Single-molecule localization microscopy as nonlinear inverse problem

Ji Yu[a,1] and Ahmed Elmokadem[a,2]

[a]Center for Cell Analysis and Modeling, University of Connecticut Health Center, Farmington, CT 06030

Edited by Melike Lakadamyali, University of Pennsylvania, Philadelphia, PA, and accepted by Editorial Board Member Yale E. Goldman September 1, 2019 (received for review July 23, 2019)

We present a statistical framework to model the spatial distribution of molecules based on a single-molecule localization microscopy (SMLM) dataset. The latter consists of a collection of spatial coordinates and their associated uncertainties. We describe iterative parameter-estimation algorithms based on this framework, as well as a sampling algorithm to numerically evaluate the complete posterior distribution. We demonstrate that the inverse computation can be viewed as a type of image restoration process similar to the classical image deconvolution methods, except that it is performed on SMLM images. We further discuss an application of our statistical framework in the task of particle fusion using SMLM data. We show that the fusion algorithm based on our model outperforms the current state-of-the-art in terms of both accuracy and computational cost.

superresolution microscopy | single-molecule localization | statistical modeling | particle fusion

Single-molecule localization microscopy (SMLM) is a set of fluorescence microscopy techniques that have enabled visualization of small biological structures that was previously considered impossible due to the resolution limit (1). They do this by repeatedly imaging a small random subset of fluorescent molecules in the sample, creating images with sparse support and thereby allowing extremely high accuracy in determining the locations of the molecules. These techniques sacrifice temporal resolution to some degree but gain enhanced spatial resolution by at least 1 order of magnitude.

An SMLM dataset, although often presented in an image format, is in fact a set of highly accurate localization coordinates. Nevertheless, the scientific interest, whenever one performs an SMLM experiment, is usually to gain knowledge about the spatial distribution of biomolecules in a specimen. Despite their accuracy, localization coordinates differ from the true localizations of the molecules by a certain error. Meanwhile, there is a finite and nonuniform probability that any 2 localizations from an experiment can be originated from the same spatial location (within error, i.e., from the same pixel), and, sometimes, from the exact same target molecule. Thus, the SMLM images are approximate measures of the true molecular distributions at best.

We are interested in the general question of what can be said about the true molecular distribution, let us call it $\boldsymbol{\theta}$, from a set of observed localizations from an SMLM experiment. Like most of the inverse problems in sciences and engineering, the problem of solving $\boldsymbol{\theta}$ is both ill-posed and without a unique solution. Therefore, the question can only be studied in a probabilistic sense, that is, by computing a probability density of $\boldsymbol{\theta}$ conditioned on the experimental observations.

An analogy can be made between the task of estimating $\boldsymbol{\theta}$ and the practice of deconvolution in traditional optical microscopy. Deconvolution is a well-known technique to reverse the effect of optical diffraction, which is the source of the resolution limit in normal optical microscopy images. In the traditional photon-based images, localization information is mediated by individual photons and carries an uncertainty arising from the physical nature of the diffraction; similarly in SMLM, localization information is mediated by single molecules and carries an uncertainty arising from fitting errors. Therefore, computing $\boldsymbol{\theta}$ is to perform image restoration on SMLM data and should have a similar effect as the deconvolution operator on traditional microscopy images. There is, however, a key mathematical difference between photon-based images and SMLM images: The blurring in photon-based images can be characterized by PSF (point spread function), which describes the full statistics of how a photon deviates from its true location of origin during the imaging process. An implicit assumption embedded in the PSF is that all photons carry the same uncertainty statistics, at least when they are detected in nearby locations. The same cannot be said for SMLM images, because the localization uncertainty in SMLM depends on signal intensity, which varies and follows a relatively broad distribution. Therefore, computing $\boldsymbol{\theta}$ would require a very different approach from that of the traditional deconvolution.

Here, we present a statistical model that allows numerical computation of the posterior distribution of $\boldsymbol{\theta}$, as well as iterative algorithms for parameter estimation based on this model. We performed some validations and numerical analyses of these algorithms using simulated imaging datasets, and we also demonstrated an application of the model in particle fusion using experimental data.

Significance

Single-molecule localization microscopy (SMLM) is quickly becoming an indispensable tool for studying biological structures. Meanwhile, we are presented with new challenges in the analyses of SMLM data arising from their unique noise statistics. In this paper, we solved a fundamental problem with regard to the modeling of SMLM data. Specifically, we established a theoretical connection between a set of experimentally acquired SMLM data and the underlying spatial distribution of fluorescent molecules. Based on this framework, we can extract higher-resolution information that is obscured by the raw data. The same framework also allows us to develop better image alignment algorithms, and thus obtain faster and more accurate results in particle fusion applications.

Author contributions: J.Y. designed research; J.Y. performed research; J.Y. and A.E. analyzed data; and J.Y. and A.E. wrote the paper.

Conflict of interest statement: A.E. is currently an employee of Metrum Research Group, a biomedical consulting firm. A.E. performed research related to this publication as a graduate student at UConn Health, prior to his employment at Metrum Research Group.

This article is a PNAS Direct Submission. M.L. is a guest editor invited by the Editorial Board.

Data deposition: The current versions of our implementation of the various algorithms have been deposited at GitHub, https://github.com/jiyuuchc/lmdeconv.

[1]To whom correspondence may be addressed. Email: jyu@uchc.edu.

[2]Present address: Metrum Research Group, Tariffville, CT 06081.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1912634116/-/DCSupplemental.

First published September 23, 2019.

Statistical Model

Our primary interest is to calculate the spatial distribution of the fluorescent molecules in the imaged area. Since the data obtained experimentally have a finite number of localizations, it is only reasonable to infer the molecular distribution also on a finite basis, i.e., via discretization. It is customary to model such a distribution on a grid of pixels. Here, we denote θ_j as the fraction of molecules located within pixel j. Therefore, we define the following:

$$\boldsymbol{\theta} \overset{def}{=} \{\theta_1, \ldots, \theta_m\}, \text{ where } \theta_j \in (0,1) \text{ and } \Sigma_j \theta_j = 1. \quad [1]$$

We assume the image has m total pixels.

It should be noted that Eq. **1** makes no assumption concerning how the pixels are arranged in real space, nor does it make any assumptions about the spatial dimension. In fact, the theory discussed here applies to data in any spatial dimension. In addition, we note that, in this paper, we are only concerned with the relative distribution of molecules and not their total number in the sample. However, this should not be viewed as a limitation. Methods to estimate the absolute molecular counts from single-molecule data have been the focus of several previous studies (2, 3) and can be performed independently of the calculations discussed in this paper. Molecular counts can then be combined with $\boldsymbol{\theta}$ to yield the absolute molecular concentrations in the sample of interest.

