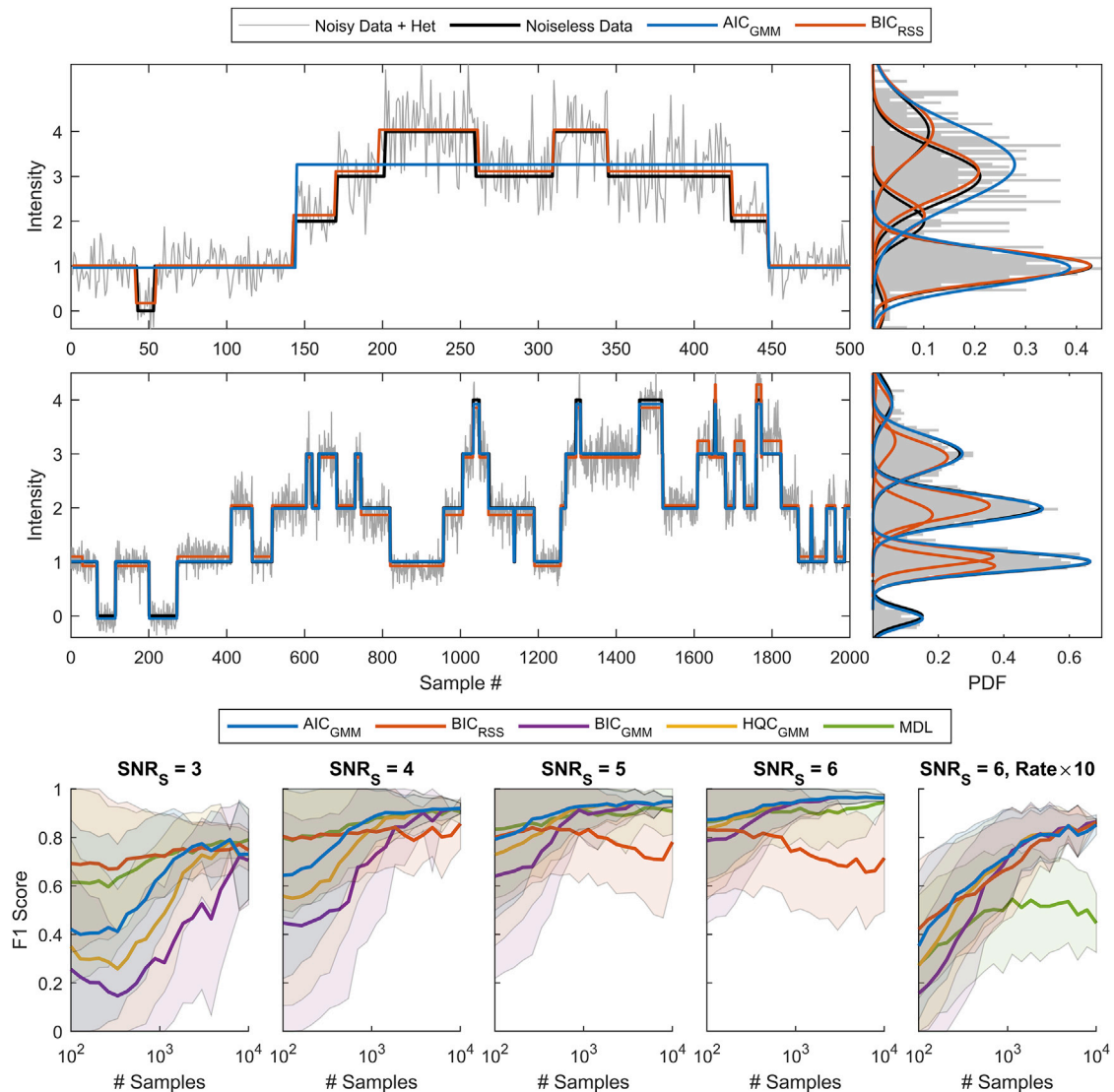
FIGURE 1 Optimal choice of AIC_{GMM} vs. BIC_{RSS} depends on experimental conditions. (*Top* and *middle*) Examples of simulated SM time series for the four-site model in Fig. S1 with (*gray*) and without (*black*) added noise and per-event state intensity heterogeneity. Traces are overlaid with DISC's idealization of the noisy data using either AIC_{GMM} (*blue*) or BIC_{RSS} (*orange*). Histograms of the noisy data are shown to the right overlaid with mixtures of Gaussians fitted to the data in each uniquely identified level for both the true noiseless series and the result of each idealization. AIC_{GMM} tends to underfit shorter series with lower per-site signal-to-noise ratios (SNR_S) (*top*, $SNR_S = 3$), whereas BIC_{RSS} tends to overfit longer series with higher SNR_S (*middle*, $SNR_S = 6$; notice that true levels are split into multiple nearby sublevels). Note that noise increases with observed intensity level such that the effective signal/ noise for events with multiple occupied sites will be less than SNR_S (see Materials and methods). (*Bottom*) Summary of performance for DISC's idealization of simulated noisy SM time series across a range of series lengths and SNR_S for $k = 0.005\,f_s$. Rate $\times$ 10 implies $k = 0.05\,f_s$. Mean (*line*) and SD (*shaded region*) for F1 scores (0–1: worst to perfect; Eq. 15) for 10–1000 simulated time series at each unique set of conditions (number of samples and SNR_S) idealized with DISC using AIC_{GMM}, BIC_{RSS}, BIC_{GMM}, HQC_{GMM}, or MDL (see Materials and methods). See Figs. S2–S18 for additional conditions and models. To see the figure in color, go online.

