LEARNING THE SPECIMEN STRUCTURE IN SINGLE-MOLECULE LOCALIZATION MICROSCOPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-molecule localization microscopy (SMLM) reconstructs three-dimensional organic specimens stained by fluorophores—molecules that emit light under laser excitation—beyond the diffraction limit by localizing active emitters in thousands of frames and accumulating their positions into a point cloud. However, conventional methods process short windows of frames independently, ignoring that all frames share the same underlying structure. In this paper, we propose a method that learns the structure of the specimen and uses it to improve emitter localization. We train a neural network on simulated data to detect and localize emitters, while simultaneously fitting a distribution over emitters to the localizations it predicts on real frames. We use this distribution to adapt the simulator throughout training, hence feeding information from the whole acquisition back into individual localizations. We jointly estimate the point spread function, the background, and the emitter photophysics via the same procedure, enabling self-calibration. After training, we obtain a network for detecting and localizing emitters, and a fitted spatial density that directly provides a continuous super-resolved representation of the specimen. On synthetic benchmarks, learning the specimen structure reduces localization errors compared with the same network trained with uniformly distributed emitters. On real acquisitions, the fitted density reveals fine specimen structure, showing smooth filaments and clearly defined nuclear pore rings.

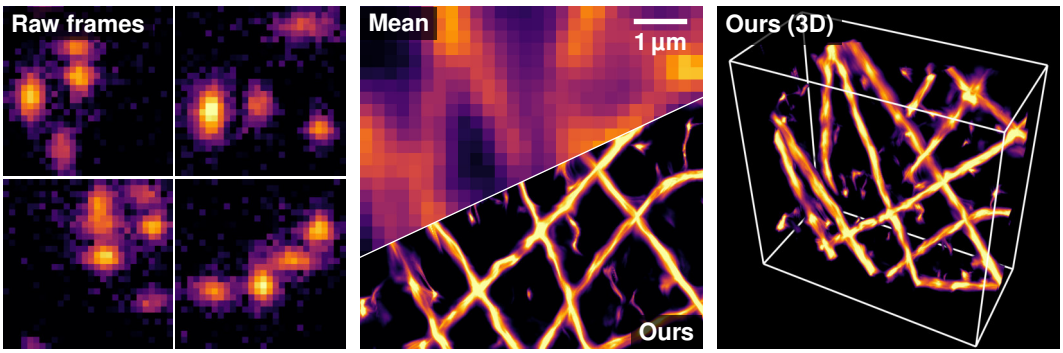

Figure 1: Our proposed super-resolved reconstruction from a real SMLM acquisition (Tubulin, see Section 5.2). **Left:** four raw frames (of $\sim 10^4$), each showing a few active emitters in the same field of view as the middle panel. **Middle:** the mean of the acquisition (top), illustrating diffraction-limited microscopy, versus our reconstruction (bottom), that is a learnt, continuous, 3D spatial density. **Right:** a 3D view of our reconstruction.

1 INTRODUCTION

Fluorescence microscopy reveals biological structures by recording light emitted by fluorescent labels stained on a specimen, called *fluorophores*, which we refer to as *emitters* throughout this manuscript. However, light diffraction blurs details below $200\,\mathrm{nm}$ for a standard

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

setup (Abbe, 1873; Schermelleh et al., 2019). Single-molecule localization microscopy (SMLM) circumvents this barrier by activating only a sparse subset of emitters per frame, so that their images rarely overlap (Betzig et al., 2006; Hess et al., 2006; Rust et al., 2006; Lelek et al., 2021). Isolated emitters can then be localized with precision far below the diffraction limit (Ober et al., 2004). Accumulating localizations over thousands of frames yields a point cloud representation of the specimen.

Formally, SMLM divides the single problem of explaining an entire acquisition into thousands of smaller problems: recovering, for each frame $\boldsymbol{y}$, the emitters $\boldsymbol{x} = \{(\boldsymbol{r}_k, \alpha_k)\}_{1 \le k \le K}$ that it contains, where $\boldsymbol{r}_k \in \mathbb{R}^3$ is a position, $\alpha_k \in \mathbb{N}$ a photon count, and K is an unknown number of emitters. State-of-the-art methods (Speiser et al., 2021; Fei et al., 2025; Séailles et al., 2026; Fu et al., 2026) rely on a simulator: they generate pairs $(\boldsymbol{x}, \boldsymbol{y}) \sim p(\boldsymbol{x}, \boldsymbol{y})$ by drawing emitter positions and photon counts, rendering their images through a point spread function (PSF) model, and adding camera noise. They then train a neural network $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ on synthetic data to approximate the posterior $p(\boldsymbol{x} \mid \boldsymbol{y})$ induced by the simulator. At inference, q_θ is applied to each frame of the acquisition, the resulting distributions are collapsed to point estimates, and all are ultimately gathered into a single point cloud.

This procedure has two limitations. First, the subproblems are treated independently, ignoring that they all image the same specimen, whose structure could inform detection in each frame. Second, existing methods almost always keep the simulator fixed throughout training, and any misspecification in the components of the generative model may degrade performance (Ward et al., 2022). In particular, their prior $p(\boldsymbol{x})$ places emitters uniformly and draws their *photophysics*, i.e. how many photons they emit and when they switch on and off, from preset distributions (Sage et al., 2019; Speiser et al., 2021). Likewise, the imaging system $p(\boldsymbol{y} \mid \boldsymbol{x})$ requires the PSF and the background (Li et al., 2018; Liu et al., 2024), except for LUNAR (Fu et al., 2026), which learns the PSF during training.

In this paper, we address both limitations by adapting the simulator to the acquisition we want to reconstruct. We introduce $q_\phi(\boldsymbol{x})$, a parametric distribution over emitters that is a learnable counterpart of the prior $p(\boldsymbol{x})$. We also introduce a distribution $q_\psi(\boldsymbol{c})$, where $\boldsymbol{c}$ contains the shared imaging parameters to estimate, including the PSF parameters and the background (Eq. (1)). Each training step minimizes two objectives. First, two neural posteriors, $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ and $q_\eta(\boldsymbol{c} \mid \boldsymbol{y})$, are trained to estimate emitters $\boldsymbol{x}$ and imaging parameters $\boldsymbol{c}$ on simulated frames $\boldsymbol{y}$. Second, their predictions on frames of the acquisition of interest are used to fit $q_\phi(\boldsymbol{x})$ and $q_\psi(\boldsymbol{c})$, which thus summarize what has been inferred across the whole acquisition: $q_\phi(\boldsymbol{x})$ captures the structure of the specimen and the photophysics of emitters, while $q_\psi(\boldsymbol{c})$ combines information about the PSF and the background. As training progresses, the simulator gradually shifts its sampling from the preset priors to these fitted distributions, feeding knowledge from all frames back into the neural posteriors. This scheme is similar to the wake-sleep algorithm (Hinton et al., 1995), and it tailors the learnt emitter distribution $q_\phi(\boldsymbol{x})$ to the acquisition we want to reconstruct.

Overall, our method makes three contributions.

1. **Structure-informed localization.** The fitted distribution $q_\phi(\boldsymbol{x})$ carries information from the whole acquisition into the inference on each frame. We show that the neural posterior detects and localizes emitters more accurately when trained under this distribution than under a fixed one.
2. **Self-calibration.** We jointly estimate the PSF, the emitter photophysics and the background from the acquisition, during training. Only the camera response parameters are supplied by the user.
3. **Reconstruction as part of training.** The spatial marginal of $q_\phi(\boldsymbol{x})$ is a density over the field of view, that is itself a super-resolved representation of the specimen, see Figure 1. It is obtained during training rather than as a post-processing step.

We show that our method improves detection and localization performance compared to the same training under a fixed prior, and compared to state-of-the-art methods (DECODE (Speiser et al., 2021), LiteLoc (Fei et al., 2025), SHOT (Séailles et al., 2026) and LUNAR (Fu et al., 2026)) on the open datasets of the SMLM 2016 Challenge (Sage et al., 2019) and on experimental acquisitions.

2 Related work

Single-molecule localization microscopy (SMLM). Among far-field super-resolution methods for microscopy (Hell & Wichmann, 1994; Gustafsson, 2000; Dertinger et al., 2009; Müller & Enderlein, 2010; Gustafsson et al., 2016; Balzarotti et al., 2017; Laine et al., 2023), SMLM reaches a resolution

of ~ 20 nm at the cost of long acquisition times (Betzig et al., 2006; Hess et al., 2006; Lelek et al., 2021). SMLM methods have moved from per-frame fitting models (Babcock et al., 2012; Ovesnỳ et al., 2014) to neural networks trained on simulations (Nehme et al., 2018; Boyd et al., 2018; Speiser et al., 2021; Séailles et al., 2026). They mainly differ in their output and training objective: a super-resolved image (or volume) (Nehme et al., 2018; 2020), a set of localizations trained under a maximum mean discrepancy (Gretton et al., 2012; Boyd et al., 2018), a pixel-wise detection map with sub-pixel localization and uncertainty trained under relaxed NPE objectives (Speiser et al., 2021; Fu et al., 2023; Fei et al., 2025; Séailles et al., 2026; Fu et al., 2026). In all cases, the reconstruction of the specimen is obtained after training by accumulating the localizations, and the prior over emitters used to simulate the training data is fixed. Ouyang et al. (2018) has trained a network under other reconstructions to complete sparse ones. Cox et al. (2012) has proposed joint Bayesian modelling of a whole acquisition, but at a computational cost that restricts it to small fields of view.