We use the notation $\boldsymbol{g}$ to represent the information contained within an SMLM dataset. Experimentally obtained SMLM datasets consist of a list of localization coordinates obtained from the fitting of many fluorescence images. Meanwhile, the experimental data usually also contain quantitative information regarding the localization uncertainty. Most of the modern SMLM fitting software can perform close to the Cramer–Rao lower bound (CRLB) (4–6); therefore, the localization uncertainties can often be estimated from the imaging data of the individual molecules. The uncertainty values are sometimes utilized in SMLM image rendering, where each localization is represented by a Gaussian spot, whose width is determined by the uncertainty values (7). Slightly complicating the problem are the cases in which 2 or more molecules are detected close to each other, resulting in a lowered fitting accuracy, the exact value of which can be difficult to determine experimentally. Here, we assume that these events are rare and that localization uncertainties on all detected localizations are available from the experiment:

$$\boldsymbol{g} \overset{def}{=} \left\{ g_{i,j} \in (0,1); i = 1, \ldots, n \quad j = 1, \ldots, m \right\} \sum_j g_{i,j} = 1 \text{ for all } i, \quad [2]$$

where $g_{i,j}$ denotes the probability that the ith localization was originated from the jth pixel due to the localization uncertainty and n is the total number of localizations obtained from the experiment. Note, in this notation, $\boldsymbol{g}$ represents the full knowledge about the data, and if we know $\boldsymbol{g}$, we know both the central positions of the localizations as well as their associated uncertainties.

Our goal is to find an efficient algorithm to quantitatively evaluate the posterior probability distribution $p(\boldsymbol{\theta}|\boldsymbol{g})$. To do that, we need to introduce a latent variable $\boldsymbol{u} \overset{def}{=} \{u_1, \ldots, u_n\}$ to represent the true localization of each detected localization, i.e., $u_i = j$ indicates that the true location of the ith localization is in the jth pixel. The distribution $p(\boldsymbol{\theta}|\boldsymbol{g})$ can then be expressed via the marginalization of $\boldsymbol{u}$,

$$p(\boldsymbol{\theta}|\boldsymbol{g}) = \frac{1}{C} p(\boldsymbol{\theta}) \sum_{\boldsymbol{u}} p(\boldsymbol{g}|\boldsymbol{\theta}, \boldsymbol{u}) p(\boldsymbol{u}|\boldsymbol{\theta}). \quad [3]$$

Here, the summation is over all possible choices of $\boldsymbol{u}$, C is a normalization constant, and $p(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$, reflecting the experimenter's knowledge about $\boldsymbol{\theta}$ before experimental data were obtained.

The exact functional form of $p(\boldsymbol{\theta})$ deserves some discussion. One obvious choice is the constant prior [$p(\boldsymbol{\theta}) = \text{constant}$], which is almost always a good choice, because it reflects the scientific desire of not assuming anything before the experiment but deriving all statistical information from the empirical data only. In fact, all of the numerical results presented in this paper were based on a constant prior. That said, in many practical situations, it may be desirable to use a more relaxed prior form. Specifically, we suggest a symmetric Dirichlet distribution for our model:

$$p(\boldsymbol{\theta}) = \text{Dir}(\boldsymbol{\theta} : \alpha_0) = \frac{\Gamma(n\alpha_0)}{[\Gamma(\alpha_0)]^n} \prod_i \theta_i^{\alpha_0 - 1}. \quad [4]$$

This prior choice has several advantages. First of all, $p(\boldsymbol{\theta})$ reduces to a constant function if $\alpha_0 = 1$; thus, the constant prior can be viewed as a special case of the Dirichlet prior. More importantly, the Dirichlet prior distribution facilitates tuning of the sparsity in the inference results. Specifically, if $\alpha_0 < 1$, $p(\boldsymbol{\theta})$ is larger when $\boldsymbol{\theta}$ is sparse (molecules are concentrated into a small number of pixels). Consequently, the model will overweight sparse $\boldsymbol{\theta}$ values in its results; and vice versa, if $\alpha_0 > 1$, sparse solutions will be underweighted in the results. This is useful if the researcher has some prior knowledge of the sample sparsity, which is not an uncommon scenario. Finally, it is well known that Dirichlet distribution is a conjugate prior to several exponential family probability distributions, which facilitates the computation of many statistical problems. Indeed, as shown in *SI Appendix*, this choice simplifies the expression of $\boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{u};\boldsymbol{g})$.

Our aim is to understand the statistical properties of $\boldsymbol{\theta}$. For that purpose, Eq. **3** is not very useful by itself, despite its relative simple form. Instead, we devised a method to numerically evaluate this distribution via a Gibbs sampling scheme. The idea is to iteratively sample $\boldsymbol{\theta}$ and $\boldsymbol{u}$ from 2 conditional probability distributions:

1) Draw one sample of $\boldsymbol{u}^{[t+1]}$ from $p(\boldsymbol{u}|\boldsymbol{\theta}^{[t]};\boldsymbol{g})$;
2) Draw one sample of $\boldsymbol{\theta}^{[t+1]}$ from $p(\boldsymbol{\theta}|\boldsymbol{u}^{[t+1]};\boldsymbol{g})$;
3) Repeat.

According to the principle of Gibbs sampling, such a scheme will produce samples from the joint distribution $p(\boldsymbol{\theta},\boldsymbol{u}|\boldsymbol{g})$. By simply discarding all $\boldsymbol{u}$ values from the results, we end up with samples from the distribution $p(\boldsymbol{\theta}|\boldsymbol{g})$.

The sampling scheme takes advantage of the fact that, unlike $p(\boldsymbol{\theta}|\boldsymbol{g})$, the 2 conditional probability distributions, $p(\boldsymbol{u}|\boldsymbol{\theta};\boldsymbol{g})$ and $p(\boldsymbol{\theta}|\boldsymbol{u};\boldsymbol{g})$, can both be expressed in simple analytical forms without complex normalization factors. As shown in *SI Appendix*,

$$p(\boldsymbol{u}|\boldsymbol{\theta};\boldsymbol{g}) = \prod_i \frac{\theta_{u_i} g_{i,u_i}}{\sum_j \theta_j g_{i,j}}, \quad [5]$$

and

$$p(\boldsymbol{\theta}|\boldsymbol{u};\boldsymbol{g}) = \text{Dir}(\boldsymbol{\theta} : \boldsymbol{\alpha}_0 + \boldsymbol{U}), \quad [6]$$

where $\boldsymbol{U}$ denotes the histogram counts of $\boldsymbol{u}$ in all pixels, i.e., U_j is the number of $\boldsymbol{u}$'s elements that equal to j.