This effect was exacerbated in simulations with low photon counts.

The relative success of AIC_{GMM} compared to BIC_{GMM} suggested that an evaluation of AIC_{RSS} (Eq. 4) was necessary to compare to BIC_{RSS}. For two of the tested models, AIC_{RSS} performs consistently worse than BIC_{RSS} across the evaluated conditions (Figs. S17 and S18). AIC_{RSS} tends to overfit in comparison to BIC_{RSS} because of its relatively smaller penalty term for added states and/or transitions. For

longer series with higher SNRs in which BIC_{RSS} already tends to overfit, AIC_{RSS} only exacerbates the overfitting. Therefore, AIC_{RSS} was not included in further analyses. Analogously, AIC_{GMM} performs better than BIC_{GMM} because it is less prone to underfitting because of its smaller penalty term. We note, however, that there was a slight preference for BIC_{GMM} over AIC_{GMM} at the highest SNRs and rates for some models. In this regime, overfitting noise fluctuations and rapid flickery transitions is mitigated by the

higher penalty term for BIC_{GMM}. Thus, for data at higher SNRs and rates than tested here, it is possible that BIC_{GMM} would be a more optimal choice. Given that most fluorescence-based SM series fall in the range of tested conditions, we focus on AIC_{GMM}.

Optimal idealization with DISC across data conditions requires selection between at least two OCs such as AIC_{GMM} or BIC_{RSS} based on the number of samples and SNR in the time series. For SM fluorescence measurements in which fluorophore bleaching and spatial or temporal heterogeneity in excitation power can lead to stochastic variability in both the number of samples and SNR across molecules, this choice should optimally be made on a per-molecule basis.

Optimization of a variable penalty hyperparameter does not outperform a simple choice between either AIC_{GMM} or BIC_{RSS}

Given the dependence of DISC's performance on the penalty term of the chosen OC, we explored whether optimizing a variable penalty term for a given OC would further enhance performance. We introduced a hyperparameter λ that scales the penalty term (Eq. 16). Using AIC_{GMM}, BIC_{RSS}, or MDL as the framework for the OC, we scaled λ from 0.001 to 100 and chose the value of λ that maximized the OC's performance for a given set of experimental conditions (see Materials and methods) (Figs. S19–S22). Notably, all the GMM-based OCs differ solely in their scaling of the penalty term, meaning λ evaluation of AIC_{GMM} is inclusive of similar evaluations with BIC_{GMM} or HQC_{GMM}. This markedly improved performance for each OC in their respective problematic regimes. However, in nearly all conditions, AIC_{GMM}, BIC_{RSS}, or MDL optimized with this approach did not outperform the better of either AIC_{GMM} or BIC_{RSS} with $\lambda = 1$. Although it is theoretically possible to fit a curve for a given OC's optimal λ at every experimental condition to guarantee peak performance, the optimal λ is relatively model dependent (Figs. S19–S22). This curve-fitting approach would therefore overfit the underlying training models, limiting its utility for many SM experiments. In contrast, a simple selection between AIC_{GMM} or BIC_{RSS} is practical and generally sufficient for optimal performance on a per-molecule basis.

A decision boundary for selecting the optimal OC on a per-molecule basis

To automate the unsupervised optimal choice of either AIC_{GMM} or BIC_{RSS} on a per-molecule basis, we used a machine learning tool, the SVM, to identify a two-dimensional linear decision boundary based on number of samples and the effective SNR in each time series (see Materials and methods) (Fig. 2). For each unique pair of conditions (number of samples and effective SNR), Fig. 2 illustrates the relative preference for AIC_{GMM} over BIC_{RSS} based on F1 score
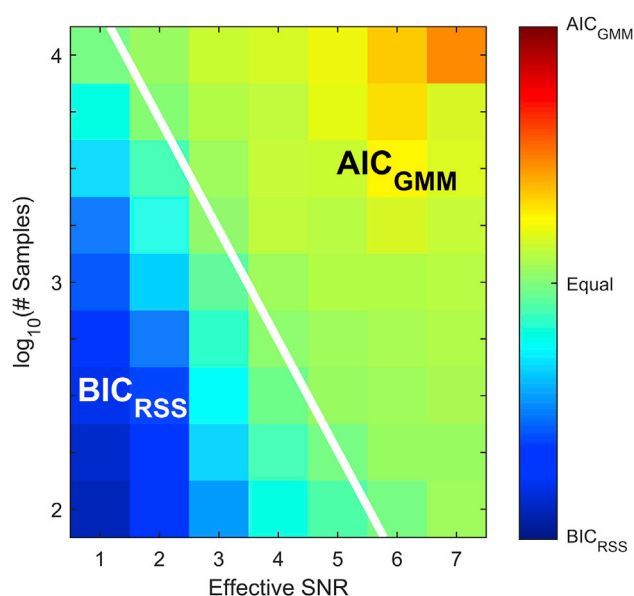

