SAND2023-01349C

Sandia National Laboratories

Spatial-Temporal Modeling for Nanoscale Resolvement of Photo-Switching Fluorophores

March 30 2023

Lekha Patel[1], **Ed Cohen**

[1]Department of Scientific Machine Learning, Center of Computing Research, Sandia National Laboratories

Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and opera
owned subsidiary of Honeywell International Inc., for the U.S. Department of E
DE-NA0003525.

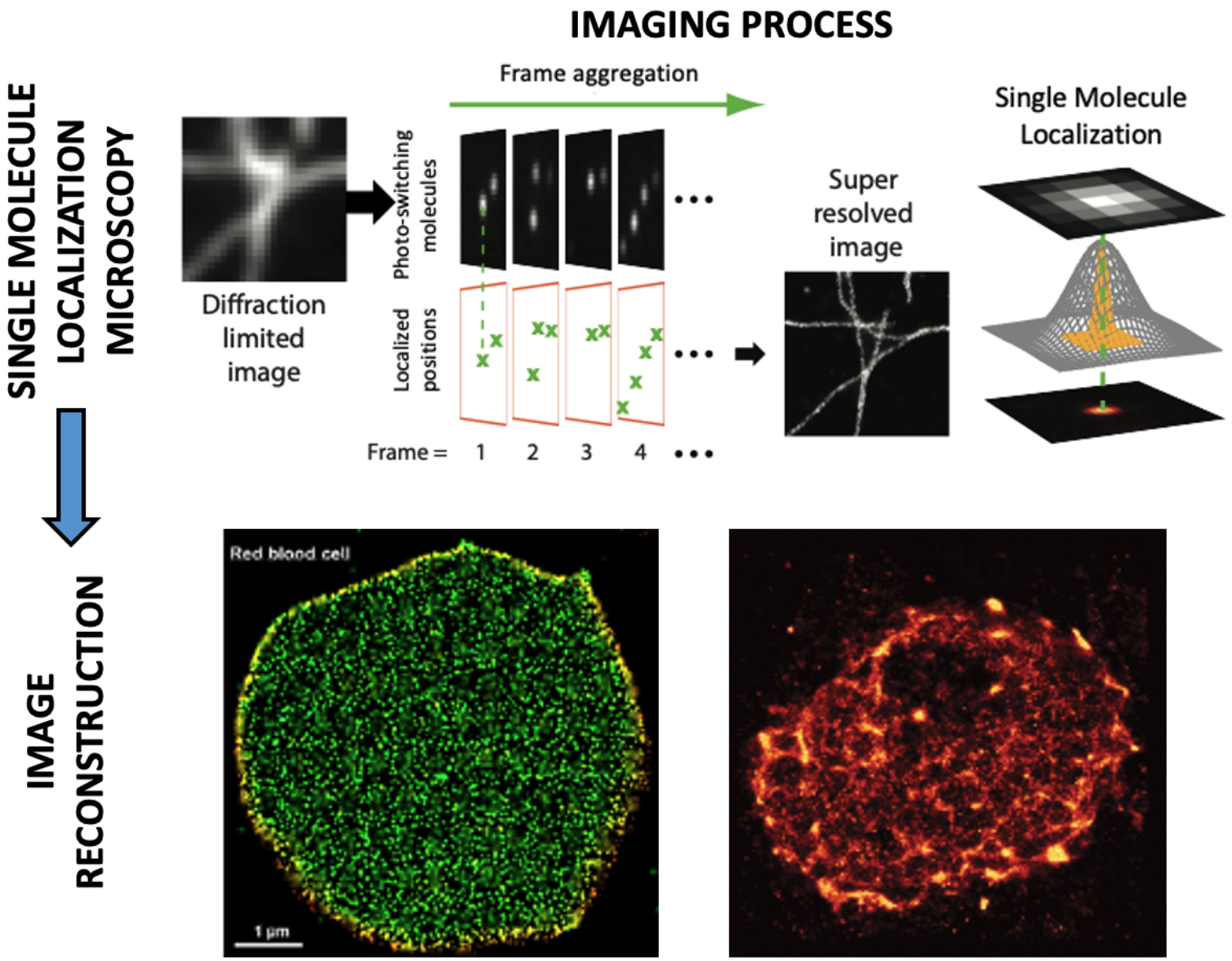
NA-0003525. SAND NO. SAND2021-2970 PE

Outline

Fluorescence Imaging I

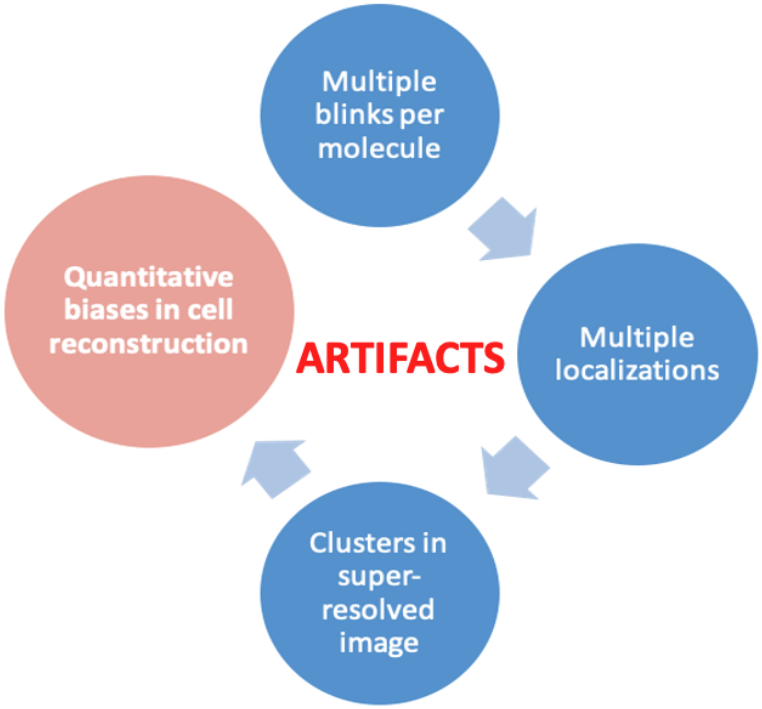
- Imaging at high optical resolutions is frequently achieved via *stochastic* photo-switching of molecules over $\approx 10^4$ frames (Betzig et al., 2006; Heilemann et al., 2008).
- Stochastic imaging takes many different forms: DNA-PAINT, (f)PALM, (d)STORM.
- Enables sparse subsets to be detected at any one time.
- Allows us to retrieve their spatial coordinates with high precision.
- Large cellular structures/mechanisms can be imaged.

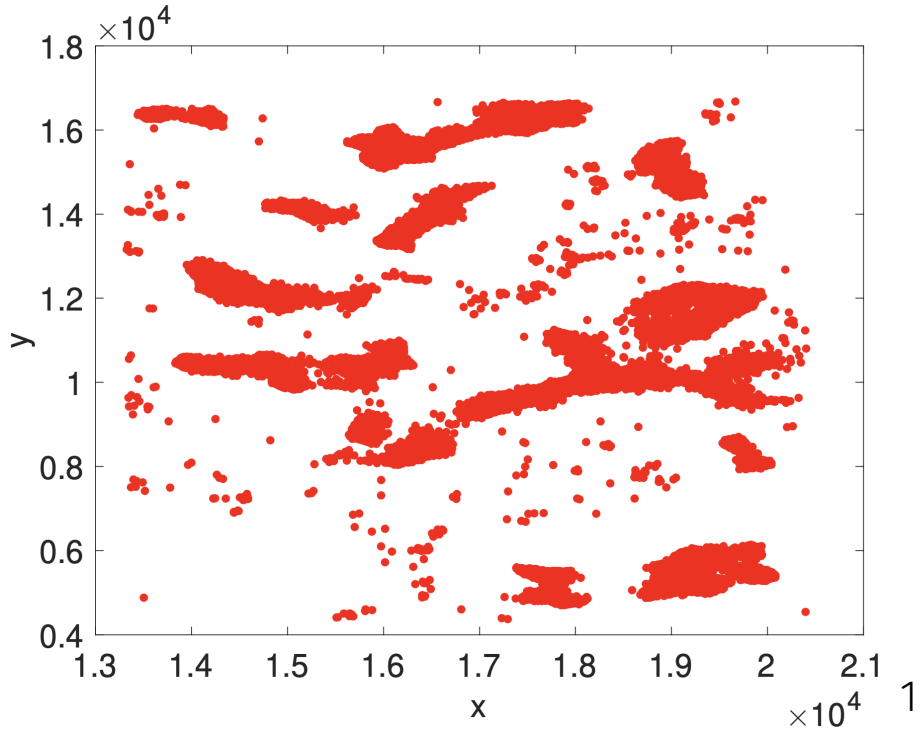
Obtaining super-resolution images I



Obtaining super-resolution images II

- Final superresolution images obtained by *superimposition* of frames.
- Final image shows dense spatial data around molecules of interest, which motivates spatial analyses.