Calibrating the PSF. PSFs are usually measured on fluorescent beads before the acquisition (Liu et al., 2013; Babcock & Zhuang, 2017; Li et al., 2018; Ferdman et al., 2020), or from the acquisition itself in a preprocessing step (Xu et al., 2020; Liu et al., 2024). LUNAR (Fu et al., 2026) fits the PSF parameters during training, alternating between learning a posterior over emitters from simulations and fitting the imaging model to real frames using the inferred localizations. By contrast, we treat PSF parameters as latent variables, like emitter positions and photon counts, without reconstructing real frames. We then fit distributions to quantities inferred by an amortized posterior: the emitter spatial density, their photophysics, and the PSF model are all recovered with the same procedure.

Neural posterior estimation (NPE). NPE (Papamakarios & Murray, 2016; Cranmer et al., 2020) is a branch of simulation-based inference (SBI) that approximates an intractable posterior by a neural network trained on data simulated from a generative model $p(\boldsymbol{x}, \boldsymbol{y})$. It powers many scientific applications (Dax et al., 2021; Vasist et al., 2023; Barret & Dupourqué, 2024), but degrades under model misspecification (Ward et al., 2022). Two lines of work are relevant to us. First, when many observations share a parameter, its posterior given all observations can be decomposed into its posteriors conditioned on individual observations (Geffner et al., 2023; Linhart et al., 2024). Hierarchical NPE (Rodrigues et al., 2021) conditions on an aggregated representation of multiple observations, but it does not scale to tens of thousands observations. Second, when the observation of interest $\tilde{\boldsymbol{y}}$ is known before training, sequential NPE achieves better sampling efficiency using priors informed by $\tilde{\boldsymbol{y}}$ — known as *proposals* — than generic priors, but requires additional correction procedures (Lueckmann et al., 2017; Greenberg et al., 2019; Deistler et al., 2022). Our treatment of the PSF parameters combines the pooling mechanism of the first line, and the correction of (Greenberg et al., 2019).

Fitting distributions to observations. Fitting a prior distribution from a population of observations is empirical Bayes (Robbins, 1992). In variational inference, fitting the prior to the aggregated posterior is the principle of VampPrior (Tomczak & Welling, 2018; Hoffman & Johnson, 2016). Alternating between fitting a recognition model on simulated data and fitting a generative model on real data is the wake-sleep algorithm (Hinton et al., 1995).

3 METHOD OVERVIEW

The main idea behind our method is to train a neural network to estimate the emitter configuration in each frame, while simultaneously adapting the simulator to the acquisition. To this end, we minimize simultaneously two objectives during every training step, each over a distinct set of parameters. The first objective trains a neural network — called the *neural posterior* — on simulated frames to estimate the posterior over emitter configurations given a few consecutive frames. The second objective is used to fit an emitter distribution to the predictions of the neural posterior on real frames. As training progresses, this fitted distribution gradually replaces the initial prior of the simulator, carrying information from the whole acquisition into individual localizations. Figure 2 illustrates both objectives. If the imaging system is unknown, we jointly estimate it in the same way, with a second neural posterior and a distribution over its parameters fitted to its predictions.

3.1 SETUP

The forward model. An SMLM acquisition $\mathbf{I}_{1:T}$ is a collection of T frames of a fixed specimen recorded through a microscope. Given a frame $\mathbf{I}$, we denote by $\{(\boldsymbol{r}_k, \alpha_k)\}_{k=1}^{K}$ its emitter

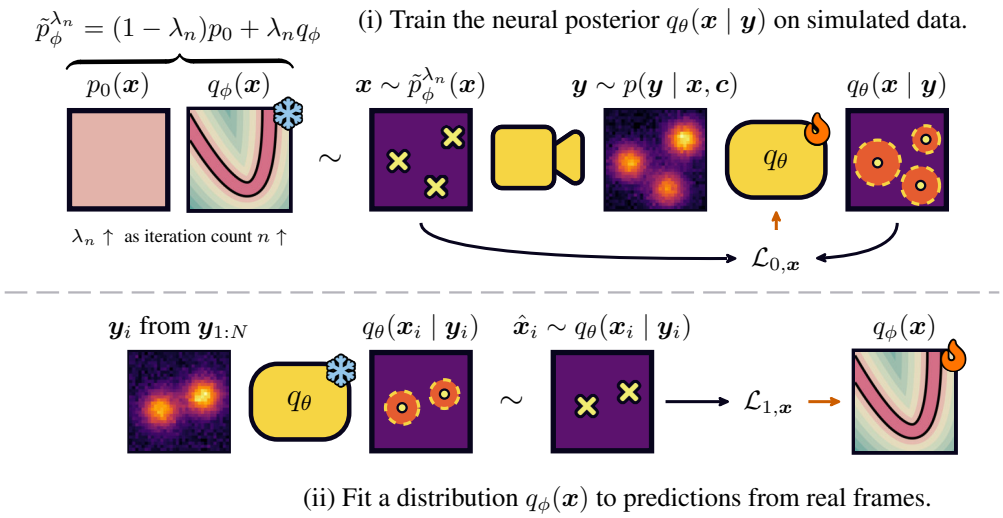

Figure 2: Illustration of a training step, which simultaneously minimizes two functions. (i) Simulated blocks train $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ with $\mathcal{L}_0$. (ii) Its samples on real blocks fit $q_\phi(\boldsymbol{x})$ with $\mathcal{L}_1$. The same pipeline also estimates the PSF and the background in Eq. (1) (omitted here for simplicity). The proposal $\tilde{p}_\phi^{\lambda_n}$ mixes the fitted distribution with a fixed prior, and λ_n progressively increases throughout training. Flames indicate updated parameters, and snowflakes indicate fixed parameters.

configuration, where K is the unknown number of emitters, $\boldsymbol{r}_k \in \Omega \subset \mathbb{R}^3$ is the position of the k-th emitter, and $\alpha_k \in \mathbb{N}$ is its photon count. The frame is generated by

$$\mathbf{I} \sim \mathcal{P}\big(\mathcal{A}(\boldsymbol{x}, \boldsymbol{c})\big), \quad \text{where } \mathcal{A}(\boldsymbol{x}, \boldsymbol{c}) = \sum_{k=1}^{K} \alpha_k \boldsymbol{h}_{\boldsymbol{c}}(\boldsymbol{r}_k) + \boldsymbol{b}_c, \tag{1}$$

$\boldsymbol{h}_{\boldsymbol{c}}(\boldsymbol{r})$ is the PSF image at position $\boldsymbol{r}$, $\boldsymbol{b}_c$ represents the background (assumed static), and $\mathcal{P}$ is the camera model (Appendix B.2). The vector $\boldsymbol{c}$ contains the shared imaging parameters, which includes the PSF parameters (Zernike coefficients, its lateral scale, and others: our complete model is in Appendix B.1) and the background pixels. Importantly, $\boldsymbol{c}$ is shared by all frames of the acquisition. We introduce two distributions over $\boldsymbol{c}$: a prior $p_0(\boldsymbol{c})$, which expresses initial knowledge of the imaging parameters, and a parametric distribution $q_\psi(\boldsymbol{c})$, which is fitted to estimates of $\boldsymbol{c}$ on real frames during training. In practice, $p_0(\boldsymbol{c})$ and $q_\psi(\boldsymbol{c})$ are diagonal Gaussians, see Section 4.

A block model. An emitter may remain active for a few consecutive frames, thus neighboring frames can help detection in a frame of interest. To use this temporal information, our method works on short blocks of F consecutive frames rather than on single frames (we use $F = 3$ in practice, following DECODE (Speiser et al., 2021)). We divide the acquisition into a dataset of N blocks, denoted $\boldsymbol{y}_{1:N}$, where $\boldsymbol{y}_i = (\mathbf{I}_{t_i}, \dots, \mathbf{I}_{t_i+F-1})$ and t_i is the starting frame of the i-th block. Each block has an emitter configuration $\boldsymbol{x}_i = \{(\boldsymbol{r}_{ik}, \boldsymbol{\alpha}_{ik})\}_{k=1}^{K_i}$, with photon count $\boldsymbol{\alpha}_{ik} \in \mathbb{N}^F$.

Standard SMLM assumptions are that emitters belong to a single dye family, that illumination is uniform and constant, and that interactions between emitters are negligible (Helmerich et al. (2022) studied the limits of this approximation). These assumptions motivate modeling the emitter configurations of each block using a common *marked point process*, whose points are emitter positions and whose marks are photon-count sequences. We denote this distribution by $q_\phi(\boldsymbol{x})$: it is the emitter distribution fitted by our second objective (its implementation is given in Section 4). As in DECODE (Speiser et al., 2021), we treat blocks as independent, even though neighboring blocks may contain the same activated emitters.

An adaptive simulator. Generating synthetic data requires a distribution over emitter configurations and shared imaging parameters, called the *proposal* (Greenberg et al., 2019). Given that we adapt our simulator to predictions, errors in these predictions can affect future training examples. For instance, missed emitters may be absent from the fitted distribution, while repeated false detections

may accumulate too much probability mass around their positions. Therefore, we use a *defensive mixture* (Hesterberg, 1995): we draw synthetic emitter configurations $\boldsymbol{x}$ and parameters $\boldsymbol{c}$ from

$$\tilde{p}_\phi^\lambda(\boldsymbol{x}) = (1-\lambda)p_0(\boldsymbol{x}) + \lambda q_\phi(\boldsymbol{x}) \;\text{ and }\; \tilde{p}_\psi^\lambda(\boldsymbol{c}) = (1-\lambda)p_0(\boldsymbol{c}) + \lambda q_\psi(\boldsymbol{c}), \tag{2}$$

where $\lambda \in [0,1]$, and $p_0(\boldsymbol{x})$ and $p_0(\boldsymbol{c})$ are reference priors respectively over emitter configurations $\boldsymbol{x}$ and the imaging system parameters $\boldsymbol{c}$. The weight λ acts as a cursor, providing a continuous transition from the priors to the fitted models. We set λ close to 0 early in training, so that early simulated data comes mostly from the fixed priors, as in other SMLM methods. As training progresses, we gradually increase λ so that the fitted distributions $q_\phi(\boldsymbol{x})$ and $q_\psi(\boldsymbol{c})$ contribute to the simulator.