FIGURE 2   A linear decision boundary for the optimal choice of either AIC_{GMM} or BIC_{RSS}. Heatmap of the degree of preference for AIC_{GMM} vs. BIC_{RSS} from 0 to 1 (see Eq. 17) is shown for simulations of the four-site binding model shown in Fig. S1 with state intensity heterogeneity. Preference was determined from 1000 simulated time series for each unique pair of conditions (SNR_S and number of samples in the time series). Given nonuniform noise across intensity levels, preference is shown as a function of the effective SNR of each trace (see Materials and methods) rather than the average simulated SNR_S. The white line denotes the linear decision boundary determined by the SVM classifier. Boundaries for additional models with and without state intensity heterogeneity are shown in Figs. S23 and S24. To see the figure in color, go online.

(Eq. 17). The boundary between regimes of high BIC_{RSS}-preference and high AIC_{GMM}-preference is well described by a line, and more complex nonlinear boundaries were unnecessary. Constraining the boundary to a linear SVM also prevented overfitting small fluctuations in the training data set.

A similar linear decision boundary was determined for each model in Fig. S1 for simulations both with and without per-event state intensity heterogeneity (Figs. S23 and S24). The boundaries for each model were largely similar, with the primary difference being a slight shift to lower SNR for the simpler two-state models at one or two sites. Nonetheless, the same boundary determined for the more complex four-site and three-state models was also appropriate for the simpler one- and two-site models given the broad region of roughly equivalent preference for AIC_{GMM} or BIC_{RSS} in the simpler models. Here, we selected the boundary for the four-site binding model as appropriate for all tested models. This boundary is given by $\log_{10}(\#samples) = -0.49 SNR_{effective} + 4.69$. For a given model, the determined boundary was somewhat dependent on transition rate, with small shifts in intercept or slope across rates (Fig. S25). However, the overall cluster of boundaries across rates was highly consistent with each model's overall boundary. Although boundaries were highly similar for data both with and without per-event state

intensity heterogeneity, BIC_{RSS} was generally either optimal or equivalent to AIC_{GMM} across tested conditions in the absence of heterogeneity (Figs. S23 and S24). Thus, the need for an automated per-molecule selection of OC is much clearer when the data include such heterogeneity. Nonetheless, use of the identified boundary does not harm idealization when no per-event heterogeneity is observed. We cannot rule out that some mechanisms may give rise to time series with very different decision boundaries. However, the set of models and range of simulated conditions covers representative data for typical SM fluorescence experiments. Thus, the identified boundary is likely to be relevant for many SM imaging data sets or similar data.

Estimation of effective SNR for SM time series

To use this decision boundary, one must estimate the average separation in intensity levels (ΔI_{avg}) and the overall noise (σ) of each time series in an experimental data set. Therefore, we developed an unsupervised approach to estimate the effective SNR of an SM time series (Fig. 3). First, we apply DISC using BIC_{RSS} to generate an initial idealization that may overfit, but likely does not underfit, the data (Fig. 3, *top left*). The SD of the residuals between this initial fit and the noisy data is taken as our overall noise estimate σ (Fig. 3, *top right*). To estimate ΔI_{avg}, we first find the absolute value of the change in intensity levels at each change point i in the idealized series (ΔI_i). We then generate an array containing n_i copies of each step ΔI_i, where n_i is the number of samples between changepoint $i - 1$ and $i + 1$. This procedure ensures that changes in intensity are weighted according to the relative fraction of the series that they define. To avoid inclusion of small intensity changes due to BIC_{RSS} overfitting noise fluctuations, we only consider change points at which $\Delta I_i > 2\sigma$. The mean of the resulting array of ΔI_i-values is taken as our signal estimate ΔI_{avg} (Fig. 3, *bottom left*). Finally, the effective SNR is estimated as the ratio $\Delta I_{avg}/\sigma$.

We tested the SNR estimation approach described above by comparing it to the known effective SNR for each simulated series for each of the models in Fig. S1 with and without state intensity heterogeneity. In all cases, this estimation procedure provided a good approximation of the known effective SNR (Fig. 3, *bottom right*). SNR tended to be slightly overestimated for traces with low effective SNRs (around 2) and underestimated for traces with high effective SNRs (around 6–7). However, over- and underestimation was marginal and is unlikely to have a major impact on choice of OC for the identified decision boundary.