The first equation (Eq. **5**) is simply the product of multiple categorical distributions. Algorithms for sampling from categorical distributions are well known. The second equation indicates that samples of $\boldsymbol{\theta}$ should be drawn from the Dirichlet distribution. The problem can be converted to a much simpler one by drawing parallel samples from multiple Gamma distributions. In Algorithm 1, we outlined the pseudocode for the implementation of this sampling scheme.

Finally, we want to point out that in the above formulation we assumed that throughout the experiment, all localizations are detected following the same molecular distribution $\boldsymbol{\theta}$, which may not be true in reality. The current cohort of SMLM techniques can be roughly divided into 2 groups: the first group, represented by PAINT (point accumulation in nanoscale topography) (8, 9), localizes single molecules through reversible bindings of fluorescent probes; while the second group, represented by STORM (stochastic optical reconstruction microscopy) (10) and PALM (photoactivation localization microscopy) (7), localizes single molecules by switching on or off the fluorescent states of the molecules. While our assumption is satisfied exactly in the first group, in the second group, each localization induces a finite probability of photobleaching; therefore, the $\boldsymbol{\theta}$ values fluctuate slowly throughout the experiment. Here, we assume that the probability distribution of interest can be approximated using an "average" $\boldsymbol{\theta}$ that is constant over time. In *Properties of* $\hat{\boldsymbol{\theta}}$ *and* $\overline{\boldsymbol{\theta}}$, we will briefly examine numerically how this assumption affects the accuracy of the algorithms.

Parameter Estimation

An important motivation to model $p(\boldsymbol{\theta}|\boldsymbol{g})$ is to obtain a numerical estimation of $\boldsymbol{\theta}$ based on experimental data. Here, we discuss 2 potential estimators.

The first method to estimate $\boldsymbol{\theta}$—since we already discussed the method to sample the posterior distribution—is by calculating the sample average $(1/N)\sum_i^N \boldsymbol{s}_i$, where $\boldsymbol{s}_i$ values are individual samples of $\boldsymbol{\theta}$. With a sufficient number of samples, the result approaches the so-called posterior expectation estimator, also known as the MMSE (minimum mean square error) estimator:

$$\overline{\boldsymbol{\theta}} \overset{def}{=} \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{g}) d\boldsymbol{\theta}. \qquad [7]$$

The second method is to compute the mode of the posterior distribution, resulting in the MAP (maximum a posteriori) estimator:

$$\hat{\boldsymbol{\theta}} \overset{def}{=} \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{g}). \qquad [8]$$

Although in principle the mode of the distribution can also be obtained from a sufficient number of samples, it is significantly faster to compute it via an EM (expectation-maximization)-type optimization algorithm. The key equation for this method is the iterative updating formula (see *SI Appendix* for the full derivation):

$$\hat{\boldsymbol{\theta}}_j^{[t+1]} \propto \sum_i \frac{\hat{\boldsymbol{\theta}}_j^{[t]} g_{i,j}}{\sum_k \hat{\boldsymbol{\theta}}_k^{[t]} g_{i,k}} + \alpha_0 - 1, \qquad [9]$$

where $\hat{\boldsymbol{\theta}}^{[t]}$ is a previous guess of the molecular distribution and $\hat{\boldsymbol{\theta}}^{[t+1]}$ is a newly updated and better guess. For this iterative algorithm to work, one needs an initial guess of $\hat{\boldsymbol{\theta}}^{[0]}$ to start the iteration. A generic choice is to use a uniform image, i.e., by setting all elements of $\hat{\boldsymbol{\theta}}^{[0]}$ to the same value, which we found to work well.

One complication in applying Eq. **9** is that if one starts with a sparse prior $(\alpha_0 < 1)$, the iterations potentially could produce negative numbers in some pixel positions. Negative numbers are outside the support of the $\boldsymbol{\theta}$ distribution. One way to avoid such a situation is to take the largest optimization step one could possibly take within the support by setting the offending pixels to zero. Therefore, the iterative rule in this situation becomes the following:

$$\hat{\boldsymbol{\theta}}_j^{[t+1]} \propto \max\left(0, \sum_i \frac{\hat{\boldsymbol{\theta}}_j^{[t]} g_{i,j}}{\sum_k \hat{\boldsymbol{\theta}}_k^{[t]} g_{i,k}} + \alpha_0 - 1\right) \text{ if } \alpha_0 < 1. \qquad [10]$$

Readers familiar with statistical models with L1-regularization (11, 12) would recognize that Eq. **10** is essentially a soft-thresholding operator commonly seen in those types of problems. This connection is somewhat reaffirming, because L1-regularization is typically used to enforce sparsity in solutions. Here, by adopting a sparse Dirichlet prior, we are trying to achieve the same goal. Therefore, the 2 approaches resulting in a similar type of mathematical operation is not entirely unexpected.

Finally, we would like to point out a connection between the 2 estimators discussed here (MAP and MMSE estimator) and the current image-rendering practices in the SMLM field. To start, one can consider an extreme scenario where all localizations obtained in an experiment are far apart from each other, so that the correlation between different localizations can be safely ignored and the parameter estimation can be done one localization at a time. For a single measured localization, the MAP estimator is the localization centroid position itself; thus, the computed $\hat{\boldsymbol{\theta}}$ image is the histogram representation of the localization data (i.e., the pixel intensity is directly proportional to the number of localizations detected in this location). Conversely, the MMSE estimator for a single localization is a Gaussian spot, whose width is governed by the localization uncertainty. Therefore, the image representation is the Gaussian-spot representation. Both representations are used in the current practices. Of course, neither the histogram representation nor the Gaussian-spot representation considers correlative information between different localizations and thus is not an accurate representation of the full information conveyed by the data.

Properties of $\hat{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\theta}}$

To evaluate the numerical properties of the algorithms proposed above, we first performed some simple computational tests on simulated SMLM data. We considered a simple ring-shaped structure (Fig. 1*A*) with 40-nm outer diameter and 20-nm inner diameter. We generated localization data following 2 different data models: The PAINT data model assumed that the molecules are localized within the ring structure with uniform probability, while the STORM data model assumed that there were a fixed number of fluorescent molecules (200) within the structure and each localization has a 5% probability to irreversibly photobleach the molecule. For both cases, we assumed the number of photons produced in each localization followed a geometric distribution, but if the intensity was lower than 20% of the mean, the data were discarded. This resulted in an average localization uncertainty of 14.3 nm, with a SD of 5.5 nm. SMLM images were generated with a various number of localizations, from 30 to 3,000.