3.2 Learning the neural posteriors from simulations

In our first objective, the fitted distributions over the emitter configurations $q_\phi(\boldsymbol{x})$ and the shared imaging parameters $q_\psi(\boldsymbol{c})$ are fixed. As in recent SMLM methods (Speiser et al., 2021; Séailles et al., 2026), we train a neural network called the *neural posterior*, whose outputs parameterize a distribution $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ over emitter configurations, to approximate the proposal posterior $\tilde{p}_\phi^\lambda(\boldsymbol{x} \mid \boldsymbol{y})$. We also add a second neural posterior $q_\eta(\boldsymbol{c} \mid \boldsymbol{y})$ over the imaging parameters, which we train to approximate the corresponding proposal posterior $\tilde{p}_\psi^\lambda(\boldsymbol{c} \mid \boldsymbol{y})$. The NPE problem we would ideally like to solve is (Papamakarios & Murray, 2016)

$$\min_{\theta,\eta} \; \mathbb{E}_{\boldsymbol{x}\sim\tilde{p}^\lambda(\boldsymbol{x}),\, \boldsymbol{c}\sim\tilde{p}^\lambda(\boldsymbol{c}),\, \boldsymbol{y}\sim p(\boldsymbol{y}\mid\boldsymbol{x},\boldsymbol{c})}\big[-\log q_\theta(\boldsymbol{x} \mid \boldsymbol{y}) - \log q_\eta(\boldsymbol{c} \mid \boldsymbol{y})\big]. \tag{3}$$

In practice, $q_\eta(\boldsymbol{c} \mid \boldsymbol{y})$ is a Gaussian distribution (see Section 4), thus $\log q_\eta(\boldsymbol{c} \mid \boldsymbol{y})$ has a closed form. By contrast, $\boldsymbol{x}$ is a set of emitters, and evaluating $\log q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ would require summing over all assignments between predicted and target emitters, which is intractable. Instead, for a minibatch of B simulated samples $\boldsymbol{x}_b \sim \tilde{p}_\phi^\lambda(\boldsymbol{x}), \boldsymbol{c}_b \sim \tilde{p}_\psi^\lambda(\boldsymbol{c})$ and $\tilde{\boldsymbol{y}}_b \sim p(\cdot \mid \boldsymbol{x}, \boldsymbol{c})$, we minimize

$$\mathcal{L}_0(\theta, \eta; \boldsymbol{x}_{1:B}, \boldsymbol{c}_{1:B}, \tilde{\boldsymbol{y}}_{1:B}) = \frac{1}{B}\sum_{b=1}^{B} \left(\ell_{0,\boldsymbol{x}}(\theta; \boldsymbol{x}_b, \tilde{\boldsymbol{y}}_b) + \ell_{0,\boldsymbol{c}}(\eta; \boldsymbol{c}_b, \tilde{\boldsymbol{y}}_b)\right), \tag{4}$$

where $\ell_{0,\mathbf{c}}(\eta; \boldsymbol{c}, \boldsymbol{y}) = -\log q_\eta(\boldsymbol{c} \mid \boldsymbol{y})$ and $\ell_{0,\boldsymbol{x}}$ is the loss function between predicted and ground truth emitters of Séailles et al. (2026) (Appendix D.1), which upper-bounds $-\log q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ (it uses a soft assignment based on optimal-transport instead of summing over all assignments).

3.3 Fitting the emitter and the imaging model distributions

In our second objective, the neural posteriors $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ and $q_\eta(\boldsymbol{c} \mid \boldsymbol{y})$ are held fixed, and we fit the distributions $q_\phi(\boldsymbol{x})$ and $q_\psi(\boldsymbol{c})$ from predictions on the acquisition $\boldsymbol{y}_{1:N}$. Starting from the empirical Bayes principle (Robbins, 1992), variational derivations followed by the proposal correction of Greenberg et al. (2019) (Appendix A) yield the fitting problem

$$\min_{\phi,\psi} \; D_{\mathrm{KL}}\big(\bar{q}_\theta(\boldsymbol{x}) \parallel q_\phi(\boldsymbol{x})\big) + \frac{1}{N} D_{\mathrm{KL}}\big(q_\psi(\boldsymbol{c}) \parallel \pi_\eta(\boldsymbol{c})\big), \tag{5}$$

where D_{KL} denotes the Kullback-Leibler (KL) divergence, and the two fixed targets are

$$\bar{q}_\theta(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} q_\theta(\boldsymbol{x} \mid \boldsymbol{y}_i) \quad \text{and} \quad \pi_\eta(\boldsymbol{c}) \propto p_0(\boldsymbol{c}) \prod_{i=1}^{N} \frac{q_\eta(\boldsymbol{c} \mid \boldsymbol{y}_i)}{\tilde{p}_{\bar{\psi}}^\lambda(\boldsymbol{c})}, \tag{6}$$

where $\bar{\psi} = \mathrm{stopgrad}(\psi)$. Equation 5 cleanly separates the fitting of $q_\phi(\boldsymbol{x})$ and $q_\psi(\boldsymbol{c})$ into two divergences, each against its respective target in Eq. (6). The first divergence relates to the emitter distribution $q_\phi(\boldsymbol{x})$, whose information is split over all $\boldsymbol{y}_{1:N}$. The inferred emitter posteriors are accumulated by an average in the *aggregated posterior* $\bar{q}_\theta(\boldsymbol{x})$ (a process known as *linear pooling* (Genest & Zidek, 1986; Hoffman & Johnson, 2016)) to which $q_\phi(\boldsymbol{x})$ is fitted. By contrast, in the second term, every block provides evidence about the shared imaging parameters $\boldsymbol{c}$. Their contributions are multiplied to obtain $\pi_\eta(\boldsymbol{c})$ (a process known as *logarithmic pooling* (Geffner et al., 2023; Linhart et al., 2024)), to which $q_\psi(\boldsymbol{c})$ is fitted. The division by $\tilde{p}_{\bar{\psi}}^\lambda(\boldsymbol{c})$ in $\pi_\eta(\boldsymbol{c})$ comes from the correction of Greenberg et al. (2019), which removes the simulation prior already contained in the posterior. Hence, no gradient should be propagated to ψ through $\tilde{p}_{\bar{\psi}}^\lambda(\boldsymbol{c})$, which is achieved by the stopgrad with $\bar{\psi}$.

Algorithm 1 Training on one acquisition

Require: Blocks $\boldsymbol{y}_{1:N}$, priors $p_0(\boldsymbol{x})$ and $p_0(\boldsymbol{c})$, batch size B, number of steps M, schedule $\{\lambda_n\}_{n=1}^{M}$.
Ensure: Trained localization network $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ and fitted spatial emitter density $q_\phi(\boldsymbol{r})$.

Initialize $\Theta = (\theta, \eta)$ and $\Phi = (\phi, \psi)$.
for $n = 1, \dots, M$ **do**
  $\boldsymbol{y}_{1:B}^{(n)} \leftarrow \mathrm{SampleBatch}(\boldsymbol{y}_{1:N}, B)$ // sample real blocks of frames
  for $b = 1, \dots, B$ **do** // sample synthetic blocks of frames, Eq. (2)
    $\boldsymbol{x}_b \sim (1-\lambda_n)p_0(\boldsymbol{x}) + \lambda_n q_\phi(\boldsymbol{x})$
    $\boldsymbol{c}_b \sim (1-\lambda_n)p_0(\boldsymbol{c}) + \lambda_n q_\psi(\boldsymbol{c})$
    $\tilde{\boldsymbol{y}}_b \sim p(\cdot \mid \boldsymbol{x}_b, \boldsymbol{c}_b)$
  end for
  $g_\Theta \leftarrow \nabla_\Theta \mathcal{L}_0(\theta, \eta; \boldsymbol{x}_{1:B}, \boldsymbol{c}_{1:B}, \tilde{\boldsymbol{y}}_{1:B})$ // neural posteriors step, Eq. (4)
  $g_\Phi \leftarrow \nabla_\Phi \mathcal{L}_1(\phi, \psi; \boldsymbol{y}_{1:B}^{(n)})$ // fitted distributions step, Eq. (7)
  $(\Theta, \Phi) \leftarrow \mathrm{AdamW}\big((\Theta, \Phi), (g_\Theta, g_\Phi)\big)$ // update all parameters
end for
return $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ and $q_\phi(\boldsymbol{r})$.

Minimizing the first divergence in Eq. (5) with respect to ϕ is equivalent to fitting $q_\phi(\boldsymbol{x})$ by maximum likelihood estimation on samples from $\bar{q}_\theta(\boldsymbol{x})$. This motivates using other methods for fitting distributions from samples, such as flow matching (Lipman et al., 2022) and loss functions based on maximum mean discrepancy (MMD) (Gretton et al., 2012). For a mini-batch of B real blocks $\boldsymbol{y}_{1:B}$, $\frac{1}{B}\sum_{b=1}^{B} q_\theta(\boldsymbol{x} \mid \boldsymbol{y}_b)$ is an unbiased estimate of $\bar{q}_\theta(\boldsymbol{x})$, and we denote by $\mathcal{L}_{1,\boldsymbol{x}}(\phi; \boldsymbol{y}_{1:B})$ the loss function of such a method on samples from it, see Section 4. Likewise, we approximate the second divergence in Eq. (5) over a mini-batch by a loss function $\mathcal{L}_{1,\boldsymbol{c}}(\psi; \boldsymbol{y}_{1:B})$. In practice, we minimize

$$\mathcal{L}_1(\phi, \psi; \boldsymbol{y}_{1:B}) = \mathcal{L}_{1,\boldsymbol{x}}(\phi; \boldsymbol{y}_{1:B}) + \mathcal{L}_{1,\boldsymbol{c}}(\psi; \boldsymbol{y}_{1:B}). \tag{7}$$

We describe $\mathcal{L}_{1,\boldsymbol{x}}$ and $\mathcal{L}_{1,\boldsymbol{c}}$ in Section 4 and give their explicit expressions in Appendix D.1.