AutoDISC: a completely unsupervised workflow for optimal per-molecule performance

With the linear decision boundary for optimal choice of either AIC_{GMM} or BIC_{RSS} and the SNR estimation
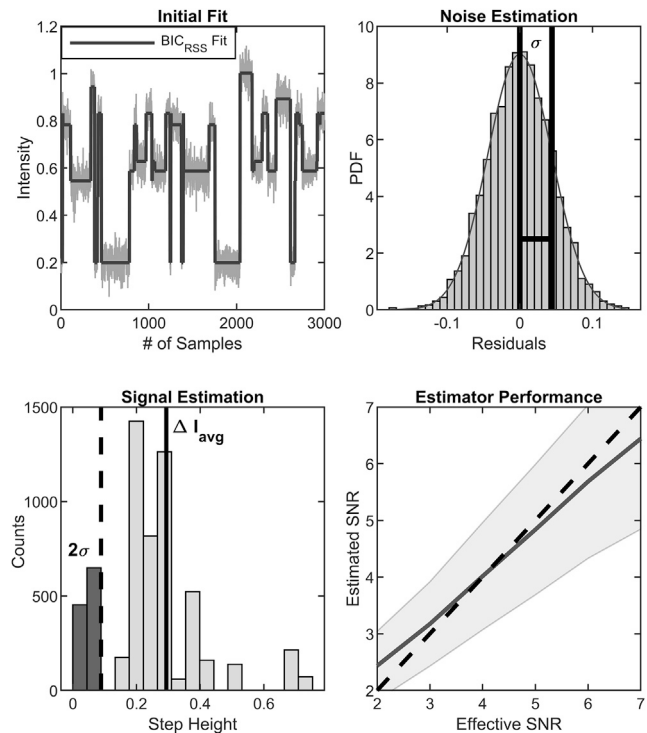

FIGURE 3 Estimation of the average intensity level separation and noise in a time series. (*Top left*) Idealization of a noisy time series using DISC with BIC_{RSS}. (*Top right*) Histogram of the residuals after subtracting the idealization from the noisy time series to the left. The SD σ of the distribution of residuals is taken as the estimate for the average noise in the observations. (*Bottom left*) Histogram of step heights from the idealized series above in which each step is duplicated according to the number of samples in the preceding and following dwells (see Results). The mean of the histogram (*solid line*) after discarding step heights less than 2σ (*dashed line*) is taken as the estimate for the average intensity level separation in the time series (ΔI_{avg}). Estimated SNR = $\Delta I_{avg}/\sigma$. (*Bottom right*) Estimated versus effective SNR (see Materials and methods) summarized across simulated time series for all tested models in Fig. S1 under all tested conditions (SNRs, number of sample points per trace, and average transition rates). Mean (solid line) and SD (shaded region) across time series. Dashed diagonal line indicated perfect estimation.

approach described above, we can now establish a completely unsupervised workflow, AutoDISC, for optimizing the DISC algorithm on a per-molecule basis (Fig. 4). The workflow for an individual time series is as follows: 1) apply DISC with BIC_{RSS} to idealize the time series. 2) Estimate the effective SNR of the series. 3) Based on the number of samples and the estimated effective SNR, select the optimal OC based on the linear decision boundary between AIC_{GMM} and BIC_{RSS}. 4) If AIC_{GMM} is optimal, apply DISC with AIC_{GMM} to idealize the time series. Otherwise, use the initial idealization. This workflow allows unsupervised per-molecule idealization of noisy SM data with stochastic variation in series duration, SNR, and event intensities typical of many SM fluorescence imaging data sets.
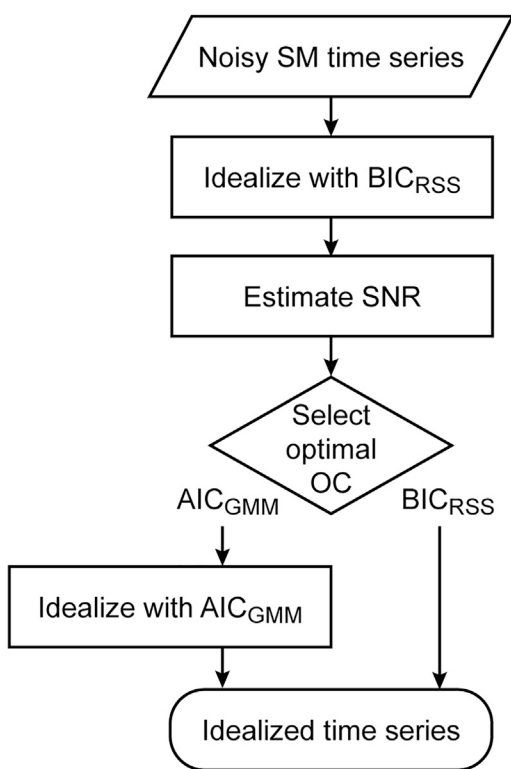
FIGURE 4   Workflow for unsupervised per-molecule optimal idealization by DISC.