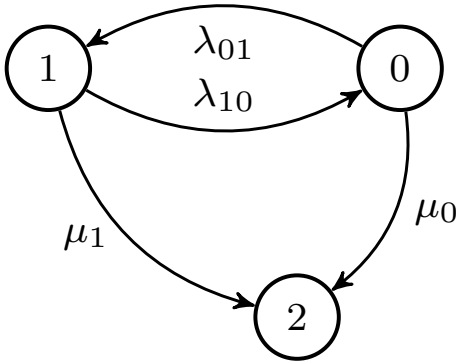
[1] Data made available courtesy of Prof. Paul French at ICL.

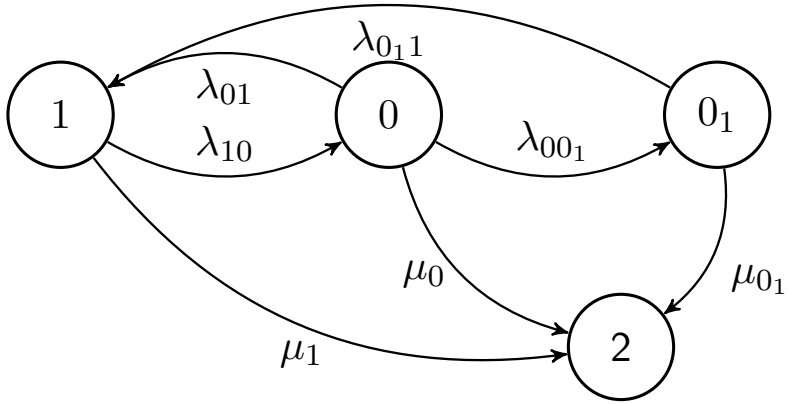
Stochastic photo-switching I

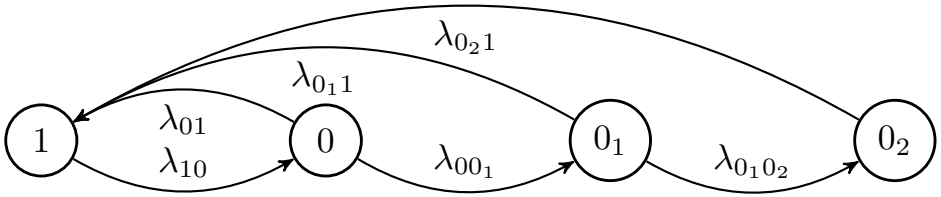
- The *photo-switching* behavior gives key information about the molecule e.g. inherent structure, pH, temperature.
- A molecule can be detected at most once in a single frame.
- Each frame gives a single representation of the mechanisms under observation.
- Utilizing the time domain in stochastic superresolution is key for analyzing underlying spatial structures.

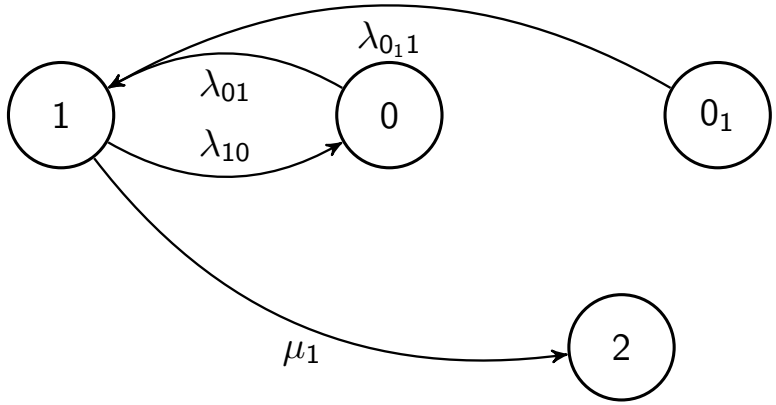
1. We can better understand molecular structures given partially-observed localizations through time. **How do we build a spatial model that leverages temporal characteristics?**
2. Can we use this to **predict photo-switching rates and underlying spatial positions, including molecular count?**

Characterizing photo-switching I



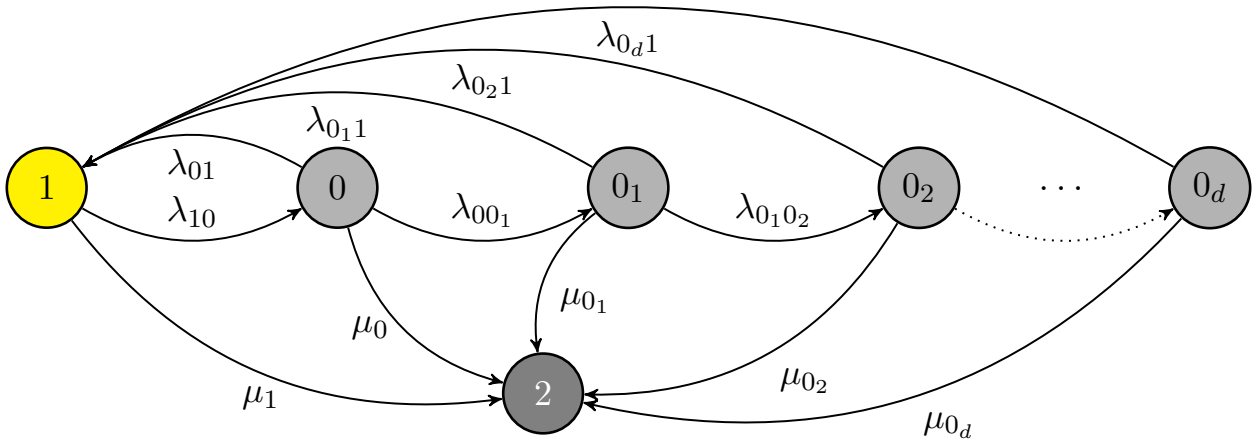






The photo-switching process $X(t)$

- For each molecule, let $\{X(t)\}_{t\in\mathbb{R}_{\geq 0}}$ be its underlying continuous time Markovian signal.
- Accesses a single on state (1), $d+1$ off states $(0, 0_1, \ldots, 0_d)$ and an absorption state (2).
- Its state space is $\mathcal{S}_X = \{0, 0_1, \ldots, 0_d, 1, 2\}$.
- $d = 0$ for DNA-PAINT, BALM.
- $d = 1$ for (f)PALM.
- $d = 2$ for (d)STORM Patel et al. (2019, 2021)



The generator G

$$G = \begin{pmatrix} -\sigma_0 & \lambda_{00_1} & 0 & 0 & 0 & 0 & \dots & \lambda_{01} & \mu_0 \\ 0 & -\sigma_{0_1} & \lambda_{0_1 0_2} & 0 & 0 & 0 & \dots & \lambda_{0_1 1} & \mu_{0_1} \\ 0 & 0 & -\sigma_{0_2} & \lambda_{0_2 0_3} & 0 & 0 & \dots & \lambda_{0_2 1} & \mu_{0_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & -\sigma_{0_d} & \lambda_{0_d 1} & \mu_{0_d} \\ \lambda_{10} & 0 & 0 & 0 & 0 & 0 & \dots & -\sigma_1 & \mu_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

- The transition probability matrix $P_t = \mathrm{e}^{Gt}$ for any $t \geq 0$ is such that $P_t(i,j) = \mathbb{P}(X_t = j | X_0 = i)$.