3.4 The complete algorithm

Algorithm 1 combines our two objectives over minibatches. We update all parameters at the same time with AdamW (Loshchilov & Hutter, 2019); all training details are given in Appendix D.2. This scheme is similar to the wake-sleep algorithm (Hinton et al., 1995): training the neural posteriors with $\mathcal{L}_0$ (Eq. (4)) on simulations corresponds to the sleep phase, while fitting the distributions to predictions on real data with $\mathcal{L}_1$ (Eq. (7)) corresponds to the wake phase. Both objectives depend on all parameters, but we differentiate $\mathcal{L}_0$ only with respect to (θ, η), without differentiating through the simulated data, and $\mathcal{L}_1$ only with respect to (ϕ, ψ), without differentiating through the neural posterior predictions.

Such a feedback loop may reinforce prediction errors. In practice, we have not observed a collapse of the fitted emitter distribution: we attribute this in part to the large number of detections and to the regularization of the spatial density, which we allow to sharpen only as the localization uncertainty decreases (Appendix C.2). The defensive mixture also preserves diversity in the simulations: we cap λ at 0.95, so that the reference priors $p_0(\boldsymbol{x})$ and $p_0(\boldsymbol{c})$ retain a weight of at least 5% (Appendix D.2).

4 Model choices

Neural posteriors. The emitter neural posterior $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ is parameterized by a convolutional neural network that outputs $\bar{K}$ candidate emitters (a hyperparameter set by the architecture of $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$). Each candidate has a detection probability s_j and a diagonal Gaussian distribution over its position and photon-count sequence. The neural posterior $q_\eta(\boldsymbol{c} \mid \boldsymbol{y})$ over the shared imaging parameters uses the same backbone and predicts the mean and diagonal covariance of a Gaussian over $\boldsymbol{c}$. The architectures are described in detail in Appendix C.

Emitter distributions. Following Sage et al. (2019), we represent the emitter configuration $\boldsymbol{x}$ of a block as a sample from a *marked point process*. Such a process is defined by three components: a distribution $p(K)$ over the number of emitters, a spatial density $p(\boldsymbol{r})$ over their positions, and a distri-

Table 1: Evaluation on the high-density 3D astigmatic datasets of the 2016 SMLM Challenge (Sage et al., 2019), with their metrics (Appendix F). Calibration settings and simulator labels are described in the main text (Section 5.1). Results are mean ± standard deviation over three training seeds, and we bold the best mean and results that do not differ significantly from it (Appendix F).

			Simulator			Metrics			
	SNR	Method	Spatial	Photophysics	PSF	Jaccard ↑	$\text{RMSE}_{\text{lat}}(\text{nm})$ ↓	$\text{RMSE}_{\text{ax}}(\text{nm})$ ↓	E_{3D} ↑
External calibration	High	DECODE	Uniform	Reference	Beads	0.858 ± 0.002	28.6 ± 0.2	33.0 ± 0.3	0.710 ± 0.001
		LiteLoc	Uniform	Reference	Beads	0.849 ± 0.001	29.0 ± 0.2	35.5 ± 0.3	0.700 ± 0.001
		SHOT	Uniform	Reference	Beads	$\mathbf{0.869 \pm 0.021}$	27.1 ± 1.7	33.4 ± 1.6	0.722 ± 0.005
		LUNAR	Uniform	Reference	Beads	0.853 ± 0.002	30.0 ± 0.3	34.4 ± 0.2	0.698 ± 0.001
		Ours-fixed	Uniform	Reference	Beads	0.888 ± 0.003	29.0 ± 0.4	33.4 ± 0.2	0.725 ± 0.003
		Ours-structure	*Learnt*	Reference	Beads	$\mathbf{0.912 \pm 0.005}$	$\mathbf{22.5 \pm 0.3}$	$\mathbf{19.3 \pm 2.7}$	$\mathbf{0.796 \pm 0.003}$
	Low	DECODE	Uniform	Reference	Beads	$\mathbf{0.765 \pm 0.005}$	46.7 ± 0.1	$\mathbf{59.3 \pm 0.3}$	0.524 ± 0.002
		LiteLoc	Uniform	Reference	Beads	$\mathbf{0.747 \pm 0.001}$	43.9 ± 0.2	$\mathbf{58.8 \pm 0.4}$	$\mathbf{0.530 \pm 0.002}$
		SHOT	Uniform	Reference	Beads	$\mathbf{0.777 \pm 0.017}$	45.4 ± 1.7	$\mathbf{58.0 \pm 1.4}$	$\mathbf{0.540 \pm 0.001}$
		LUNAR	Uniform	Reference	Beads	0.731 ± 0.002	43.8 ± 0.2	$\mathbf{57.6 \pm 0.6}$	0.521 ± 0.001
		Ours-fixed	Uniform	Reference	Beads	$\mathbf{0.752 \pm 0.001}$	47.0 ± 0.8	$\mathbf{57.6 \pm 0.6}$	0.521 ± 0.005
		Ours-structure	*Learnt*	Reference	Beads	$\mathbf{0.750 \pm 0.010}$	$\mathbf{41.5 \pm 0.7}$	$\mathbf{48.9 \pm 4.7}$	$\mathbf{0.560 \pm 0.013}$
Acquisition-only	High	DECODE	Uniform	Preset	uiPSF	0.686 ± 0.010	50.7 ± 0.8	52.1 ± 0.9	0.473 ± 0.011
		LiteLoc	Uniform	Preset	uiPSF	0.743 ± 0.003	50.1 ± 0.5	54.8 ± 0.7	0.507 ± 0.002
		SHOT	Uniform	Preset	uiPSF	0.765 ± 0.014	45.3 ± 3.4	48.4 ± 1.3	0.551 ± 0.010
		LUNAR	Uniform	Preset	*Learnt*	$\mathbf{0.836 \pm 0.007}$	$\mathbf{36.4 \pm 1.8}$	$\mathbf{35.6 \pm 2.5}$	$\mathbf{0.657 \pm 0.016}$
		Ours	*Learnt*	*Learnt*	*Learnt*	$\mathbf{0.830 \pm 0.012}$	$\mathbf{33.6 \pm 0.8}$	42.4 ± 2.4	$\mathbf{0.653 \pm 0.003}$
	Low	DECODE	Uniform	Preset	uiPSF	0.701 ± 0.001	57.4 ± 0.3	69.7 ± 0.5	0.421 ± 0.001
		LiteLoc	Uniform	Preset	uiPSF	0.666 ± 0.002	52.9 ± 0.3	68.2 ± 0.5	0.424 ± 0.001
		SHOT	Uniform	Preset	uiPSF	$\mathbf{0.714 \pm 0.005}$	52.0 ± 0.4	66.2 ± 0.6	$\mathbf{0.459 \pm 0.002}$
		LUNAR	Uniform	Preset	*Learnt*	0.696 ± 0.006	53.3 ± 0.9	61.3 ± 0.9	0.453 ± 0.003
		Ours	*Learnt*	*Learnt*	*Learnt*	$\mathbf{0.698 \pm 0.036}$	$\mathbf{42.9 \pm 0.9}$	$\mathbf{48.0 \pm 1.3}$	$\mathbf{0.521 \pm 0.026}$

bution $p(\boldsymbol{\alpha})$ over their photon-count sequences. To sample an emitter configuration $\boldsymbol{x}$, we first draw

$$K \sim p(K), \quad \text{then for all } k \in \{1, \dots, K\}, \quad \boldsymbol{r}_k \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{r}) \text{ and } \boldsymbol{\alpha}_k \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{\alpha}). \tag{8}$$

For the fitted point process $q_\phi(\boldsymbol{x})$, we use a binomial distribution over $\bar{K}$ trials and a learnable success probability for $q_\phi(K)$, where $\bar{K}$ is the maximal number of emitters per block, a continuous normalizing flow parameterized by a neural velocity field for $q_\phi(\boldsymbol{r})$, and a learnable photon-count model for $q_\phi(\boldsymbol{\alpha})$ (details in Appendix C). Their loss function $\mathcal{L}_{1,\boldsymbol{x}}$ combines three loss functions: conditional flow matching for $q_\phi(\boldsymbol{r})$, MMD for $q_\phi(\boldsymbol{\alpha})$ and a KL divergence between Poisson approximations of the inferred and fitted count distributions for $q_\phi(K)$. Their analytical forms are given in Appendix D.1. For the reference prior $p_0(\boldsymbol{x})$, we set $p_0(K) = \mathcal{U}(\{0, \dots, \bar{K}\})$, $p_0(\boldsymbol{r}) = \mathcal{U}(\Omega)$, and $p_0(\boldsymbol{\alpha})$ to a simple photophysics model (Appendix C).