Comparison to other unsupervised idealization methods

AutoDISC was benchmarked by comparison to the unsupervised idealization methods STaSI and AutoStepfinder (Fig. 5) (23,31). Idealization of simulated series for two-state dynamics at one or four sites and three-state cyclic dynamics was evaluated using either AutoDISC, STaSI, or AutoStepfinder across a range of conditions (SNR_S, number of samples per trace, and transition rates) according to F1 score. AutoDISC either outperformed or was equivalent to all other methods across models and nearly all conditions tested both with and without per-event state intensity heterogeneity (Figs. 5 and S26–S31).

The best competitor, STaSI, had similar or even in a few cases slightly better F1 scores to AutoDISC for shorter series with low SNR_S and slow transition rates but was much less favorable for longer traces with higher SNR_S and faster transition rates, consistent with previous observations based on BIC_{GMM} (13). This performance can be understood based on two observations. First, STaSI tends to miss many short-lived events, which is exacerbated under conditions of faster dynamics. In cases in which all sojourns in an intensity level are brief, STaSI may ignore the level completely (Fig. 5, middle). Second, in the presence of per-event state intensity heterogeneity, STaSI assigns numerous unique levels for each of these events rather than a single global level as correctly identified by AutoDISC (Fig. 5, top). This results in overly complex idealized series with extra states and spurious intrastate transitions that challenge automated screening or analysis.

AutoStepfinder performed similarly to AutoDISC and STaSI for shorter series with slow dynamics. However, F1 scores for AutoStepfinder were consistently lower than those for either STaSI or AutoDISC for nearly all other conditions tested. The relatively poorer performance of AutoStepfinder was primarily the result of overfitting noise fluctuations, thereby resulting in the misidentification of short-lived dwells in nonexistent states (Fig. 5, top and middle). Such overfitting occurred in both the presence and absence of per-event state intensity heterogeneity, suggesting that its penalty for additional states is often too lenient. In the presence of per-event state intensity heterogeneity, AutoStepfinder behaves similarly to STaSI in that it assigns numerous unique levels for each event rather than identifying the true single global level.

AutoDISC performs well at faster transition rates regardless of heterogeneity for more complex models with F1 scores substantially larger than the next-best tested method as series length increases. By optimally selecting the appropriate OC, equivalent or better performance is also achieved under most other conditions. Furthermore, AutoDISC is the only tested method that is consistently robust to per-event state intensity heterogeneity. Given the presence of such features in many SM fluorescence experiments, AutoDISC is an ideal choice for automated analysis or screening of these data sets.