In Fig. 1*B*, we show examples of the $\boldsymbol{\theta}$ inference results. The original localization data were plotted in both the histogram representation as well as the Gaussian-spot representation, which can be compared with $\hat{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\theta}}$ images, respectively. It can be seen that the $\boldsymbol{\theta}$ images deviate from their corresponding SMLM representations at even the lowest localization density. More importantly, at a higher localization density, the $\boldsymbol{\theta}$ images reflect the ground truth with higher fidelity than the direct renderings of the SMLM data. In the images directly rendered from the raw dataset, the center holes are obscured due to the localization errors, in contrast to the computed $\boldsymbol{\theta}$ images, in which the holes are clearly visible. This improvement cannot be achieved by simply excluding the localizations with larger errors in the direct rendering. In fact, direct renderings after removing up to 80% of the localizations (*SI Appendix*, Fig. S1) do not significantly

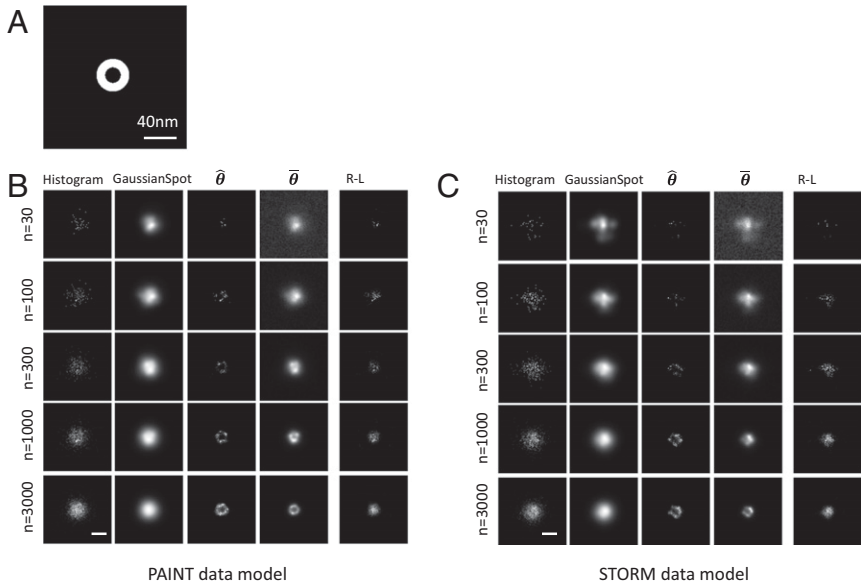

Fig. 1. Numerical validation of θ estimation algorithms with synthetic SMLM data. (*A*) The ground truth for the spatial distribution of the imaging targets. (*B*) Simulated SMLM images in either the histogram representation (first column) or the Gaussian-spot representation (second column). The $\hat{\theta}$ images (third column) were computed from the SMLM data with 50 iterations. The $\bar{\theta}$ images (fourth column) were computed from a single run with 10,000 sampling iterations following a 1,000-iteration burn-in run. The Richardson–Lucy deconvolution results (fifth column) of the same SMLM image (histogram representation) were also plotted for comparison. All computations were performed on an 8-nm grid. The *n* values represent different numbers of localizations used for each SMLM image. (*C*) Same as *B*, except the SMLM data were generated via STORM data model instead of PAINT data model. See the main text for details of the 2 data models.

improve the image quality with this dataset. Instead, the statistical algorithm achieves its results mainly by identifying correlations between multiple localizations and "correct" localization errors based on the correlative information. Similar simulations on dataset derived from other types of structures (e.g., cross and disk; *SI Appendix*, Figs. S2 and S3) produced images with similar qualitative improvements. Combined, these results are consistent with the notion that $\boldsymbol{\theta}$ images computed from the model represent a kind of "deconvoluted" images of the original SMLM data. Despite this connection, performing standard Richardson–Lucy deconvolution (Fig. 1*B*), using a Gaussian PSF corresponding to the average localization uncertainty, produced inaccurate results, indicating that accurately accounting for the variations of localization uncertainties is important for the computation.

Visual comparison between the PAINT model and the STORM model suggests that the algorithms produced $\boldsymbol{\theta}$ estimations with similar accuracy in those 2 cases. More quantitative comparisons can be made by computing mean square error (MSE) of the estimators against the ground truth (Fig. 2). Our results show that the estimators computed from the STORM data exhibit a slightly larger MSE, likely due to the aforementioned approximation of the model. Nevertheless, the algorithms are effective in both cases in producing images with reduced errors.

The MAP estimator $\hat{\boldsymbol{\theta}}$ exhibited a biphasic convergence behavior, in which MSE decreased in early iterations and increased again later (Fig. 2 *A* and *B*). It is well known that nonlinear deconvolution algorithms, such as the Richardson–Lucy algorithm, exhibit a similar behavior (13, 14). The mechanism can be understood by examining the noise statistics at specific spatial frequencies, as shown in Fig. 3. Here, we computed the signal-to-noise ratio (SNR), defined as the ratio of the total variance of the ground truth signal against the variance of the noise (difference between the ground truth and the estimators), at each iteration. It can be seen that there was significant SNR recovery at lower frequencies in early iterations, a delayed recovery at intermediate frequencies, and a continuous deterioration at higher frequencies. In other words, the early iterations recovered the signal (at lower frequencies) and the later iteration tried to overfit the noise (at higher frequencies). Interestingly, the results also indicated that the main reason that the estimator performed worse for STORM data is due to a faster deterioration at higher frequencies (Fig. 3). In the deconvolution literature, it is clear that the most effective way to stop overfitting of the noise is to perform the optimization with a regularization term (15, 16), which requires some prior knowledge about the image. We expect a similar strategy will be effective in the SMLM case, due to the similarity of the 2 problems.

The $\bar{\boldsymbol{\theta}}$ approaches a minimal MSE asymptotically, but also exhibits a much larger residue MSE (Fig. 2 *C* and *D*), and therefore may not be the preferred estimator in most cases. However, for very high localization densities (e.g., $n = 3{,}000$), $\bar{\boldsymbol{\theta}}$ can reach an asymptotic error close to that of $\hat{\boldsymbol{\theta}}$.

One experimental factor that can potentially impact the accuracy of the $\boldsymbol{\theta}$ estimation is the accuracy of uncertainty estimations. As mentioned earlier, most of the SMLM software today provide estimations for each localization uncertainty; however, they typically do this by computing the theoretical CRLB limit assuming shot-noise–limited signals and a uniform background. Furthermore, one of the most important experimental factors affecting the localization uncertainty is the number of photons detected, which needs to be computed from calibrations. Extra photon noise, uneven background, or inaccurate calibration all can introduce errors in the reported localization uncertainty. To assess how these errors impact the accuracy of the $\boldsymbol{\theta}$ estimations, we performed computations on simulated data with various amounts of relative errors in the estimations of the localization uncertainties (*SI Appendix*, Fig. S4). We found that small relative errors (below 20%) have virtually no effects on the accuracy of the $\boldsymbol{\theta}$ estimators (based on the MSE measures). However, once the uncertainty errors reach ~30%, there is a significant jump in the MSEs of the estimators, especially $\hat{\boldsymbol{\theta}}$, pointing to an upper limit of the allowed uncertainty errors for this algorithm.