Imaging model. We use the PSF model of Liu et al. (2026), described in Appendix B.1, which mainly relies on a Zernike expansion of the pupil phase. The background has the same shape as the frames, and each of its pixels is included in $\boldsymbol{c}$ and learnt jointly with the PSF parameters. We use diagonal Gaussians for both the reference prior $p_0(\boldsymbol{c})$ and the learnt distribution $q_\psi(\boldsymbol{c})$. The loss function $\mathcal{L}_{1,\boldsymbol{c}}$ can be evaluated in closed form with the Gaussian approximation to the proposal described in Appendix D.1. The camera response parameters, which define $\mathcal{P}$ in Eq. (1), are supplied by the user, typically from manufacturers' datasheets (Appendix B.2). By default, lateral coordinates are expressed in pixels and axial coordinates on a normalized scale. Converting positions to nanometers requires the pixel size and a calibration of the axial scale.

5 Experiments

5.1 Synthetic benchmarks

We compare our method with DECODE (Speiser et al., 2021), LiteLoc (Fei et al., 2025), SHOT (Séailles et al., 2026) and LUNAR (Fu et al., 2026) on the publicly available, low and high SNR, high-density (2.0 activations/μm^2/frame), 3D astigmatic datasets of the 2016 SMLM

Table 2: Evaluation on experimental datasets (Section 5.2). # Dets. is the number of localizations (in millions), FRC is the Fourier ring correlation resolution estimate, and RSP the resolution-scaled Pearson coefficient (Appendix F). Best metric values for each dataset are in bold.

	Tubulin			**NPC-Nup107**			**NPC-Nup96**		
Method	# Dets. ($\times 10^6$)	FRC (nm) ↓	RSP ↑	# Dets. ($\times 10^6$)	FRC (nm) ↓	RSP ↑	# Dets. ($\times 10^6$)	FRC (nm) ↓	RSP ↑
DECODE	22.38	23.1	0.722	1.50	14.2	0.856	1.93	25.1	0.763
LiteLoc	10.04	23.0	0.720	0.95	16.2	0.842	0.48	23.6	0.702
SHOT	14.24	21.6	0.783	1.18	14.1	**0.901**	0.56	22.0	0.776
LUNAR	15.05	**17.9**	0.791	1.26	12.1	0.826	0.90	**19.5**	0.705
Ours	11.02	18.7	**0.799**	1.08	**11.8**	0.779	0.41	21.0	**0.861**

Challenge (Sage et al., 2019). We study two calibration settings, detailed below: *External calibration*, to isolate the effect of learning the specimen structure, and *Acquisition-only*, to evaluate performance using only the acquisition and camera parameters. Code to reproduce these experiments is in the supplementary material.

Metrics. We report the standard metrics used by Sage et al. (2019), detailed in Appendix F. In a few words, we match predicted localizations to ground-truth emitters within each frame, and count matched pairs below the challenge's distance threshold as true positives. We then compute the Jaccard index, which measures detection performance, and the lateral and axial root mean square errors (RMSE_{lat} and RMSE_{ax}) over matched emitters, which measure localization performance. We also report the 3D efficiency (E_{3D}), which combines these metrics into a single score.

Results over these metrics are reported in Table 1. The columns *Spatial*, *Photophysics*, and *PSF* respectively describe the spatial density, the photophysics of emitters, and the source of the PSF for each simulator. We explain their entries below, giving the corresponding labels in parentheses. Only components marked *Learnt* are updated during training.

External calibration. This setting compares all methods under a common reference calibration. We map each method simulator to the public values of the challenge (*Reference*), and set the PSF to a common model estimated from beads (*Beads*) with SMAP (Ries, 2020). *Ours-fixed* follows the same setup, and is almost equivalent to SHOT (Séailles et al., 2026) without its recursive architecture and with localization uncertainties estimates for each detection. While these methods use uniform emitter positions and a preset number of emitters, *Ours-structure* differs only by learning the spatial density $q_\phi(\boldsymbol{r})$ and $q_\phi(K)$ from the acquisition. It reduces both lateral and axial localization errors at both SNR levels, and improves detection at high SNR.

Acquisition-only calibration. In practice, a separate PSF calibration and ground-truth photophysics values may be unavailable. Therefore, this setting only uses the acquisition and available camera parameters. The emitter photophysics and density are set according to each method's documentation (*Preset*), using either their dedicated toolbox or the calibration plugin of SMAP (Ries, 2020). For DECODE, LiteLoc and SHOT, a PSF model is pre-estimated from the image stack using *uiPSF* (Liu et al., 2024). Our full method jointly learns the emitter-count distribution $q_\phi(K)$, spatial density $q_\phi(\boldsymbol{r})$, photophysics $q_\phi(\boldsymbol{\alpha})$, and distribution over the imaging system $q_\psi(\boldsymbol{c})$. Its results are either the best or statistically on par with the best for Jaccard, RMSE_{lat}, and E_{3D} at both SNR levels.

5.2 Real acquisitions

We compare the same methods on three open SMLM datasets: Tubulin and NPC-Nup107 from Li et al. (2018); Sage et al. (2019), and NPC-Nup96 from Fei et al. (2025). We applied the calibration process of the acquisition-only setting described in Section 5.1. Table 2 evaluates reconstructions from the localization network, while Figure 3 shows the fitted spatial density.

Quantitative results. In the absence of ground-truth emitter positions, we use Fourier ring correlation (FRC) (Banterle et al., 2013) and the resolution-scaled Pearson coefficient (RSP) (Culley et al., 2018), detailed in Appendix F. The former estimates lateral resolution from images reconstructed using localizations from two disjoint subsets of the acquisition. The latter measures agreement with a lower-resolution reference after blurring and downsampling the reconstruction.

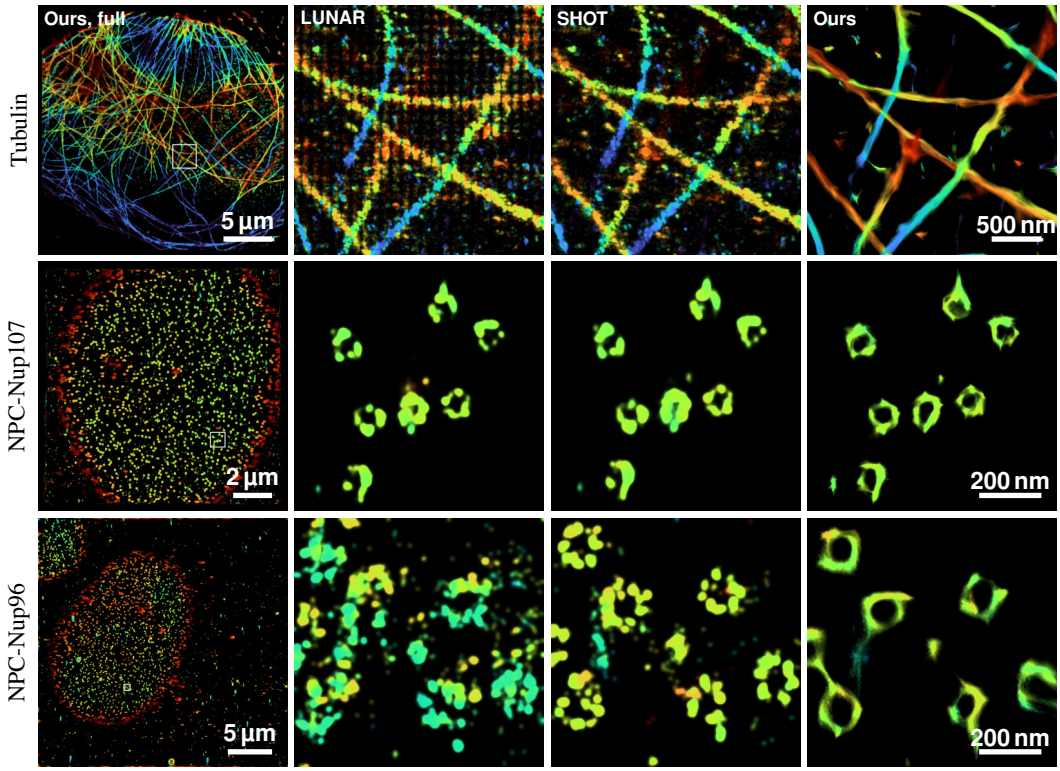

Figure 3: Reconstructions of the Tubulin, NPC-Nup107 and NPC-Nup96 acquisitions. For each dataset, we show our reconstruction over the full field of view, and crops reconstructed by LUNAR, SHOT and our method. While LUNAR and SHOT's reconstructions are point clouds, our method shows the fitted spatial density $q_\phi(\boldsymbol{r})$. Color encodes depth.

We evaluate our method with the reconstruction from the localizations of $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$, so that all methods are compared on an equal footing using point clouds. Results are reported in Table 2. Our method has the highest RSP on Tubulin and NPC-Nup96, and the second-lowest FRC on both, behind LUNAR. On NPC-Nup107, it has the best FRC, but also the worst RSP, which illustrates the limits of assessing the reconstruction with either metric alone.

Qualitative results. Figure 3 shows reconstructions on all three datasets. Our method uses the fitted spatial density $q_\phi(\boldsymbol{r})$. On Tubulin, ours shows smoother filaments and less signal between them than LUNAR and SHOT. LUNAR also shows a grid artefact, which does not appear with SHOT and ours. On NPC-Nup107 and NPC-Nup96, our reconstruction shows clearer ring-like NPC structures.

6 Conclusion and discussion

In this paper, we propose an SMLM method that learns the specimen structure from the acquisition and uses it to improve emitter localization. We fit an emitter distribution to predictions on real frames, and use it to adapt the simulator during training, while jointly estimating the PSF, the background and the emitter photophysics. The fitted spatial density also provides a direct, continuous, super-resolved representation of the specimen. With external calibration, learning the spatial density and emitter-count distribution reduces localization errors. On real acquisitions, our fitted spatial density reveals fine specimen structure, with smooth filaments and clearly defined nuclear pore rings, using only the acquisition and camera parameters.