The non-zero transition rates $\boldsymbol{\lambda}_G$ in G are *unknown*.

The observation process I

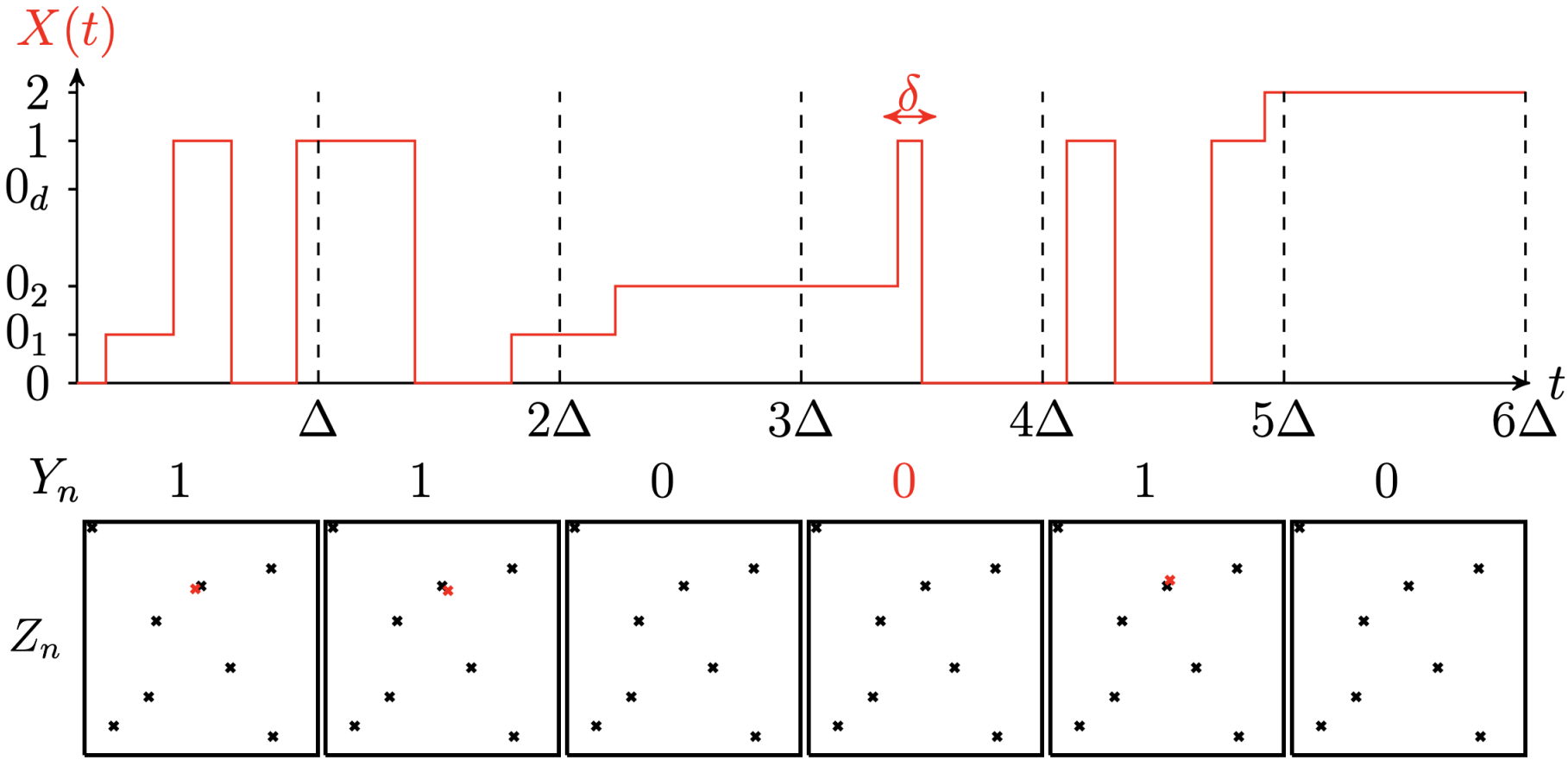
- Consider K (known for now) molecules filmed over N_F frames.
- Frame length Δ, minimum localization time is $\delta \in [0, \Delta)$.
- Let $\{Y_{n,k} \in \{0,1\}, n = 1, \dots, N_F\}$ be the discrete-time binary process indicating whether emitter k is localized in frame n.
- $\{Z_{n,k} \in W \subset \mathbb{R}^m, n = 1, \dots, N_F\}$ is the *spatial localization process*. If $Y_{n,k} = 0$, no observation is made. Otherwise, an observation about its true spatial position $\mathbf{c}_k \in W$ is made.

$$T_{n,k} = \int_{(n-1)\Delta}^{n\Delta} \mathbb{1}_{\{1\}}(X_k(t))\, \mathrm{d}t, \quad Y_{n,k} = \mathbb{1}_{[\delta,\Delta)}(T_{n,k}).$$

$$Z_{n,k} = \begin{cases} \emptyset & \text{if } Y_{n,k} = 0 \\ \mathbf{z} \sim \mathcal{N}\left(\mathbf{c}, \frac{\Delta}{N_p T_{n,k}} I_2\right) & \text{if } Y_{n,k} = 1, \end{cases}$$

N_p is the expected number of continuously emitted photons on Δ.

The observation process II



Ideal data tensor

Ideal data tensor $\mathcal{Z}$ uses $\mathcal{Y}$ to give *exact* fluorescent/binding spatial localizations for *each* molecule:

$$\mathcal{Z} = \begin{array}{c} \\ \text{No. frames} \downarrow \end{array} \begin{array}{c} \text{No. molecules} \rightarrow \\ \begin{array}{cccccc} 1 & 2 & 3 & \dots & K-1 & K \end{array} \\ \begin{pmatrix} \mathbf{z}_{1,1} & \mathbf{z}_{1,2} & \mathbf{z}_{1,3} & \dots & \mathbf{z}_{1,K-1} & \emptyset \\ \mathbf{z}_{2,1} & \emptyset & \emptyset & \dots & \mathbf{z}_{2,K-1} & \emptyset \\ \mathbf{z}_{3,1} & \mathbf{z}_{3,2} & \emptyset & \dots & \mathbf{z}_{3,K-1} & \mathbf{z}_{3,K} \\ \emptyset & \emptyset & \emptyset & \dots & \emptyset & \emptyset \\ \emptyset & \mathbf{z}_{5,2} & \emptyset & \dots & \emptyset & \emptyset \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \emptyset & \emptyset & \emptyset & \dots & \emptyset & \mathbf{z}_{N_F-1,K} \\ \emptyset & \mathbf{z}_{N_F,2} & \emptyset & \dots & \emptyset & \emptyset \end{pmatrix} \end{array} \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ \vdots \\ N_F - 1 \\ N_F \end{array}$$

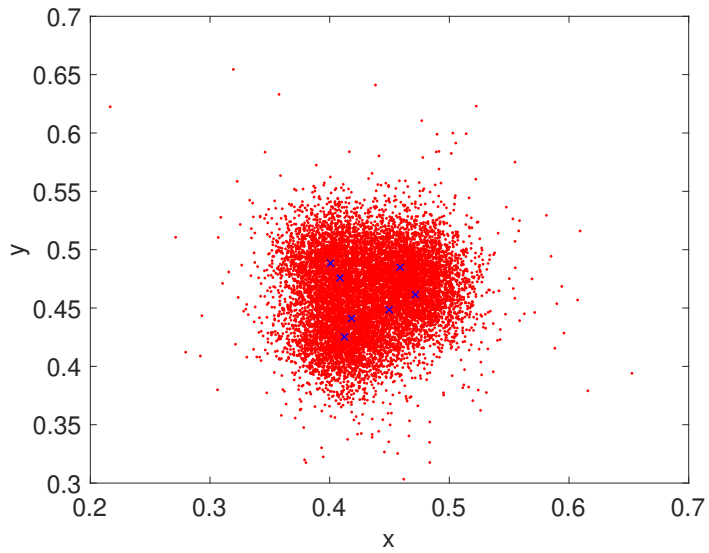
Why is the ideal case ideal?