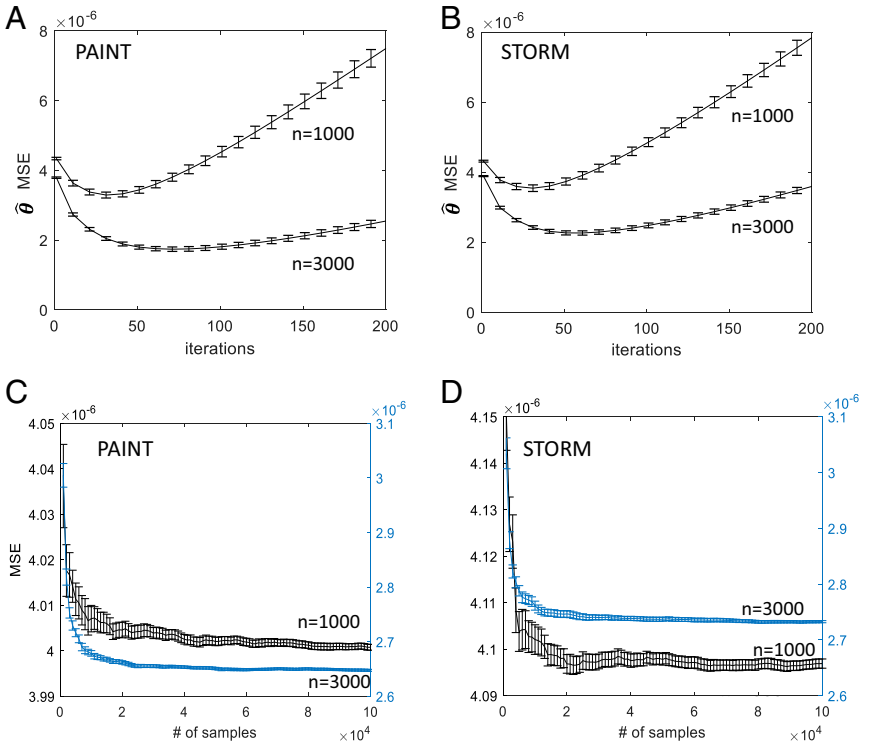

Fig. 2. MSEs of the $\boldsymbol{\theta}$ estimators. (*A* and *B*) MSEs of the $\hat{\boldsymbol{\theta}}$ estimators computed against the ground truth at every iteration for data according to either the PAINT data model (*A*) or STORM data model (*B*). The results shown were the average of 50 independent simulated images. (*C* and *D*) MSEs of the $\bar{\boldsymbol{\theta}}$ estimators computed against the ground truth at every 1,000 samples for data according to either PAINT data model (*C*) or STORM data model (*D*). The results are from 30 independent runs.

Next, we tried to get a rough estimation of the effective resolution of $\hat{\boldsymbol{\theta}}$. The effective resolution of the original SMLM data can be evaluated via computing Fourier ring correlation (FRC) between 2 subsampled SMLM images (17, 18). However, this cannot be done with $\hat{\boldsymbol{\theta}}$. Instead, we performed FRC between $\hat{\boldsymbol{\theta}}$ and the ground truth image (Fig. 4). In this case, the commonly used FRC threshold of 0.143 corresponds to a corrected threshold of 0.5 (19), due to the absence of the subsampling. Based on this measure, we found that the $\hat{\boldsymbol{\theta}}$ images exhibit a resolution that is roughly 2 times the original data when the localization density is high (Fig. 4, *Right*; localization density, ~1.06 nm^{-2}). Lower localization density also allows recovery of some high-resolution features (Fig. 4, *Left*), but the correlation factor does not reach the threshold of 0.5, indicating that the confidence of the recovered features is low. Most SMLM experiments available today do not have a localization density that is comparable to the high end of the simulations here. This limitation will impact the performance of $\boldsymbol{\theta}$ computation from experimental data. For example, in *SI Appendix*, Fig. S5, we plotted experimental STORM images of microtubules in COS cells, as well as the corresponding $\boldsymbol{\theta}$ images. The computed images exhibited some expected improvements. For example, the hollow cores of the microtubules were more prominent in the $\hat{\boldsymbol{\theta}}$ images. Meanwhile, other spatial features in $\hat{\boldsymbol{\theta}}$, such as the spotty appearance along the longitudinal axis of the microtubules, were unexpected and may be due to overfitting artifacts.

Application: Particle Fusion

Although direct $\boldsymbol{\theta}$ estimation is the most obvious application of the proposed model, we would like to suggest that many other inference problems involving SMLM will also benefit from our model. In our view, the significance of the model is that it provides a mathematical foundation for interpreting SMLM datasets, which is generally useful when working with SMLM data. Here, we demonstrate one such idea in the case of SMLM particle fusion.

Particle fusion aligns many noisy images of identical particles to produce a high-quality image. The technique originated from the electron microscopy (EM) field but recently has been adopted to superresolution optical microscopy data (20–22). Despite its origin in EM, it has been recognized that the existing EM algorithms may not work well for SMLM data, due to the vastly different noise characteristics (23, 24). Heydarian et al. (23) recently published an algorithm tailored for SMLM, which performs significantly better than previous software. The core concept of the new method is to register single-particle SMLM images reconstructed with a Gaussian kernel, which allows fast image registration computations. In addition, to improve accuracy, registrations are performed between all pairs of particles. It therefore has a $O(N^2)$ type complexity versus the particle number.

We hypothesized that a better result may be obtained by registering particle images against $\boldsymbol{\theta}$, instead of other SMLM images, because $\boldsymbol{\theta}$ is a better "resolved" image. Furthermore, in this case, the data likelihood serves as a good cost function for image registration:

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{g},\boldsymbol{t}) \overset{def}{=} -\log p(\boldsymbol{g}|\boldsymbol{\theta},\boldsymbol{t}) = -\sum_i \log p\left(g_{i,\cdot}|\boldsymbol{\theta},\boldsymbol{t}\right). \qquad [11]$$

Here, $\boldsymbol{t}$ represents all parameters of the rigid affine transformation used for the registration. It typically has only 3 degrees of freedom, 2 for the translations and 1 for the rotation. Furthermore, the term $p(g_{i,\cdot}|\boldsymbol{\theta},\boldsymbol{t})$ can be intuitively understood as the following: If we perform, on $\boldsymbol{\theta}$, a Gaussian blur operation,