DISCUSSION

The performance of the DISC algorithm with each of six different OCs (BIC_{GMM}, BIC_{RSS}, AIC_{GMM}, AIC_{RSS}, HQC_{GMM}, and MDL) was investigated for simulations reflecting typical SM fluorescence time series observations under variable experimental conditions. Each OC differentially impacts DISC's performance for different conditions, with the SNR and number of samples in a time series being the primary determinants for the optimal OC. For nearly all conditions tested, the OC with the best F1 score was either BIC_{RSS} or one of the GMM-based OCs such as AIC_{GMM}. BIC_{RSS} is typically optimal for time series with few samples and/or a low SNR, whereas AIC_{GMM} is typically optimal for time series with a thousand or more samples and high SNR. This difference in behavior can largely be attributed to the different penalty terms for each OC and the GMM likelihood distributions. However, optimization of a variable penalty term does not generally outperform either AIC_{GMM} or BIC_{RSS}, suggesting that a simple choice between the two OCs is sufficient for optimal performance by DISC in nearly all cases. Examination of the optimal OC under variable conditions (i.e., SNR and number of samples) reveals a general linear decision boundary that can be used to select the optimal OC for a
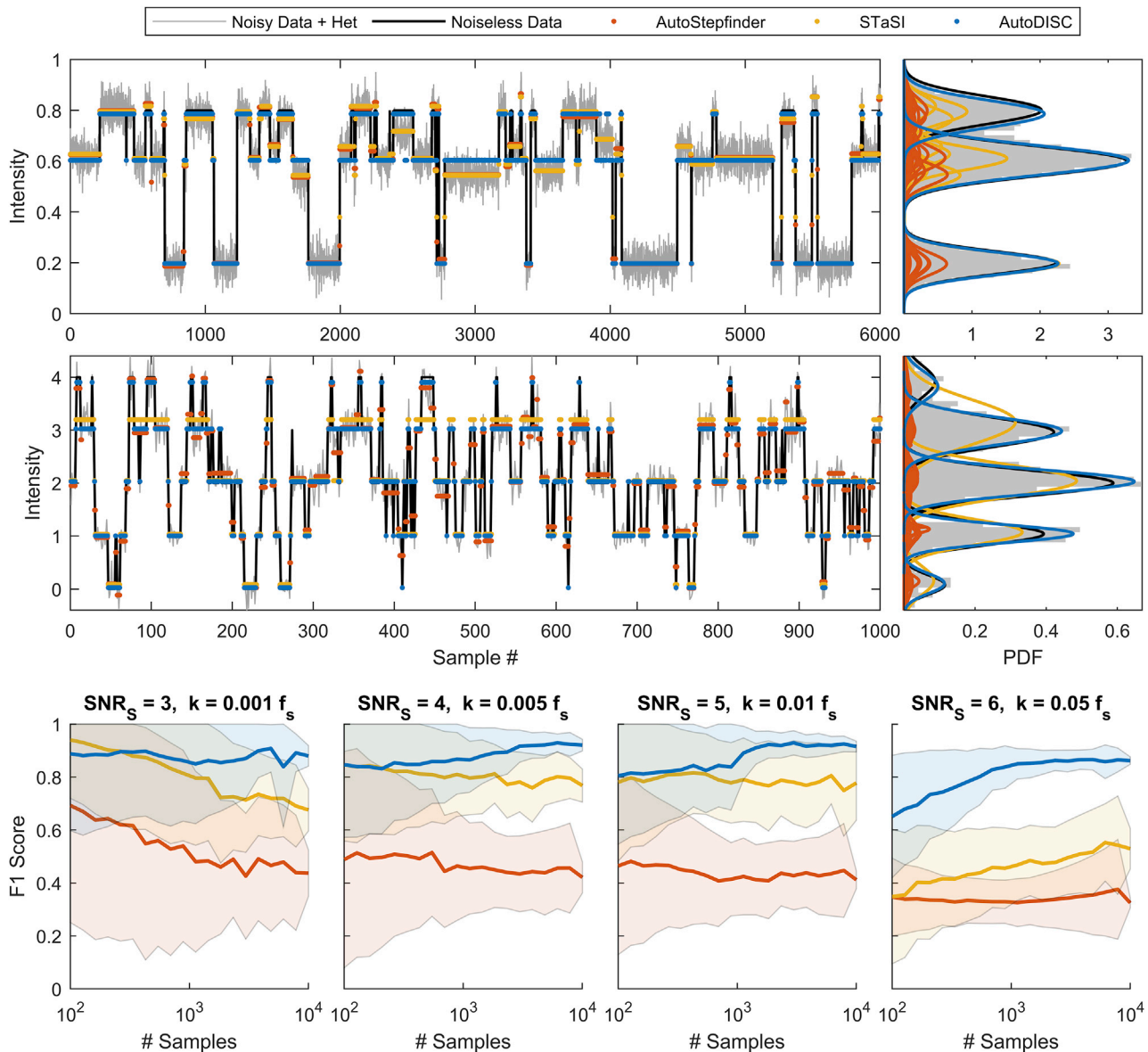
FIGURE 5 Comparison of AutoDISC, STaSI, and AutoStepfinder. (*Top* and *middle*) Examples of simulated SM time series for the three-state cyclic model (*top*; $SNR_S = 7$, $k_f = 0.01 f_s$) and four-site model (*middle*; $SNR_S = 6$, $k = 0.05 f_s$) shown in Fig. S1 with (*gray*) and without (*black*) added noise and per-event state intensity heterogeneity. Simulated traces are overlaid with idealizations from AutoDISC (*blue*), STaSI (*yellow*), and AutoStepfinder (*orange*) (23,31). Histograms of the noisy data are shown to the right of each series overlaid with mixtures of Gaussians fitted to the data in each uniquely identified level for both the true noiseless series and the idealizations. (*Bottom*) Summary of each method's idealization performance for varying series length, SNR_S, and rate constant k using the four-site model. Mean (*line*) and SD (*shaded region*) for F1 scores (0–1: worst to perfect; Eq. 15) for 10–1000 simulated time series at each unique set of conditions (number of samples and SNR_S) are given. See Figs. S26–S31 for additional conditions and models. To see the figure in color, go online.

given set of conditions. We further developed an estimator for the effective SNR of experimental SM time series that relates intensity level separation and Gaussian noise. Together, these developments establish AutoDISC: a completely unsupervised workflow for the DISC algorithm that automates selection of the optimal OC on a per-molecule basis. AutoDISC performs favorably compared to other contemporary unsupervised idealization approaches on time series typical of SM fluorescence experiments.

Fluorescence imaging is a common approach for massively parallel observations of SM time series (1,3,6,15). Critically, effective SNRs in SM fluorescence data sets are often near our derived decision boundary, at which optimal choice of OC is most impactful. Furthermore, stochastic fluorophore bleaching and nonuniformities in the optical pathways result in per-molecule variation in properties such as the observation time window and the ratio of the fluorescence intensity separation between distinct states to the noise within each state

(24,25). In addition to uncertainty regarding the underlying mechanism, per-molecule variation in state intensity levels challenges analysis with approaches such as HMMs having user-defined state intensity distributions. Furthermore, for such large data sets, it is often the case that only a subset of molecules exhibits the behavior of interest. It is therefore ideal to avoid computationally intensive analyses such as HMMs on the potentially large fraction of nonrelevant molecules. Rapid, unsupervised approaches are thus ideal in many cases, either as a complete analysis or for prescreening to identify a subset of molecules for further analysis.