The format of data gives us:

- The exact number of imaged molecules.
- The frame in which each molecule is localized.
- The spatial localization of each molecule.

Spatial reconstruction of true molecular positions, or *cluster positions* $\mathcal{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K)$ can be computed via Gaussian MLE:

$$\hat{\mathbf{c}}_k = \frac{1}{\sum_{n:\mathcal{Z}_{n,k}\neq\emptyset} 1} \sum_{n:\mathcal{Z}_{n,k}\neq\emptyset} \mathbf{z}_{n,k}.$$



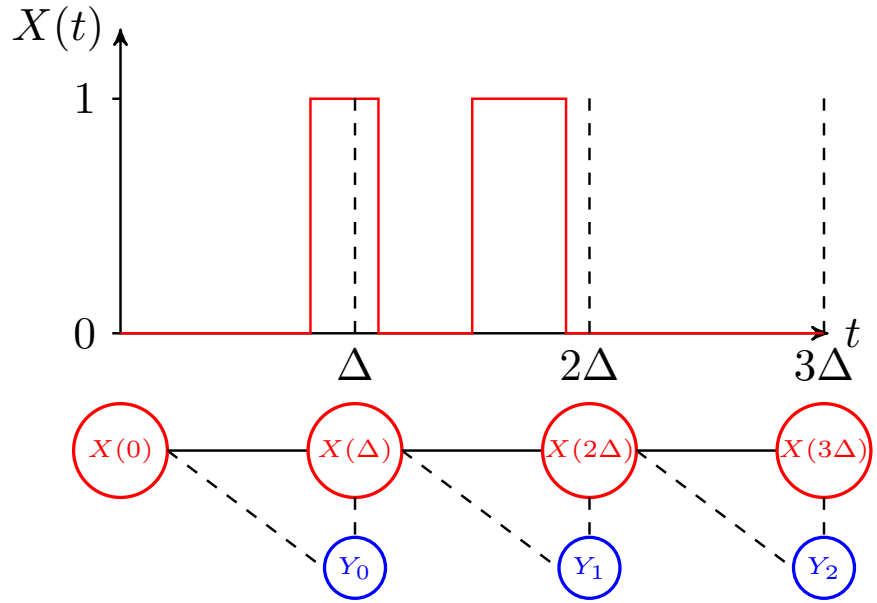
Why is the ideal case ideal?

The format of data gives us:

- The exact number of imaged molecules.
- The frame in which each molecule is localized.
- The spatial localization of each molecule.

Fluorescence/binding rates in G can be estimated using $\mathcal{Y}$ with the photo-switching Hidden Markov Model (PSHMM) (Patel et al., 2019), characterized with:

- **Initial probability vector**
  $\boldsymbol{\nu}_X := \nu_i = \mathbb{P}(X(0) = i)$.
- **Transmission matrices**
  $B_\Delta^{(l)} := B_\Delta^{(l)}(i,j) = \mathbb{P}(Y_n = l, X(\Delta) = j | X(0) = i)$.



Biological imaging data

- If molecules are sufficiently well-separated in space, post-processing image recognition algorithms can be used to accurately obtain data as $\mathcal{Y}$ and $\mathcal{Z}$ (Lin et al., 2015).
- In dense spatial structures of interest, obtaining sufficiently good spatial separation between molecules is highly challenging, if not impossible, to engineer.
- In this case, obtaining $\mathcal{Y}$, $\mathcal{Z}$ and K (number of molecules) is more difficult to obtain from imaging data alone.
- In each frame, localization algorithms fitting point-spread function to high-photon intensity regions will obtain spatial localizations; *including false positive localizations.*
- In each frame, a *set* of spatial localizations is obtained, without specific molecular labels.

Spatial localization set time series

Aggregate (set union) localizations over K

No. frames

$$
\begin{array}{ll}
\mathbf{Z_1} = \{\mathbf{z_{1,1}}, \mathbf{z_{1,2}}, \mathbf{z_{1,3}}\} & \begin{pmatrix} 1 & 2 & 3 & \dots & K-1 & K \\ \mathbf{z}_{1,1} & \mathbf{z}_{1,2} & \mathbf{z}_{1,3} & \dots & \emptyset & \emptyset \\ \mathbf{z}_{2,1} & \emptyset & \emptyset & \dots & \mathbf{z}_{2,K-1} & \emptyset \\ \emptyset & \mathbf{z}_{3,2} & \emptyset & \dots & \mathbf{z}_{3,K-1} & \emptyset \\ \emptyset & \emptyset & \emptyset & \dots & \emptyset & \emptyset \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \emptyset & \emptyset & \emptyset & \dots & \emptyset & \mathbf{z}_{N_F-1,K} \\ \emptyset & \mathbf{z}_{N_F,2} & \emptyset & \dots & \emptyset & \emptyset \end{pmatrix} \\
\mathbf{Z_2} = \{\mathbf{z_{2,1}}, \mathbf{z_{2,K-1}}\} & \\
\mathbf{Z_3} = \{\mathbf{z_{3,2}}, \mathbf{z_{3,K-1}}\} & \\
\mathbf{Z_4} = \emptyset & \\
\vdots & \\
\mathbf{Z_{N_F-1}} = \{\mathbf{z_{N_F-1,K}}\} & \\
\mathbf{Z_{N_F}} = \{\mathbf{z_{N_F,2}}\} &
\end{array}
$$

Instead of the ideal data tensor, we have the **spatial localization set time series** $\{Z_n : n = 1, \dots, N_F\}$ of localizations in each frame by molecular and false-positive aggregation (set union):

$$Z_n = A_n \cup \left\{ \bigcup_{k=1}^{K} \mathcal{Z}_{n,k} \right\}.$$

Inference problem

Using the spatial localization sets $\{Z_n\}_{n=1}^{N_F}$, we want to estimate:

- The number of molecules K.
- Their spatial positions $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$.
- Fluorescence/binding rates.

Spatial point process

A (simple, finite) spatial point process $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$ contains random spatial points with density $f(\mathbf{c})$ over W **with** cardinality distribution $p_K(k)$ on $\mathbb{Z}_{\geq 0}$. Its density function is

$$f(\mathcal{C}) := k! p_K(k) \prod_{\mathbf{c} \in \mathcal{C}} f(\mathbf{c}).$$

Since observations sets Z_n are related to $\mathcal{C}$, we aim to infer:

$$f(\mathcal{C} | \{Z_n\}_{n=1}^{N_F}) \propto f(\{Z_n\}_{n=1}^{N_F} | \mathcal{C}) \pi(\mathcal{C}).$$

A pure spatial birth process I

- Before a molecule first fluoresces/binds, it ceases to be part of the underlying molecular configuration.
- We introduce a *hidden spatial point process* $\{C_n : n = 1, \ldots N_F\}$ which gives the number and positions of molecules that have already been activated before or in frame n.
- Each C_n can be probabilistically compared with Z_n at each frame.
- C_n has the property that $C_n \xrightarrow{P} \mathcal{C}$.