However, our two training objectives do not minimize a single joint objective, and errors in the inferred localizations can propagate to the fitted emitter distribution, hence deriving a unified objective is a natural next step. Evaluation on real acquisitions also remains difficult without ground-truth localizations, calling for methods that better assess the quality of recovered structures.

AI USE STATEMENT

This project was assisted by generative AI tools, mainly Claude (Anthropic) and ChatGPT (OpenAI).

We developed the research questions, method and experimental design based on our analysis of the limitations of existing SMLM methods. Claude helped us search for related work, connect our ideas with other fields (variational inference, empirical Bayes, proposal correction in simulation-based inference), and derive the maths of Appendix A. We checked all references manually, and verified all derivations by hand.

Regarding the codebase, Claude wrote snippets for functions and methods we asked for, helped with technical implementation details (parallelization, speed-ups), and helped us reproduce the results of previous methods from their authors' released code. We built and kept full control of the codebase, as all generated code was reviewed and tested.

We wrote the manuscript and used Claude and ChatGPT to proofread it and to improve overall clarity. We also used its feedback to revise the manuscript, similarly to a peer-review process where the LLM played the role of a reviewer. We used the ICLR pre-submission review tools to obtain additional feedback on the manuscript.

Overall, we believe this project could have been done without modern AI tools, but they definitely sped up the research process. We take full responsibility for the manuscript, code and results.

REPRODUCIBILITY STATEMENT

We provide the source code and model implementation in the supplementary material. It includes training configurations and our tools for exporting predictions and evaluating reconstructions. These materials enable reproduction of the experiments reported in Section 5. All datasets are publicly available, and remain subject to their original licences and terms of use set by their original authors. Further implementation, training and evaluation details are provided in Appendices.

REFERENCES

Ernst Abbe. Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung. *Archiv für mikroskopische Anatomie*, 9(1):413–468, 1873.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint*, 1607.06450, 2016.

Hazen Babcock, Yaron M Sigal, and Xiaowei Zhuang. A high-density 3D localization algorithm for stochastic optical reconstruction microscopy. *Optical Nanoscopy*, 1:1–10, 2012.

Hazen P Babcock and Xiaowei Zhuang. Analyzing single molecule localization microscopy data using cubic splines. *Scientific Reports*, 7(552), 2017.

Francisco Balzarotti, Yvan Eilers, Klaus C Gwosch, Arvid H Gynnå, Volker Westphal, Fernando D Stefani, Johan Elf, and Stefan W Hell. Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science*, 355:606–612, 2017.

Niccolo Banterle, Khanh Huy Bui, Edward A Lemke, and Martin Beck. Fourier ring correlation as a resolution criterion for super-resolution microscopy. *Journal of Structural Biology*, 183:363–367, 2013.

Didier Barret and Simon Dupourqué. Simulation-Based Inference with Neural Posterior Estimation applied to X-ray spectral fitting: Demonstration of working principles down to the Poisson regime. *Astronomy & Astrophysics*, 686:A133, 2024.

Eric Betzig, George H Patterson, Rachid Sougrat, O Wolf Lindwasser, Scott Olenych, Juan S Bonifacino, Michael W Davidson, Jennifer Lippincott-Schwartz, and Harald F Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 313:1642–1645, 2006.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Nicholas Boyd, Eric Jonas, Hazen Babcock, and Benjamin Recht. DeepLoco: Fast 3D Localization Microscopy Using Neural Networks. *BioRxiv preprint*, pp. 267096, 2018.

Susan Cox, Edward Rosten, James Monypenny, Tijana Jovanovic-Talisman, Dylan T Burnette, Jennifer Lippincott-Schwartz, Gareth E Jones, and Rainer Heintzmann. Bayesian localization microscopy reveals nanoscale podosome dynamics. *Nature Methods*, 9(2):195–200, 2012.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Siân Culley, David Albrecht, Caron Jacobs, Pedro Matos Pereira, Christophe Leterrier, Jason Mercer, and Ricardo Henriques. Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nature Methods*, 15:263–266, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 26:2292–2300, 2013.

Maximilian Dax, Stephen R Green, Jonathan Gair, Jakob H Macke, Alessandra Buonanno, and Bernhard Schölkopf. Real-Time Gravitational Wave Science with Neural Posterior Estimation. *Physical review letters*, 127(24):241103, 2021.

Michael Deistler, Pedro J Goncalves, and Jakob H Macke. Truncated proposals for scalable and hassle-free simulation-based inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:23135–23149, 2022.

Thomas Dertinger, Ryan Colyer, Gopal Iyer, Shimon Weiss, and Jörg Enderlein. Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). *Proceedings of the National Academy of Sciences (PNAS)*, 106(52):22287–22292, 2009.

Yue Fei, Shuang Fu, Wei Shi, Ke Fang, Ruixiong Wang, Tianlun Zhang, and Yiming Li. Scalable and lightweight deep learning for efficient high accuracy single-molecule localization microscopy. *Nature Communications*, 16(7217), 2025.

Boris Ferdman, Elias Nehme, Lucien E Weiss, Reut Orange, Onit Alalouf, and Yoav Shechtman. VIPR: vectorial implementation of phase retrieval for fast and accurate microscopic pixel-wise pupil estimation. *Optics Express*, 28(7):10179–10198, 2020.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. POT Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Shuang Fu, Wei Shi, Tingdan Luo, Yingchuan He, Lulu Zhou, Jie Yang, Zhichao Yang, Jiadong Liu, Xiaotian Liu, Zhiyong Guo, et al. Field-dependent deep learning enables high-throughput whole-cell 3D super-resolution imaging. *Nature Methods*, 20(3):459–468, 2023.

Shuang Fu, Wei Shi, Eugene A Katrukha, Xi Chen, Yue Fei, Ke Fang, Ruixiong Wang, Tianlun Zhang, Donghan Ma, and Yiming Li. Aberration-aware 3D localization microscopy via self-supervised neural-physics learning. *Nature Communications*, 17:6504, 2026.

Tomas Geffner, George Papamakarios, and Andriy Mnih. Compositional Score Modeling for Simulation-Based Inference. In *International Conference on Machine Learning (ICML)*, pp. 11098–11116, 2023.

Christian Genest and James V Zidek. Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science*, 1:114–135, 1986.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic Posterior Transformation for Likelihood-Free Inference. In *International Conference on Machine Learning (ICML)*, pp. 2404–2414, 2019.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Mats GL Gustafsson. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *Journal of Microscopy*, 198:82–87, 2000.

Nils Gustafsson, Siân Culley, George Ashdown, Dylan M Owen, Pedro Matos Pereira, and Ricardo Henriques. Fast live-cell conventional fluorophore nanoscopy with ImageJ through super-resolution radial fluctuations. *Nature Communications*, 7(1):12471, 2016.

Ernst Hairer, Gerhard Wanner, and Syvert P Nørsett. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, 1993.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Stefan W Hell and Jan Wichmann. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics Letters*, 19(11):780–782, 1994.

Dominic A Helmerich, Gerti Beliu, Danush Taban, Mara Meub, Marcel Streit, Alexander Kuhlemann, Sören Doose, and Markus Sauer. Sub-10 nm fluorescence imaging. *BioRxiv preprint*, pp. 479592, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *arXiv preprint*, 1606.08415, 2016.

Samuel T Hess, Thanu PK Girirajan, and Michael D Mason. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophysical Journal*, 91(11):4258–4272, 2006.

Tim Hesterberg. Weighted Average Importance Sampling and Defensive Mixture Distributions. *Technometrics*, 37:185–194, 1995.

Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. The "Wake-Sleep" Algorithm for Unsupervised Neural Networks. *Science*, 268:1158–1161, 1995.

Michael Hirsch, Richard J Wareham, Marisa L Martin-Fernandez, Michael P Hobson, and Daniel J Rolfe. A Stochastic Model for Electron Multiplication Charge-Coupled Devices – From Theory to Practice. *PloS one*, 8(1):e53671, 2013.

Matthew D Hoffman and Matthew J Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Romain F Laine, Hannah S Heil, Simao Coelho, Jonathon Nixon-Abell, Angélique Jimenez, Theresa Wiesner, Damián Martínez, Tommaso Galgani, Louise Régnier, Aki Stubb, et al. High-fidelity 3D live-cell nanoscopy through data-driven enhanced super-resolution radial fluctuation. *Nature Methods*, 20:1949–1956, 2023.

Mickaël Lelek, Melina T Gyparaki, Gerti Beliu, Florian Schueder, Juliette Griffié, Suliana Manley, Ralf Jungmann, Markus Sauer, Melike Lakadamyali, and Christophe Zimmer. Single-molecule localization microscopy. *Nature Reviews Methods Primers*, 1(39), 2021.

Yiming Li, Markus Mund, Philipp Hoess, Joran Deschamps, Ulf Matti, Bianca Nijmeijer, Vilma Jimenez Sabinina, Jan Ellenberg, Ingmar Schoen, and Jonas Ries. Real-time 3D single-molecule localization using experimental point spread functions. *Nature Methods*, 15: 367–369, 2018.

Julia Linhart, Gabriel Victorino Cardoso, Alexandre Gramfort, Sylvain Le Corff, and Pedro LC Rodrigues. Diffusion posterior sampling for simulation-based inference in tall data settings. *arXiv preprint*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. *arXiv preprint*, 2022.