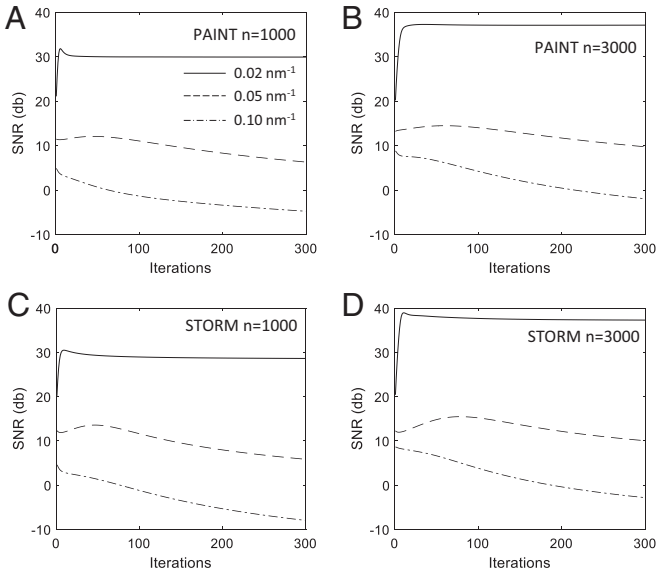

Fig. 3. SNR recovery by $\hat{\theta}$ at different spatial frequencies. Frequency-dependent SNRs were computed with the power spectra of the ground truth image and the computed $\hat{\theta}$ images using a 0.22 nm^{-1} integration bandwidth. The results from PAINT data model are shown in *A* and *B*. The results from STORM data model are shown in *C* and *D*. Note that the peak SNR recovery for STORM data are comparable to PAINT data at low and intermediate frequencies. However, the deterioration of the SNR at high frequencies is worse for the STORM data.

parametrized by the localization uncertainty specified in $g_{i,\cdot}$, followed by an affine transformation according to the parameters in $\boldsymbol{t}$, then the probability we are looking for is the pixel value at the observed position of the ith localization (see *SI Appendix* for more detailed discussions). Thus, the cost function we use here takes into consideration the relative importance of high accuracy localizations versus low accuracy localizations. For high accuracy localizations, the $\boldsymbol{\theta}$ image should have sharp features that vary quickly over space, and therefore a small change in the transformation parameters would incur a large penalty/gain in the cost function value. In contrast, for low accuracy localizations, the transformations would have a much smaller impact on the cost function values due to the Gaussian blurring.

With this insight, we designed an iterative particle fusion algorithm. We first performed a rough alignment of particles using 1-to-N registration and used the result to compute the posterior distribution of $\boldsymbol{\theta}$. We then tried to get a better registration of each particle iteratively, namely:

$$\boldsymbol{t}^{k,[i+1]} \leftarrow \arg\min_{\mathbf{t}} \int \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{g}^k, \boldsymbol{t}) \, p\left(\boldsymbol{\theta} | \boldsymbol{g}^1, \boldsymbol{g}^2, \ldots, \boldsymbol{g}^N, \boldsymbol{t}^{1,[i]}, \boldsymbol{t}^{2,[i]}, \ldots, \boldsymbol{t}^{|N,[i]}\right) d\boldsymbol{\theta}. \quad [12]$$

Here, we use superscript k to denote particle numbers. The integration over the complete posterior distribution would ensure the convergence of the iterative algorithm. This is again an EM-type iteration algorithm. A brief proof of correctness is supplied in *SI Appendix*. Because direct integration in Eq. **12** is intractable, we approximate the integration by summation over samples drawn from the $\boldsymbol{\theta}$ distribution using the Gibbs sampler. The optimization of $\boldsymbol{t}$ itself can be achieved using any generic purpose optimization algorithms. We implemented it using the simple gradient descent function in MATLAB. In addition, as we pointed out previously in a related work concerning image alignment (25), a trick to improve convergence speed is to artificially lower effective resolution in early iterations. We implement this here by scaling the localization uncertainty up by a few folds in the first iteration and reducing it to real values in later iterations.

To test the validity of this algorithm, we performed particle fusion calculation on a set of experimentally acquired SMLM particle data on DNA origami structures, also published by Heydarian et al. (23). In this set of data, the original authors designed a DNA origami that folds in a way to display 37 total oligo binding sites, which in turn are organized to spell out a "TUD" logo of the authors' institute. The locations of these sites were probed via the DNA-PAINT method; however, a certain number of the sites were purposely inactivated to mimic the low degree of labeling (DOL) effect in many SMLM experiments. The complete dataset contains both high-quality data with 80% DOL as well as low-quality data with 50% and 30% DOL.

Fig. 5 shows our computed particle fusion results for the 30% DOL particles. The initial registration is generated by an 1-to-N registration, i.e., we simply aligned all particles against one randomly chosen one from the ensemble. The result, as can be seen in Fig. 5, is quite inaccurate. Despite this poor starting point, the iterations rapidly improved the registration quality and the algorithm converged with 4 iterations, after which the FRC resolution no longer improved (Fig. 5*C*). We then removed the particles (27 out of 549) that are considered outliers based on their likelihood values, and obtained a final FRC resolution of 3.5 ± 0.3 nm (Fig. 5*B*), which is a statistically significant improvement over the original result (5 nm) (23). We believe this

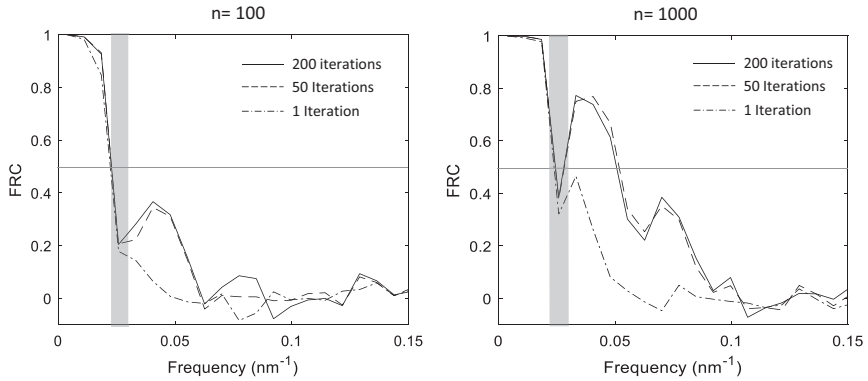

Fig. 4. FRC resolution of $\hat{\theta}$. FRC against the ground truth image were computed for $\hat{\theta}$ images recovered at low (*Left*) or high (*Right*) localization densities, using PAINT data model. The horizontal line denotes the expected 0.5 FRC threshold, equivalent to the commonly used 0.143 threshold when data are subsampled. The vertical bars indicate the FRC resolution of the original SMLM images. The width of the bar represents the SD.