Under any $\boldsymbol{\nu}_X$ we define for $n \geq 1$

$$C_n = C_{n-1} \cup B_n \qquad C_0 = \emptyset$$

$$Z_n = \Phi(C_n) \cup A_n$$

A pure spatial birth process II

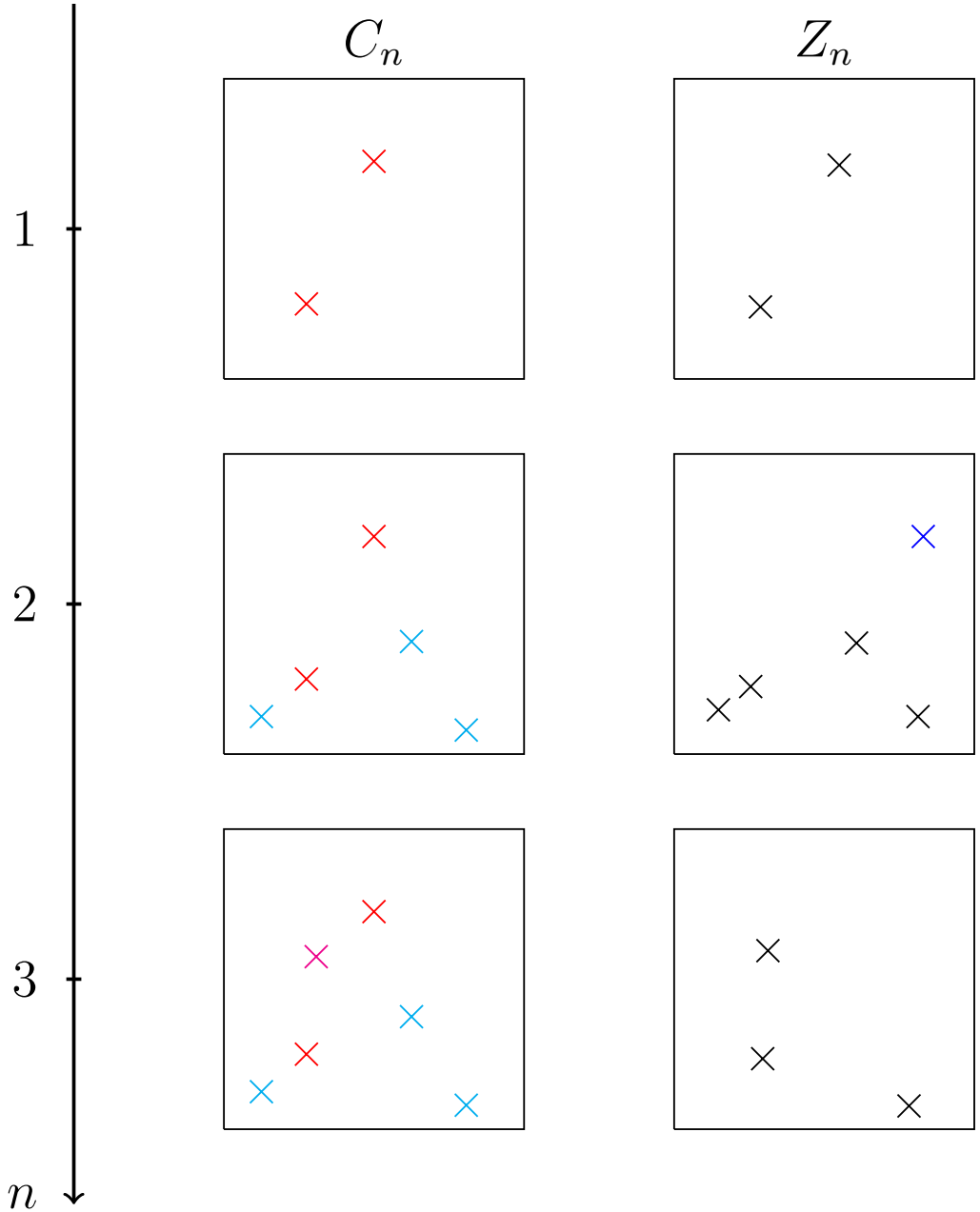
- B_n denotes the point process of molecules that are first activated in frame n.
- A_n denotes the point process of false positive observations, with rate α, in frame n.
- Each $\mathbf{c} \in C_n$ generates an offspring $\mathbf{z} \in \Phi(C_n)$, where $\mathbf{z}$ is a variate from the density

$$f(\mathbf{z}^*|\mathbf{c}) = \mathcal{N}\left(\mathbf{c}, \frac{\Delta}{N_p T_n} I_2\right),$$

with probability $p_{D,n}$ and is empty with probability $1 - p_{D,n}$.

$$\mathbf{z} = \begin{cases} \emptyset & \text{with probability } 1 - p_{D,n} \\ \{\mathbf{z}^*\} & \text{with probability } p_{D,n} f(\mathbf{z}^*|\mathbf{c}). \end{cases}$$

An illustration

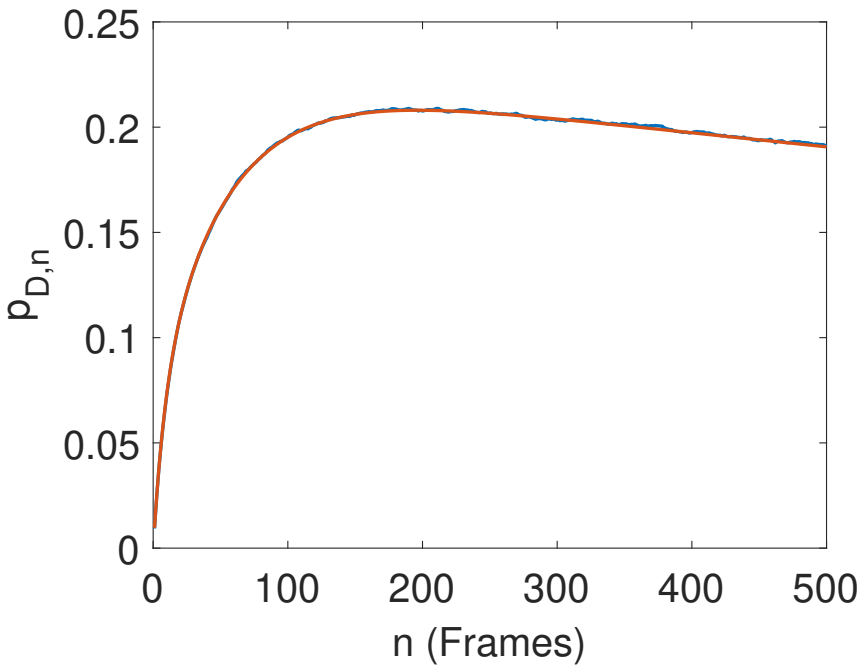


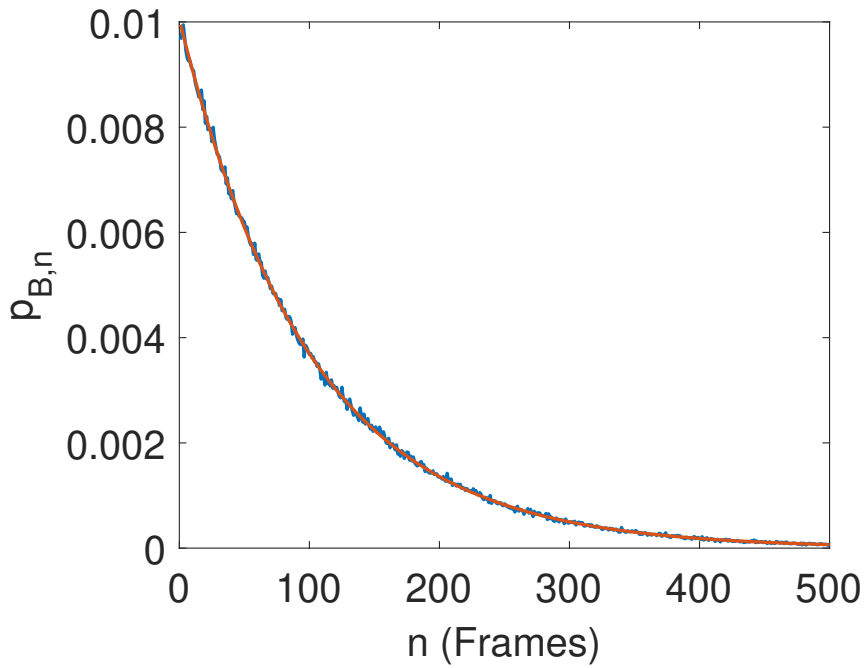
Characterizations I

For $n \geq 1$, let $p_{B,n}(\boldsymbol{\theta})$ to be the probability a molecule first appears, and $p_{D,n}(\boldsymbol{\theta})$ to be the probability of detecting a single molecule in frame n, dependent on $\boldsymbol{\theta} = \{\boldsymbol{\lambda}_G, \delta, \alpha\}$.

$$p_{B,n}(\boldsymbol{\theta}) = \boldsymbol{\nu}_X^{*\top}(B^{(0)}_{T^*,\delta=0})(B^{(0)}_{\Delta,\delta=0})^{n-1}B^{(1)}_{\Delta,\delta=0}\mathbf{1}_{d+3}$$

$$p_{D,n}(\boldsymbol{\theta}) = \boldsymbol{\nu}_X^{\top}\mathrm{e}^{G(n-1)\Delta}B^{(1)}_{\Delta}\mathbf{1}_{d+3}.$$





Characterizations II

- $K \sim \text{Poi}(\lambda_K)$. $N_{B,n} \sim \text{Poi}(\lambda_K p_{B,n}(\boldsymbol{\theta}))$:
- $f_{B_n}(C') = \mathrm{e}^{-\lambda_K p_{B_n}(\boldsymbol{\theta})} \prod_{\mathbf{c}\in C'} \lambda_K p_{B_n}(\boldsymbol{\theta}) b(\mathbf{c})$
- $f_{A_n}(C') = \mathrm{e}^{-\alpha} \prod_{\mathbf{c}\in C'} \alpha a(\mathbf{c})$.
- Spatial prior over W, e.g. $a(\mathbf{c}) = b(\mathbf{c}) = 1/|W|$ (CSR).