DISC is a recently developed, largely unsupervised approach for idealization of noisy SM time series that automates detection of the intensity levels within each individual time series without requiring postulation of a specific mechanistic model (13). When possible, the additional constraint imposed by global optimization of a specific molecular mechanism or a singular set of state intensity distributions should be preferred. However, this approach is not appropriate for unknown mechanisms or experimental data with per-molecule variability in state intensity emissions. DISC not only handles these data efficiently but is also robust to within-state intensity fluctuations that can arise from changing orientation in polarized and/or exponentially decaying excitation fields or dye photodynamics (14,15,34). Furthermore, DISC is orders of magnitude faster than HMM methods while maintaining state-of-the-art accuracy, precision, and recall (13). However, DISC relies on a user-specified OC to guide its idealization. Here, we show that the optimal choice of OC depends primarily on both the number of sample points and the signal and noise properties in each time series. Thus, maximizing the performance of DISC on experimental data sets with variability in these parameters requires selection of the optimal OC on a per-molecule basis.

By automating the per-molecule choice of OC, we develop a fully unsupervised workflow, AutoDISC, that maximizes DISC's performance across data sets with stochastic variation in observation conditions. AutoDISC either outperforms or matches two other unsupervised idealization methods (STaSI and AutoStepfinder) across nearly all conditions tested. One of the advantages of AutoDISC over these other methods is its ability to robustly identify global intensity levels in the presence of per-event intensity heterogeneity. For data that include such heterogeneity, this is highly beneficial for automated identification and analysis of the dynamics of interest. However, if one's goal is to identify the observed local intensity of each event rather than the global state level, then an alternate method such as STaSI or AutoStepfinder may be preferred. We also note that we only evaluated data with up to five unique states similar to smFRET experiments, whereas some observations such as stepwise motion of molecular motors or extension in molecular tweezers can give rise to many more levels (35,36). In principle, DISC should continue to perform well with many additional levels, provided they are reasonably well separated. Another advantage of AutoDISC is its relatively better performance for series with faster dynamics, data that have historically been reliant on computationally expensive approaches such as HMMs (37,38).

As compared to DISC, AutoDISC's computational cost is either marginal or at most doubled owing to potentially rerunning DISC with AIC_{GMM}. Thus, this workflow does not negate the benefit of DISC's computational speed (13), making it an attractive approach for large high-throughput data sets. Given the prevalence of per-molecule stochastic variation in SM fluorescence observations and effective SNRs that are often near our derived decision boundary, AutoDISC provides an immediately useful tool to both optimize and speed exploration and analysis of these and similar SM data sets.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2021.08.045.

AUTHOR CONTRIBUTIONS

A.B. performed all simulations and analysis and contributed to writing the manuscript. M.P.G.-O. conceived, designed, and oversaw the project and contributed to writing of the manuscript.

REFERENCES

1. Chen, J., R. V. Dalal, …, J. D. Puglisi. 2014. High-throughput platform for real-time monitoring of biological processes by multicolor single-molecule fluorescence. *Proc. Natl. Acad. Sci. USA.* 111:664–669.

2. Soltani, M., J. Lin, …, M. D. Wang. 2014. Nanophotonic trapping for precise manipulation of biomolecular arrays. *Nat. Nanotechnol.* 9:448–452.

3. Yan, R., S. Moon, …, K. Xu. 2018. Spectrally resolved and functional super-resolution microscopy via ultrahigh-throughput single-molecule spectroscopy. *Acc. Chem. Res.* 51:697–705.

4. Altman, R. B., D. S. Terry, …, S. C. Blanchard. 2011. Cyanine fluorophore derivatives with enhanced photostability. *Nat. Methods.* 9:68–71.

5. Halabi, E. A., D. Pinotsi, and P. Rivera-Fuentes. 2019. Photoregulated fluxional fluorophores for live-cell super-resolution microscopy with no apparent photobleaching. *Nat. Commun.* 10:1232.

6. Juette, M. F., D. S. Terry, …, S. C. Blanchard. 2016. Single-molecule imaging of non-equilibrium molecular ensembles on the millisecond timescale. *Nat. Methods.* 13:341–344.

7. Miller, H., Z. Zhou, …, M. C. Leake. 2018. Single-molecule techniques in biophysics: a review of the progress in methods and applications. *Rep. Prog. Phys.* 81:024601.

8. Zhou, X., and S. T. C. Wong. 2006. Informatics challenges of high-throughput microscopy. *IEEE Signal Process. Mag.* 23:63–72.