Sheng Liu, Emil B Kromann, Wesley D Krueger, Joerg Bewersdorf, and Keith A Lidke. Three dimensional single molecule localization using a phase retrieved pupil function. *Optics express*, 21(24):29462–29487, 2013.

Sheng Liu, Jianwei Chen, Jonas Hellgoth, Lucas-Raphael Müller, Boris Ferdman, Christian Karras, Dafei Xiao, Keith A Lidke, Rainer Heintzmann, Yoav Shechtman, et al. Universal inverse modeling of point spread functions for SMLM localization and microscope characterization. *Nature Methods*, 21(6):1082–1093, 2024.

Yan Liu, Vasiliki Stergiopoulou, Jonathan Chuah, Eric Bezzam, Gert-Jan Both, Michael Unser, Daniel Sage, and Jonathan Dong. Revisiting PSF models: Unifying framework and high-performance implementation. *Journal of Microscopy*, 301(3):362–374, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.

Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Claus B Müller and Jörg Enderlein. Image Scanning Microscopy. *Physical review letters*, 104(19): 198101, 2010.

Elias Nehme, Lucien E Weiss, Tomer Michaeli, and Yoav Shechtman. Deep-STORM: super-resolution single-molecule microscopy by deep learning. *Optica*, 5(4):458–464, 2018.

Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning. *Nature Methods*, 17(7):734–740, 2020.

Raimund J Ober, Sripad Ram, and E Sally Ward. Localization Accuracy in Single-Molecule Microscopy. *Biophysical journal*, 86(2):1185–1200, 2004.

Wei Ouyang, Andrey Aristov, Mickaël Lelek, Xian Hao, and Christophe Zimmer. Deep learning massively accelerates super-resolution localization microscopy. *Nature Biotechnology*, 36: 460–468, 2018.

Martin Ovesnỳ, Pavel Křížek, Josef Borkovec, Zdeněk Švindrych, and Guy M Hagen. ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics*, 30(16):2389–2390, 2014.

George Papamakarios and Iain Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016.

Gabriel Peyré, Marco Cuturi, et al. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. *Advances in Neural Information Processing Systems (NeurIPS)*, 20, 2007.

Jonas Ries. SMAP: a modular super-resolution microscopy analysis platform for SMLM data. *Nature Methods*, 17:870–872, 2020.

Herbert E Robbins. An Empirical Bayes Approach to Statistics. In *Breakthroughs in Statistics*, pp. 388–394. 1992.

Pedro Rodrigues, Thomas Moreau, Gilles Louppe, and Alexandre Gramfort. HNPE: Leveraging Global Parameters for Neural Posterior Estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:13432–13443, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Michael J Rust, Mark Bates, and Xiaowei Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods*, 3:793–796, 2006.

Daniel Sage, Thanh-An Pham, Hazen Babcock, Tomas Lukes, Thomas Pengo, Jerry Chao, Ramraj Velmurugan, Alex Herbert, Anurag Agrawal, Silvia Colabrese, et al. Super-resolution fight club: assessment of 2D and 3D single-molecule localization microscopy software. *Nature Methods*, 16:387–395, 2019.

Lothar Schermelleh, Alexia Ferrand, Thomas Huser, Christian Eggeling, Markus Sauer, Oliver Biehlmaier, and Gregor PC Drummen. Super-resolution microscopy demystified. *Nature cell biology*, 21:72–84, 2019.

Artur Speiser, Lucas-Raphael Müller, Philipp Hoess, Ulf Matti, Christopher J Obara, Wesley R Legant, Anna Kreshuk, Jakob H Macke, Jonas Ries, and Srinivas C Turaga. Deep learning enables fast and dense single-molecule localization with high accuracy. *Nature Methods*, 18:1082–1090, 2021.

Romain Séailles, Jean-Baptiste Masson, Jean Ponce, and Julien Mairal. Optimal transport unlocks end-to-end learning for single-molecule localization. *International Conference on Learning Representations (ICLR)*, 2026.

Jakub M Tomczak and Max Welling. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1214–1223, 2018.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport. *arXiv preprint*, 2023.

Malavika Vasist, François Rozet, Olivier Absil, Paul Mollière, Evert Nasedkin, and Gilles Louppe. Neural posterior estimation for exoplanetary atmospheric retrieval. *Astronomy & Astrophysics*, 672:A147, 2023.

Daniel Ward, Patrick Cannon, Mark Beaumont, Matteo Fasiolo, and Sebastian Schmon. Robust Neural Posterior Estimation and Statistical Model Criticism. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 2022.

Fan Xu, Donghan Ma, Kathryn P MacPherson, Sheng Liu, Ye Bu, Yu Wang, Yu Tang, Cheng Bi, Tim Kwok, Alexander A Chubykin, et al. Three-dimensional nanoscopy of whole cells and tissues with in situ point spread function retrieval. *Nature Methods*, 17(5):531–540, 2020.

A Derivations

A.1 The second objective of Eq. (7)

The evidence bound. We seek to maximize the evidence with respect to ϕ, which is the empirical-Bayes criterion (Robbins, 1992), where $\boldsymbol{y}_{1:N}$ is a dataset of N i.i.d. samples given the shared value $\boldsymbol{c}$:

$$\max_{\phi} \; p_\phi(\boldsymbol{y}_{1:N}). \tag{9}$$

We have the following product model

$$p_\phi(\boldsymbol{y}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{c}) = p_0(\boldsymbol{c}) \prod_{i=1}^{N} p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{c})\, q_\phi(\boldsymbol{x}_i), \tag{10}$$

whose evidence is

$$p_\phi(\boldsymbol{y}_{1:N}) = \int p_0(\boldsymbol{c}) \prod_{i=1}^{N} \left[\int p(\boldsymbol{y}_i \mid \boldsymbol{x}, \boldsymbol{c}) q_\phi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right] \mathrm{d}\boldsymbol{c}. \tag{11}$$

We approximate the acquisition posterior by

$$q(\boldsymbol{x}_{1:N}, \boldsymbol{c}) = q_\psi(\boldsymbol{c}) \prod_{i=1}^{N} q_\theta(\boldsymbol{x}_i \mid \boldsymbol{y}_i). \tag{12}$$

This mean-field family (Blei et al., 2017) separates the per-block emitter configurations from the remaining uncertainty in the common PSF. It can be useful when evidence from many blocks concentrates the PSF posterior.

For fixed emitter factors, define

$$g_\theta(\boldsymbol{c}) = p_0(\boldsymbol{c}) \exp\left\{ \sum_{i=1}^{N} \mathbb{E}_{q_\theta(\boldsymbol{x}|\boldsymbol{y}_i)}[\log p(\boldsymbol{y}_i \mid \boldsymbol{x}, \boldsymbol{c})] \right\},$$
$$Z_\theta = \int g_\theta(\boldsymbol{c}) \, \mathrm{d}\boldsymbol{c}, \qquad \pi_\theta(\boldsymbol{c}) = g_\theta(\boldsymbol{c})/Z_\theta. \tag{13}$$

The density π_θ is the unrestricted optimal PSF factor for these fixed emitter factors.

Lemma 1. *For any ϕ, ψ and θ:*

$$-\log p_\phi(\boldsymbol{y}_{1:N}) \le N \, D_{\mathrm{KL}}(\bar{q}_\theta \,\|\, q_\phi) + D_{\mathrm{KL}}(q_\psi \,\|\, \pi_\theta) + C(\theta), \tag{14}$$

where $\bar{q}_\theta = N^{-1} \sum_i q_\theta(\boldsymbol{x} \mid \boldsymbol{y}_i)$ and

$$C(\theta) = N\mathcal{H}(\bar{q}_\theta) - \sum_{i=1}^{N} H(q_\theta(\boldsymbol{x} \mid \boldsymbol{y}_i)) - \log Z_\theta. \tag{15}$$

Here, $\mathcal{H}$ is the differential entropy operator, and C is a function independent of ϕ and ψ.

Proof. The variational identity gives

$$\begin{aligned} \mathcal{F}(q, \phi) &:= \mathbb{E}_q[\log q - \log p_\phi(\boldsymbol{y}_{1:N}, \boldsymbol{x}_{1:N}, \boldsymbol{c})] \\ &= -\log p_\phi(\boldsymbol{y}_{1:N}) + D_{\mathrm{KL}}(q \,\|\, p_\phi(\boldsymbol{x}_{1:N}, \boldsymbol{c} \mid \boldsymbol{y}_{1:N})) \,. \end{aligned} \tag{16}$$

The last divergence is nonnegative. It remains to rearrange $\mathcal{F}$ under Eq. (12).

The emitter-prior terms are

$$\begin{aligned} \sum_{i=1}^{N} \mathbb{E}_{q_\theta(\boldsymbol{x}|\boldsymbol{y}_i)}[-\log p_\phi(\boldsymbol{x})] &= N\mathbb{E}_{\bar{q}_\theta}[-\log q_\phi(\boldsymbol{x})] \\ &= N D_{\mathrm{KL}}(\bar{q}_\theta \,\|\, q_\phi) + N H(\bar{q}_\theta). \end{aligned} \tag{17}$$

This equality produces the aggregated posterior: the common emitter prior is fitted to an arithmetic average of the block distributions (Hoffman & Johnson, 2016).

The PSF prior, PSF entropy, and expected log-likelihoods combine as

$$\mathbb{E}_{q_\psi}[\log q_\psi(\boldsymbol{c}) - \log g_\theta(\boldsymbol{c})] = D_{\mathrm{KL}}(q_\psi \,\|\, \pi_\theta) - \log Z_\theta. \tag{18}$$

The remaining terms are $-\sum_i H(q_\theta(\boldsymbol{x} \mid \boldsymbol{y}_i))$. Substitution into Eq. (16) gives the stated bound. □

The two aggregation mechanisms therefore follow from the two scopes of the latent variables. Different blocks contain different emitter configurations, whose distributions enter the same prior fit through a sum. The PSF is one shared quantity, so the expected log-likelihood contributions add before exponentiation in Eq. (13). These are the arithmetic and logarithmic forms of pooling (Genest & Zidek, 1986).