We can then determine $f(C_{N_F}|\{Z_n\}_{n=1}^{N_F})$ by iteratively computing:

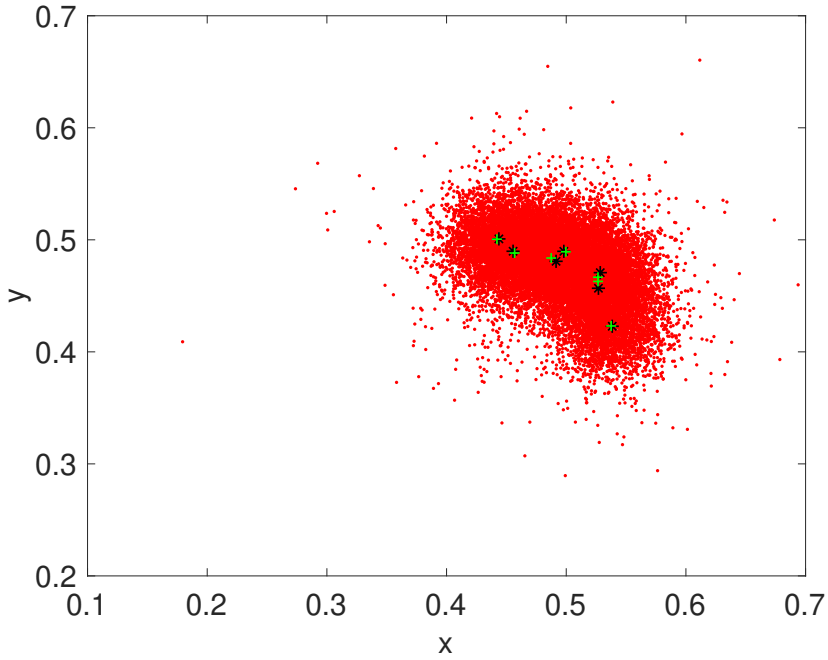
$$f(C_n|\{Z_{n'}\}_{n'=1}^{n}\boldsymbol{\theta}) \propto f(C_n|\{Z_{n'}\}_{n'=1}^{n-1}, \boldsymbol{\theta}) f(Z_n|C_n, \boldsymbol{\theta}).$$

- First term requires finding all smaller subsets of C_n that give rise to different birth sets in each frame.
- Second term requires finding all subsets of Z_n that could come from false positive observations and C_n. Computed offline: sets of size k yield observations of size $\leq k$, rest are false-positives.

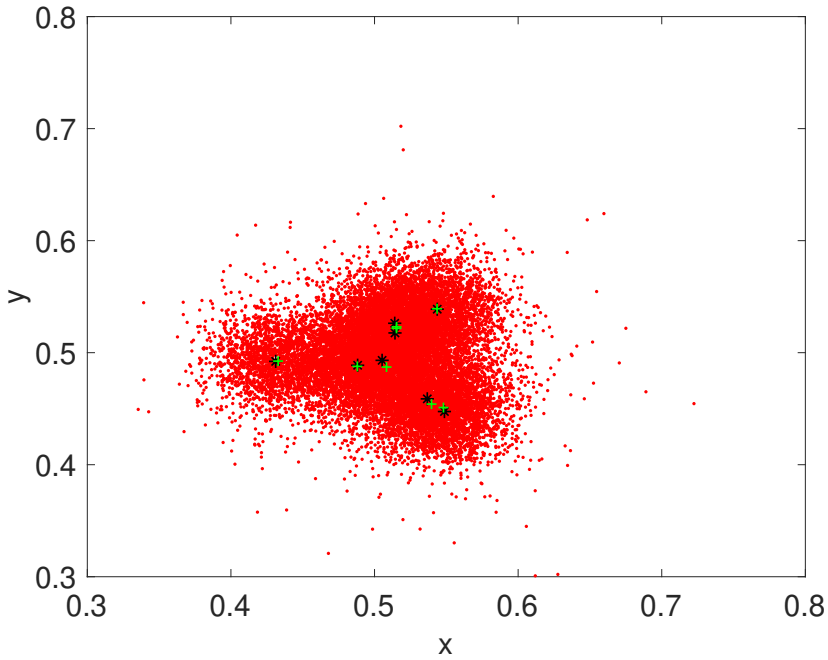
Inference I

- The computation of $f(C_n|\{Z_{n'}\}_{n'=1}^{n-1}, \boldsymbol{\theta})$ can become computationally intractable for large sets C_n.
- For molecules that are already activated before imaging, e.g. in dSTORM, $p_{B,n} = 0$ for $n > 1$, which allows for fast computation.
- We use a birth-death-shift MCMC sampler to sample C_{N_F} from $\{Z_n\}_{n=1}^{N_F}$ (i.e. underlying spatial positions and its cardinality).
- $\boldsymbol{\theta} = \{\boldsymbol{\lambda}_G, \delta, \alpha\}$ can also be estimated within the algorithm through its contribution in $p_{D,n}$.
- In real applications the localization standard deviation is determined through the localization algorithm, and is therefore known.

Simulations I



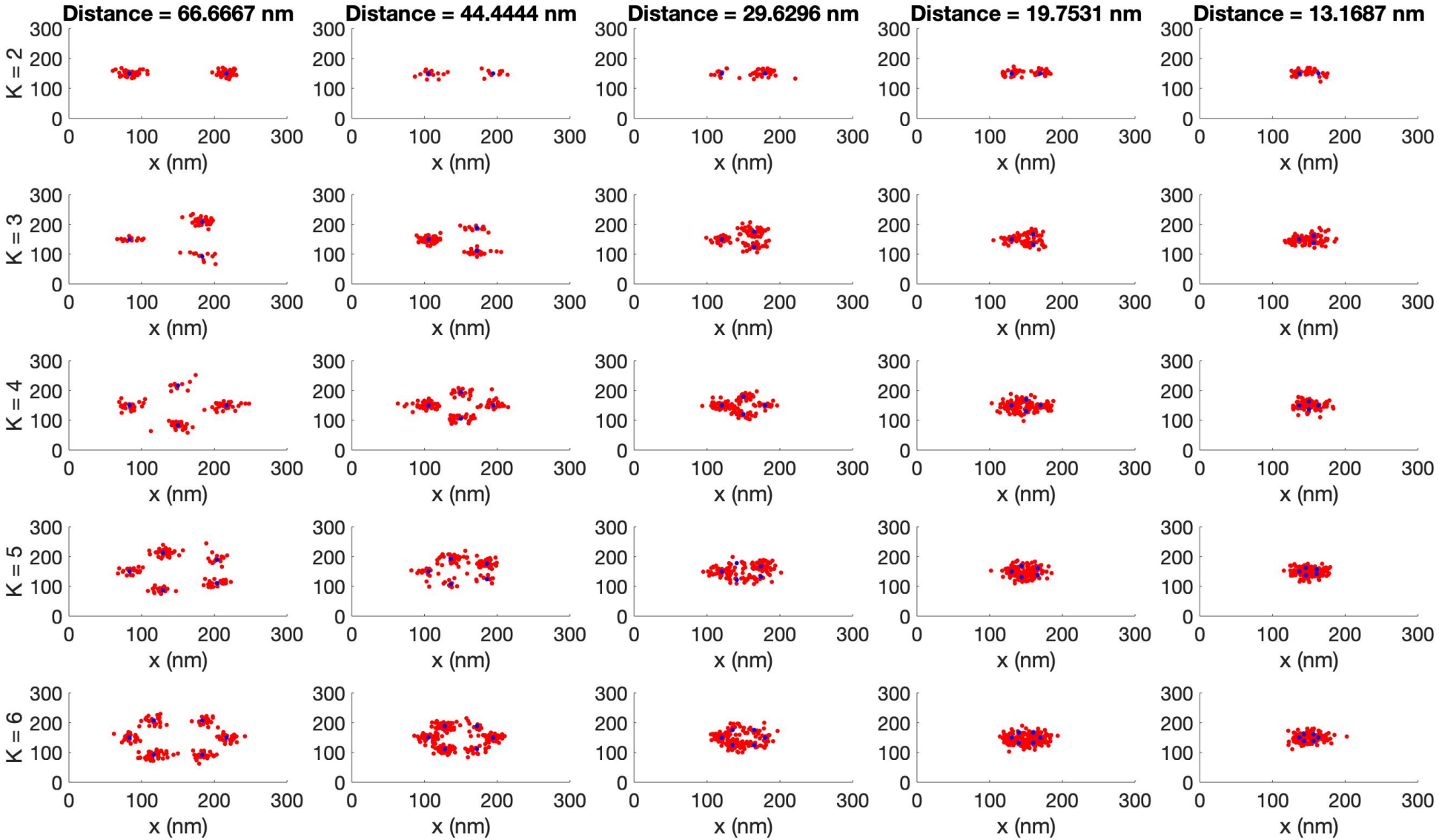
Simulation 1.