9. Blanco, M., and N. G. Walter. 2010. Analysis of complex single-molecule FRET time trajectories. *Methods Enzymol.* 472:153–178.

10. Bronson, J. E., J. Fei, …, C. H. Wiggins. 2009. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* 97:3196–3205.

11. McKinney, S. A., C. Joo, and T. Ha. 2006. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.* 91:1941–1951.

12. Blanco, M. R., A. E. Johnson-Buck, and N. G. Walter. 2013. Hidden Markov modeling in single-molecule biophysics. *In* Encyclopedia of Biophysics. G. C. K. Roberts, ed. Springer, pp. 971–975.

13. White, D. S., M. P. Goldschen-Ohm, …, B. Chanda. 2020. Top-down machine learning approach for high-throughput single-molecule analysis. *eLife.* 9:e53357.

14. Dempsey, G. T., M. Bates, …, X. Zhuang. 2009. Photoswitching mechanism of cyanine dyes. *J. Am. Chem. Soc.* 131:18192–18193.

15. Levene, M. J., J. Korlach, …, W. W. Webb. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science.* 299:682–686.

16. Greenfeld, M., D. S. Pavlichin, …, D. Herschlag. 2012. Single Molecule Analysis Research Tool (SMART): an integrated approach for analyzing single molecule data. *PLoS One.* 7:e30024.

17. Hines, K. E., J. R. Bankston, and R. W. Aldrich. 2015. Analyzing single-molecule time series via nonparametric Bayesian inference. *Biophys. J.* 108:540–556.

18. Sgouralis, I., and S. Pressé. 2017. An introduction to infinite HMMs for single-molecule data analysis. *Biophys. J.* 112:2021–2029.

19. Celik, N., F. O'Brien, …, R. Barrett-Jolley. 2020. Deep-channel uses deep neural networks to detect single-molecule events from patch-clamp data. *Commun. Biol.* 3:3.

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

21. Xu, J., G. Qin, …, X. Fang. 2019. Automated stoichiometry analysis of single-molecule fluorescence imaging traces via deep learning. *J. Am. Chem. Soc.* 141:6976–6985.

22. Kadane, J. B., and N. A. Lazar. 2004. Methods and criteria for model selection. *J. Am. Stat. Assoc.* 99:279–290.

23. Shuang, B., D. Cooper, …, C. F. Landes. 2014. Fast step transition and state identification (STaSI) for discrete single-molecule data analysis. *J. Phys. Chem. Lett.* 5:3157–3161.

24. Holden, S. J., S. Uphoff, …, A. N. Kapanidis. 2010. Defining the limits of single-molecule FRET resolution in TIRF microscopy. *Biophys. J.* 99:3102–3111.

25. Van Oostveldt, P., F. Verhaegen, and K. Messens. 1998. Heterogeneous photobleaching in confocal microscopy caused by differences in refractive index and excitation mode. *Cytometry.* 32:137–146.

26. Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19:716–723.

27. Hannan, E. J., and B. G. Quinn. 1979. The determination of the order of an autoregression. *J. R. Stat. Soc. B.* 41:190–195.

28. Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Stat.* 6:461–464.

29. Priestley, M. B. 2004. Spectral Analysis and Time Series. Elsevier, London.

30. Bishop, C. 2006. Pattern Recognition and Machine Learning. Springer, New York.

31. Loeff, L., J. W. J. Kerssemakers, …, C. Dekker. 2021. *AutoStepfinder*: a fast and automated step detection method for single-molecule analysis. *Patterns (N Y).* 2:100256.

32. Van Rijsbergen, C. J. 1979. Information Retrieval, Second Edition. Butterworths, London.

33. Goldschen-Ohm, M. P., V. A. Klenchin, …, B. Chanda. 2016. Structure and dynamics underlying elementary ligand binding events in human pacemaking channels. *eLife.* 5:e20797.

34. Stennett, E. M. S., M. A. Ciuba, …, M. Levitus. 2015. Demystifying PIFE: the photophysics behind the protein-induced fluorescence enhancement phenomenon in Cy3. *J. Phys. Chem. Lett.* 6:1819–1823.

35. Vlijm, R., J. S. J. Smitshuijzen, …, C. Dekker. 2012. NAP1-assisted nucleosome assembly on DNA measured in real time by single-molecule magnetic tweezers. *PLoS One.* 7:e46306.

36. Isojima, H., R. Iino, …, M. Tomishige. 2016. Direct observation of intermediate states during the stepping motion of kinesin-1. *Nat. Chem. Biol.* 12:290–297.

37. Yoo, J., J.-Y. Kim, …, H. S. Chung. 2020. Fast three-color single-molecule FRET using statistical inference. *Nat. Commun.* 11:3336.

38. Tang, J., Y. Sun, …, K. Y. Han. 2017. Spatially encoded fast single-molecule fluorescence spectroscopy with full field-of-view. *Sci. Rep.* 7:10945.