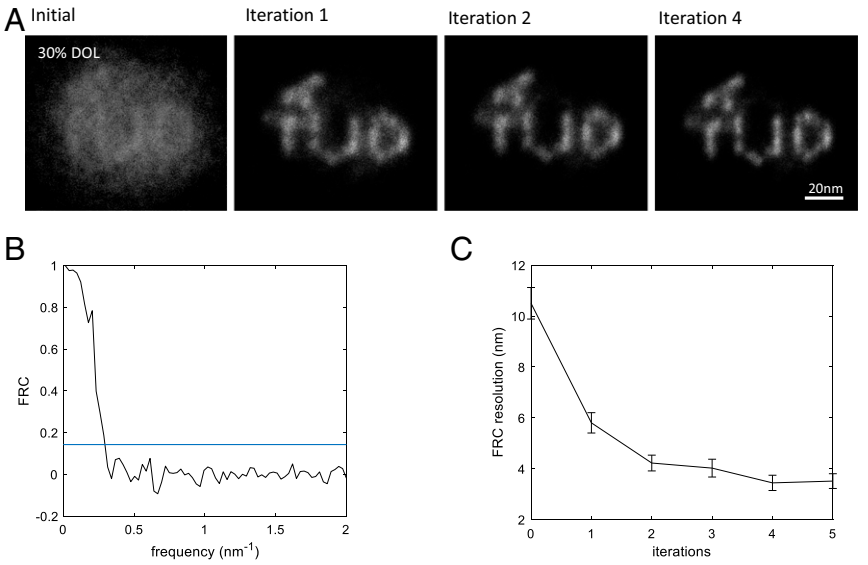

Fig. 5. Particle fusion of origami images. (*A*) The results of the particle fusion at each iteration for the low-quality (DOL 30%) origami SMLM data. The initial registration (*Left*) was generated by simple 1-to-*N* registration of all available particles using mutual information as a cost function. The rest of the images showed the fusion result after 1, 2, and 4 iterations using the algorithm described in the main text. Each iteration drew 25,000 samples from the posterior distribution using a 1.3-nm grid and used 5,000 (uniform thinning) for registration computation. (*B*) Final FRC computed after iteration 4. (*C*) The FRC resolution of the particle data at each iteration, showing no further improvement after fourth iteration.

improvement is mainly due to the fact that we performed registration against $\boldsymbol{\theta}$, instead of other SMLM images. Furthermore, since the algorithm scales with particle numbers with O(N) complexity, it runs reasonably fast on cheap equipment (1 to 2 h per iteration on an i5 laptop).

Performing the fusion calculation on high-quality data (DOL = 80%) also yielded satisfactory results (Fig. 6). The computation converged in 2 iterations as the starting point was significantly more accurate. The final FRC resolution, after removing 9 outliers out of original 383 particles, is 3.3 ± 0.2 nm, which is identical to that of the original report.

Finally, we present the $\hat{\boldsymbol{\theta}}$ images computed from the fusion results (Fig. 6). These images exhibited reduced blurriness as expected. However, it is evident that the binding sites, which physically are point-objects at the scale of interest, still exhibited relatively large spatial spread on the order of several nanometers. We suspected that this is primarily due to the dynamic molecular motion of individual sites and not the registration errors. The published localization uncertainty is determined from image fitting and CRLB; therefore, it does not take into account the fact that the positions of the target sites may fluctuate in different rounds of localizations due to low-frequency motions. Therefore, the reported localization errors are likely an underestimation. We tested whether we could compensate for such errors by slightly increasing the localization uncertainty (varying from 0.4 to 1 nm^2 in terms of additional variance). We found that such changes had relatively little effect on the particle fusion accuracy, i.e., the final FRC resolution did not change in a statistically significant manner. However, improvements in the $\hat{\boldsymbol{\theta}}$ images computed from the fusion results are evident visually (Fig. 6). In particular, the individual binding sites in the 30% DOL particles are much better resolved when this factor is introduced. It is interesting that the histogram representation of the fusion particle itself exhibited little visual indications that the data were in fact originated from point-sources. Nevertheless, the apparently blurry appearance is misleading, partly because it ignores correlations between multiple localizations, and partly because it displays high-accuracy localizations on equal footing with low-accuracy localizations. Once the localization accuracies were taken into consideration, the data contain sufficient statistical information at very high spatial frequency, allowing individual sites to be resolved.

Conclusion and Discussion

The rise and rapid adoption of SMLM have led to many new problems that are linked to these techniques' unique noise characteristics. To some degree, even a simple question such as "what is the resolution" is a reflection of the fact that there exists a complex and nondeterministic relationship between the observed data and the underlying molecular distribution. It is these

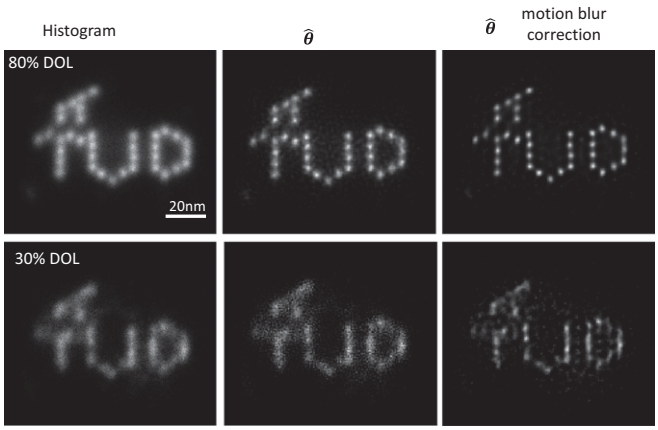

Fig. 6. Post-particle fusion θ restoration. Particle fusion results for both the high-quality (DOL 80%) data (top row) and the low-quality (DOL 30%) data (bottom row). The left column shows the histogram representation of the SMLM data after 2 (*Top*) and 4 (*Bottom*) iterations of registration. The middle column shows the corresponding $\hat{\theta}$ images (50 iterations) computed using the originally supplied localization uncertainty data, which were computed based on CRLB. The right column shows the $\hat{\theta}$ images (50 iterations) by using a slightly increased localization uncertainty to compensate for potential molecular motions. For 80% DOL, the variance of the localization uncertainty were increased by 0.8 nm^2, and for 30% DOL, 1.0 nm^2.

questions that motivated us to investigate the statistical properties of SMLM data. To this end, we believe our current model is a useful one, as we have shown that it provides reasonably accurate inference results on both synthetic and experimental data.