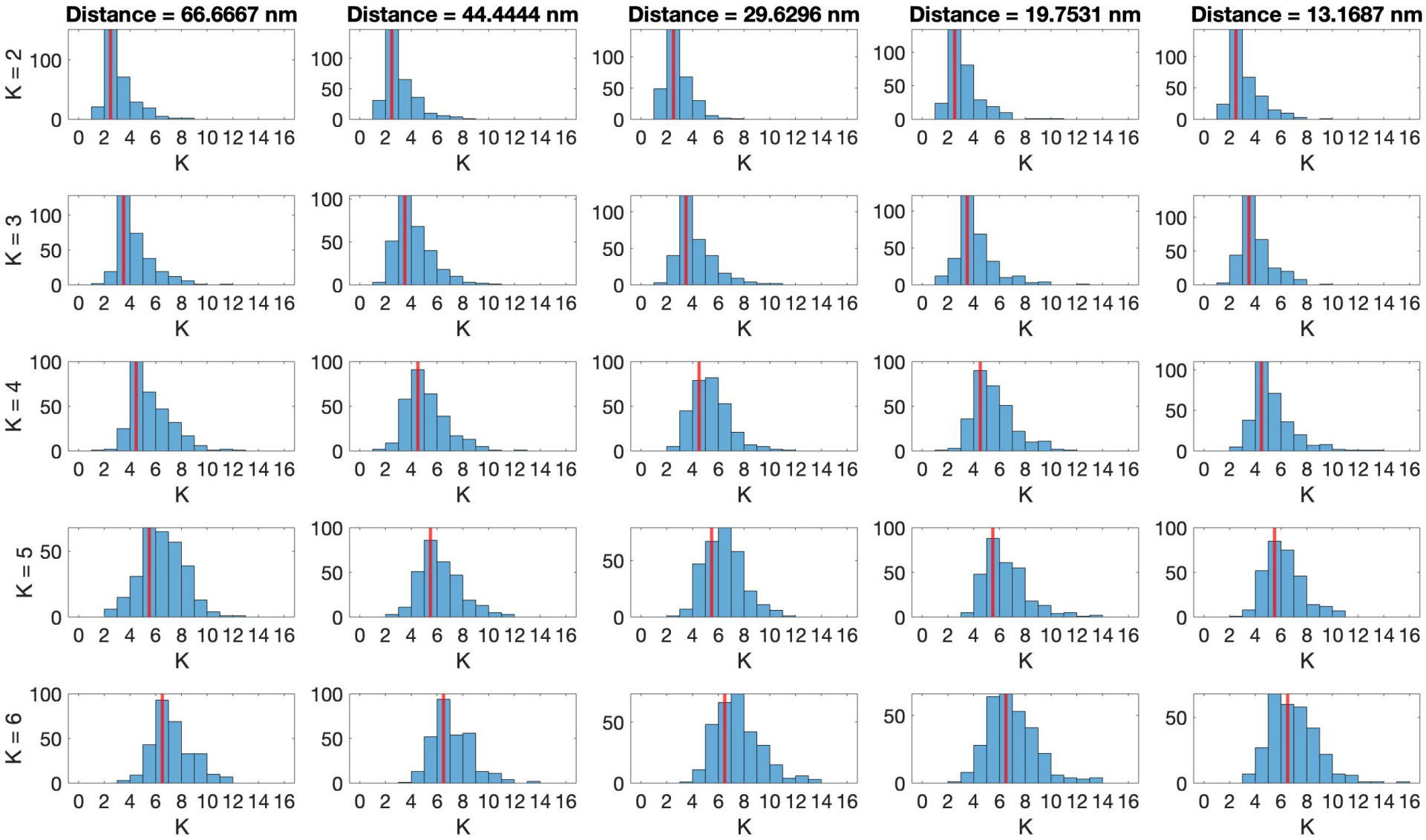
Simulation 2.

Spatial estimates of 2 simulation studies under the $d = 2$ model.

Simulations II



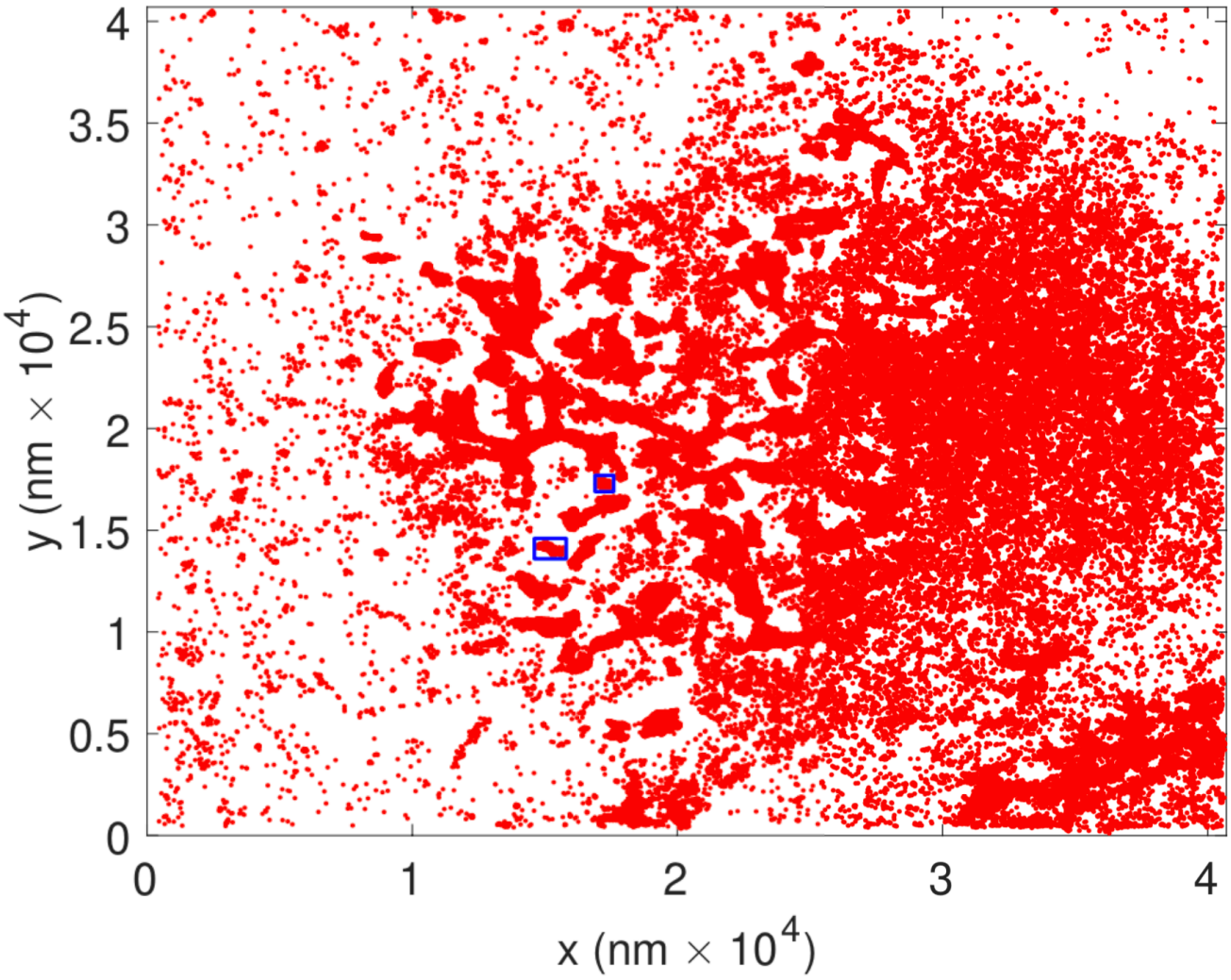
Simulations III



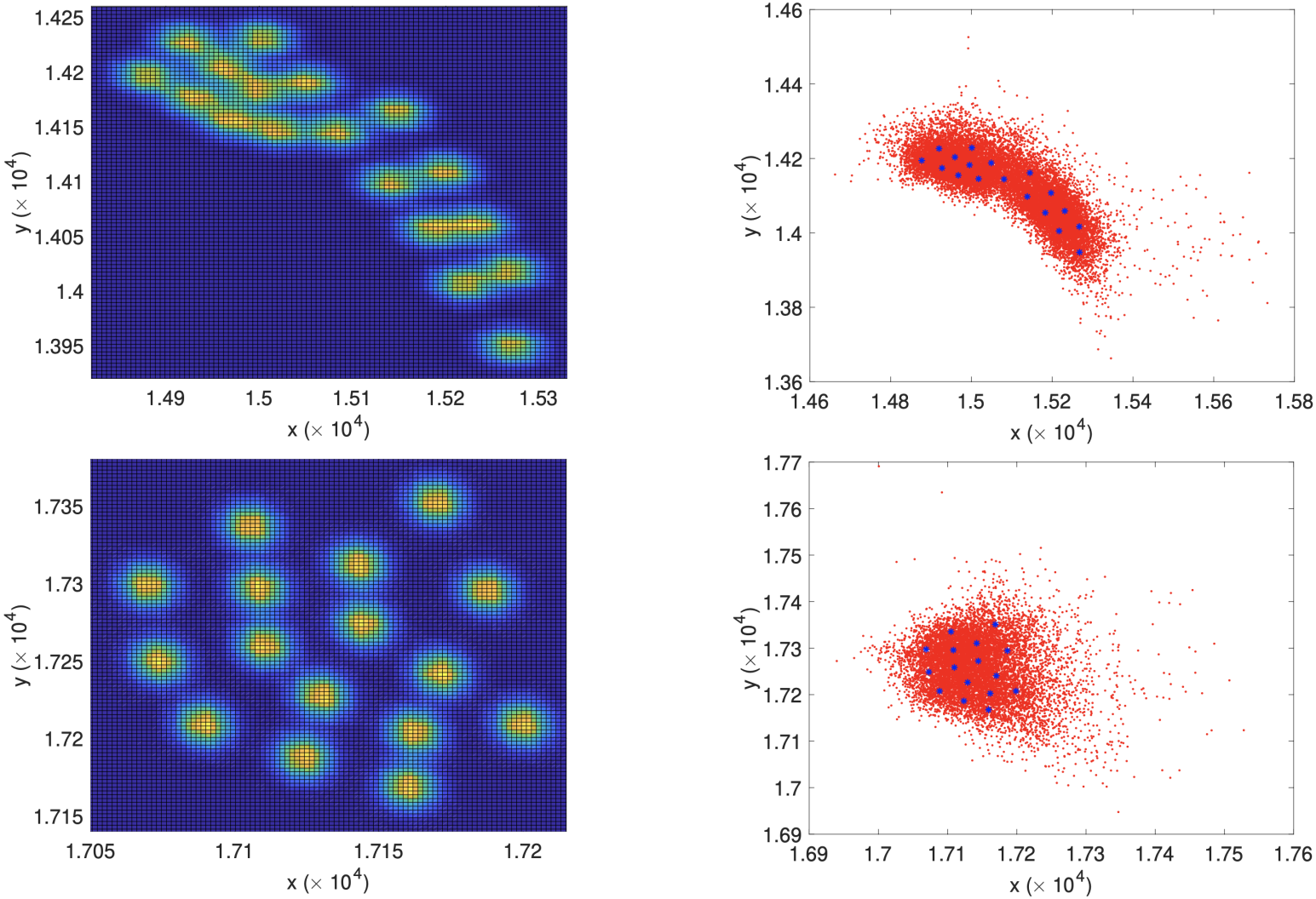
Application to Real Data I

- Chromosomes labeled with Alexa Fluor 488 fluorophores embedded in Polyvinyl Alcohol (PVA) resin considered.[2]
- Imaged over 10^4 frames, at a frame rate of $\Delta = 0.03$ seconds, with dSTORM.
- Investigated 2 clusters of molecules using subsets of raw image.
- The algorithm of Ovesný et al. (2014) applied to find high photon intensity spots, and localizations determined via Gaussian PSFs.
- Under all experiments, the $d = 2$ model was picked, with $\mu_0 = \mu_{0_1} = \mu_{0_2} = 0$.

Application to Real Data II



Application to Real Data III



Spatial estimates (with cardinality) of two clusters of image. Spatial regularity of about between fluorophores, with structure, is observed.

[2] Dataset was made available by Professor Paul French, Imperial College London.

Model summary

- Leveraging temporal characteristics of molecules imaged in stochastic superresolution is key for understanding underlying spatial structures.
- Data of imaging observations has been described - both ideal and real cases.
- In the ideal case, spatial analysis is straight-forward.
- In the real case, a spatial-temporal joint point process can be formulated that utilizes photo-switching/binding temporal characteristics.
- This model is used to describe a pure birth *hidden* process of molecular positions and count as they become activated.
- Hidden process related to observation point process which includes false positive measurements.
- Inference via MCMC can be utilized to recover hidden process.
- Robust method shown on simulations and real data.

References I

E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess. Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science*, 313(5793):1642–1645, 2006.

M. Heilemann, S. Van de Linde, M. Schüttpelz, R. Kasper, B. Seefeldt, A. Mukherjee, P. Tinnefeld, and M. Sauer. Subdiffraction - Resolution Fluorescence Imaging with Conventional Fluorescent Probes. *Angewandte Chemie International Edition*, 47(33):6172–6176, 2008.

Y. Lin, J. J. Long, F. Huang, W. C. Duim, S. Kirschbaum, Y. Zhang, L. K. Schroeder, A. A. Rebane, M. G. M. Velasco, A. Virrueta, D. W. Moonan, J. Jiao, S. Y. Hernandez, Y. Zhang, and J. Bewersdorf. Quantifying and Optimizing Single-Molecule Switching Nanoscopy at High Speeds. *Plos One*, 10(5):e0128135, 2015.

References II

M. Ovesný, J. Borkovec P. Křížek, Z. Švindrych, and G.M Hagen. ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics*, 30(16):2389–2390, 2014.

L. Patel, N. Gustafsson, Y. Lin, R. Ober, R. Henriques, and E. Cohen. A hidden Markov model approach to characterizing the photo-switching behavior of fluorophores. *Annals of Applied Statistics*, 13(3):1397–1429, 2019.

L. Patel, D. Williamson, D. M. Owen, and E. A. K. Cohen. Blinking statistics and molecular counting in direct stochastic reconstruction microscopy (dSTORM). *Bioinformatics*, 37(17): 2730–2737, 02 2021.