A direct application of the proposed model is using algorithms to perform inverse calculation from experimental data to $\boldsymbol{\theta}$. We have shown that the usual representations of the SMLM data, either the histogram type or the Gaussian-spot type, can obscure the true resolving power of the data, especially at high localization density (>0.1 localization per nm^2). At the current stage, SMLM data that surpass this density may still be rare, but as the techniques further improve, either because of the development of better PAINT probes or due to further reduction of photobleaching in STORM-type experiments, the status may change, at which point, a $\boldsymbol{\theta}$-based representation may be more suitable.

There is one area where the localization density is already high enough to warrant $\boldsymbol{\theta}$ representation: namely, particle fusion. Here, we show that our model is particularly useful for this type of data and can further be extended to infer image registration parameters. Indeed, the registration method based on our model appears to outperform existing algorithms in both speed and accuracy.

When localization density is not at a very high level, we found that the MAP estimator $\hat{\boldsymbol{\theta}}$ usually outperforms the MMSE estimator $\bar{\boldsymbol{\theta}}$ in its ability to resolve finer details. However, one should be careful with the interpretation of the MAP estimators due to their tendency of overfitting the data, which results in artifacts at high spatial frequencies. In principle, one could compute the confidence level of any spatial feature by drawing samples from the posterior distribution and finding out what percentages of the samples exhibit the feature of interest. This assumes that a mathematical test could be devised for the said feature (not always feasible). Ultimately, empirical researchers will likely rely on properly designed control experiments for the validation purpose.

An important requirement for the model is that the data should have reasonably accurate estimation of localization uncertainties for each localization. While most SMLM fitting software today provide such numbers, the uncertainty values are usually computed from theoretical limits and not validated with experimental evidence. One potential method to (partially) validate the uncertainty numbers is to examine the subset of localizations in the dataset where the same molecule is detected in multiple consecutive frames. Presumably, the true location of the molecule is not changing in those frames. Thus, one can test whether the estimated uncertainties are consistent with the measured fluctuations. Nevertheless, it is fair to say that, over the last decade, most efforts in SMLM data analyses have been focused on improving fitting accuracy, and little attention has been paid to accurate estimation of the uncertainties themselves. Hopefully in the future, this problem will attract more attention and better methods for validating uncertainty measurements will be available.

Methods

All algorithms discussed in this paper were implemented on a mixed MATLAB/C++ platform and were tested with the linux64 version of MATLAB R2016R. Parameters for generating specific results shown in figures are described in their respective figure legends.

Code Availability. The current versions of our implementation of the various algorithms are available at https://github.com/jiyuuchc/lmdeconv (26).

1. Y. M. Sigal, R. Zhou, X. Zhuang, Visualizing and discovering cellular structures with super-resolution microscopy. *Science* **361**, 880–887 (2018).
2. G. C. Rollins, J. Y. Shin, C. Bustamante, S. Pressé, Stochastic approach to the molecular counting problem in superresolution microscopy. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E110–E118 (2015).
3. D. Nino, N. Rafiei, Y. Wang, A. Zilman, J. N. Milstein, Molecular counting with localization microscopy: A Bayesian estimate based on fluorophore statistics. *Biophys. J.* **112**, 1777–1785 (2017).
4. R. J. Ober, S. Ram, E. S. Ward, Localization accuracy in single-molecule microscopy. *Biophys. J.* **86**, 1185–1200 (2004).
5. C. S. Smith, N. Joseph, B. Rieger, K. A. Lidke, Fast, single-molecule localization that achieves theoretically minimum uncertainty. *Nat. Methods* **7**, 373–375 (2010).
6. A. R. Small, R. Parthasarathy, Superresolution localization methods. *Annu. Rev. Phys. Chem.* **65**, 107–125 (2014).
7. E. Betzig *et al.*, Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006).
8. R. Jungmann *et al.*, Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. *Nat. Methods* **11**, 313–318 (2014).
9. A. Sharonov, R. M. Hochstrasser, Wide-field subdiffraction imaging by accumulated binding of diffusing probes. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 18911–18916 (2006).
10. M. Bates, B. Huang, G. T. Dempsey, X. Zhuang, Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science* **317**, 1749–1753 (2007).
11. S. Chen, D. Donoho, M. Saunders, Atomic decomposition by basis pursuit. *SIAM Rev.* **43**, 129–159 (2001).
12. K. Bredies, D. A. Lorenz, Linear convergence of iterative soft-thresholding. *J. Fourier Anal. Appl.* **14**, 813–837 (2008).
13. D. L. Snyder, M. I. Miller, L. J. Thomas, D. G. Politte, Noise and edge artifacts in maximum-likelihood reconstructions for emission tomography. *IEEE Trans. Med. Imaging* **6**, 228–238 (1987).
14. S. Prasad, Statistical-information-based performance criteria for Richardson-Lucy image deblurring. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **19**, 1286–1296 (2002).
15. N. Dey *et al.*, Richardson-Lucy algorithm with total variation regularization for 3D confocal microscope deconvolution. *Microsc. Res. Tech.* **69**, 260–266 (2006).
16. M. Arigovindan *et al.*, High-resolution restoration of 3D structures from widefield images with extreme low signal-to-noise-ratio. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17344–17349 (2013).
17. R. P. J. Nieuwenhuizen *et al.*, Measuring image resolution in optical nanoscopy. *Nat. Methods* **10**, 557–562 (2013).
18. N. Banterle, K. H. Bui, E. A. Lemke, M. Beck, Fourier ring correlation as a resolution criterion for super-resolution microscopy. *J. Struct. Biol.* **183**, 363–367 (2013).
19. P. B. Rosenthal, R. Henderson, Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
20. A. Löschberger *et al.*, Super-resolution imaging visualizes the eightfold symmetry of gp210 proteins around the nuclear pore complex and resolves the central channel with nanometer resolution. *J. Cell Sci.* **125**, 570–575 (2012).
21. X. Shi *et al.*, Super-resolution microscopy reveals that disruption of ciliary transition-zone architecture causes Joubert syndrome. *Nat. Cell Biol.* **19**, 1178–1188 (2017).
22. A. Szymborska *et al.*, Nuclear pore scaffold structure analyzed by super-resolution microscopy and particle averaging. *Science* **341**, 655–658 (2013).
23. H. Heydarian *et al.*, Template-free 2D particle fusion in localization microscopy. *Nat. Methods* **15**, 781–784 (2018).
24. J. Schnitzbauer *et al.*, Correlation analysis framework for localization-based superresolution microscopy. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3219–3224 (2018).
25. A. Elmokadem, J. Yu, Optimal drift correction for superresolution localization microscopy with Bayesian inference. *Biophys. J.* **109**, 1772–1780 (2015).
26. J. Yu, Matlab/C++ code for analyzing SMLM data. GitHub. https://github.com/jiyuuchc/lmdeconv/tree/fffa9dfcb9c4c3dd290af2f84e1b06e9a61058d9. Deposited 23 July 2019.