Proposal-corrected PSF evidence. The variational target π_θ requires expected image log-likelihoods for every block. The PSF head instead gives access to each block's marginal evidence under the simulation model. To make this relation explicit, define

$$\ell_{\tilde{p}^\lambda}(\boldsymbol{y}_i \mid \boldsymbol{c}) = \int p(\boldsymbol{y}_i \mid \boldsymbol{x}, \boldsymbol{c}) \tilde{p}^\lambda(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{19}$$

If the PSF head is exact for the current proposal, Bayes' rule gives

$$q_\eta(\boldsymbol{c} \mid \boldsymbol{y}_i) = \frac{\tilde{p}^\lambda(\boldsymbol{c}) \ell_{\tilde{p}^\lambda}(\boldsymbol{y}_i \mid \boldsymbol{c})}{\tilde{p}^\lambda(\boldsymbol{y}_i)}. \tag{20}$$

Dividing by the simulation prior therefore removes its contribution from the block posterior (Greenberg et al., 2019). Multiplying the resulting evidence contributions and applying the acquisition prior once gives

$$\pi_\eta(\boldsymbol{c}) \propto p_0(\boldsymbol{c}) \prod_{i=1}^{N} \frac{q_\eta(\boldsymbol{c} \mid \boldsymbol{y}_i)}{\tilde{p}^\lambda(\boldsymbol{c})}. \tag{21}$$

For exact heads, this is the PSF posterior of the working product model with emitter prior $\tilde{p}^\lambda(\boldsymbol{x})$ (Geffner et al., 2023; Linhart et al., 2024). The notation leaves its dependence on the fixed proposal implicit. Logarithmic pooling here means multiplication of corrected evidence with one prior; it does not mean an uncorrected geometric average of the block posteriors.

B The image formation model

B.1 Point spread function

We use a scalar Fourier model of the pupil, following the framework of Liu et al. (2026). As in Fu et al. (2026), Zernike coefficients describe the pupil phase. The implemented parameter vector $\boldsymbol{c}$ also contains an apodization parameter and an offset controlling lateral scale.

Pupil and propagation. Let $\boldsymbol{u} = (u_1, u_2)$ be a point on the pupil grid, $\rho = \|\boldsymbol{u}\|$, and $s_z(\boldsymbol{u}) = \sqrt{\max(1-\rho^2, 10^{-3})}$. The complex pupil is

$$P_{\boldsymbol{c}}(\boldsymbol{u}) = A_\epsilon(\rho)\, s_z(\boldsymbol{u})^{\beta(a)} \exp\left(\mathrm{i} \sum_{m \in \mathcal{J}} c_m Z_m(\boldsymbol{u}) \right), \qquad \beta(a) = \tfrac{1}{2} + \tanh a, \tag{22}$$

where Z_m are the sampled Zernike modes. The exponent satisfies $-1/2 < \beta(a) < 3/2$. The aperture is tapered over its outer $\epsilon = 0.1$ fraction to reduce ringing:

$$A_\epsilon(\rho) = \begin{cases} 1, & 0 \le \rho \le 1-\epsilon, \\ \frac{1}{2}\left[1 + \cos\left(\pi \frac{\rho - 1 + \epsilon}{\epsilon}\right)\right], & 1-\epsilon < \rho \le 1, \\ 0, & \rho > 1. \end{cases} \tag{23}$$

For a normalized axial coordinate $z \in [-1, 1]$, the raw intensity plane is

$$I_{\boldsymbol{c}}(\cdot, z) = \left| \mathcal{T}_{\zeta_{\boldsymbol{c}}}[P_{\boldsymbol{c}}(\boldsymbol{u}) \exp(2\pi \mathrm{i}\, \sigma(\boldsymbol{c}) z\, s_z(\boldsymbol{u}))] \right|^2, \qquad \zeta_{\boldsymbol{c}} = \exp(\ell_{\mathrm{ref}} + \ell). \tag{24}$$

Here $\mathcal{T}_\zeta$ denotes the sampled inverse Fourier transform used by the renderer, with its lateral sampling controlled by ζ. The reference scale ℓ_{ref} is set by a quick calibration pass, see Appendix E. The dimensionless constant and transform convention belong to the renderer's coordinate system.

The axial scale $\sigma(\boldsymbol{c})$ is set by a convention on the on-axis intensity. Define

$$J_{\boldsymbol{c}}(s) = \left| \sum_{\boldsymbol{u}} P_{\boldsymbol{c}}(\boldsymbol{u}) e^{2\pi \mathrm{i} s s_z(\boldsymbol{u})} \right|^2, \qquad R_{\boldsymbol{c}}(s) = \frac{\sqrt{J_{\boldsymbol{c}}(s) J_{\boldsymbol{c}}(-s)}}{J_{\boldsymbol{c}}(0)}, \tag{25}$$

We set $\sigma(\boldsymbol{c})$ to the first crossing of $R_{\boldsymbol{c}}(s) = 0.1$ on $[0, 6]$, or to 6 if no crossing is found. When a crossing is found, the geometric mean of the on-axis intensities at the two axial boundaries, $z = \pm 1$, is 10% of the on-axis intensity at $z = 0$. This specifies a normalized defocus range.

The renderer computes an axial stack, linearly interpolates between adjacent planes, and applies a bilinear subpixel shift in the image plane. The stack is normalized by the total intensity of the plane at $z = 0$.

Modes and reference distribution. For a maximum radial order n_Z, the implementation retains the following Zernike modes

$$\mathcal{J} = \{0, \ldots, (n_Z+1)(n_Z+2)/2 - 1\} \setminus \{0, 1, 2, 4\}. \tag{26}$$

Excluding $\{0, 1, 2, 4\}$ removes piston, the two tilt modes, and the defocus mode in the convention used by the code. Piston has no effect on intensity, and tilt only shifts the PSF laterally, and we therefore exclude these modes by convention. Excluding defocus reduces coupling between pupil phase and the axial coordinate.

The latent $\boldsymbol{c}$. The parameters vector $\boldsymbol{c}$ is defined as

$$\boldsymbol{c} = (\{c_m\}_{m\in\mathcal{J}}, a, \ell, \boldsymbol{\beta}), \tag{27}$$

where $\boldsymbol{\beta}$ models the background, see Appendix E. These parameters are shared by all frames of the acquisition.

For $p_0(\boldsymbol{c})$, we use a diagonal Gaussian distribution:

$$p_0(\boldsymbol{c}) = \mathcal{N}(\boldsymbol{c}; \boldsymbol{\mu}_0, s_0^2 \boldsymbol{I}), \tag{28}$$

whose mean is zero except for the vertical-astigmatism coefficient, indexed as mode 5, which has mean -1. The default standard deviation is $s_0 = 0.3$ for all entries of $\boldsymbol{c}$, including a and ℓ. This prior favors one astigmatism sign, which is also a convention.

B.2 Camera model

We consider EMCCD and sCMOS cameras — the most common cameras used in SMLM — using the standard noise model adopted in SMLM simulations (Sage et al., 2019; Speiser et al., 2021). Let μ denote the expected number of photons incident on one camera pixel, as given by the optical model $\mathcal{A}$ in Eq. (1). The number of photoelectrons is modeled as

$$n_e \sim \mathrm{Poisson}(q\mu + c), \tag{29}$$

where q is the quantum efficiency and c the mean spurious charge. This expression combines photon shot noise, quantum efficiency, and spurious charge into a single Poisson variable.

An EMCCD camera amplifies these electrons through a stochastic multiplication register (Hirsch et al., 2013). Given n_e input electrons, we model the output signal as

$$n_a \sim \mathrm{Gamma}(n_e, g_{\mathrm{EM}}), \tag{30}$$

where g_{EM} is the electron multiplication gain, and the Gamma distribution is parameterized by its shape and scale. For an sCMOS camera, there is no such amplification and we simply set $n_a = n_e$.

Finally, both cameras are subject to additive readout noise,

$$n_r = n_a + \varepsilon, \quad \text{where } \varepsilon \sim \mathcal{N}(0, \sigma_R^2), \tag{31}$$

and σ_R is the readout noise. The electrical signal is converted to analog-to-digital units (ADU) according to

$$y = \min(\max(\lfloor g_{\mathrm{ADU}} n_r + b_{\mathrm{ADU}} \rfloor, 0), y_{\mathrm{sat}}), \tag{32}$$

where g_{ADU} is the conversion factor in ADU/electron, b_{ADU} is the camera baseline, and y_{sat} is its saturation level (e.g. $y_{\mathrm{sat}} = 65535$ for images coded with 16 bits). Ignoring additive offsets and saturation, the resulting photon-to-ADU gain is therefore

$$G = \begin{cases} q\, g_{\mathrm{EM}}\, g_{\mathrm{ADU}}, & \text{EMCCD,} \\ q\, g_{\mathrm{ADU}}, & \text{sCMOS.} \end{cases} \tag{33}$$

Table 3 summarizes the parameters used for two widely used SMLM cameras.

C Distribution models

We describe here the two neural posteriors of our first objective, and the parametric distribution families we fit in our second objective.

C.1 Neural posteriors

The emitter configuration posterior is similar to Séailles et al. (2026), without their recursive branch. We let F be the block size, and $H \times W$ the frame size, such that $\boldsymbol{y} \in \mathbb{R}^{F\times H\times W}$. We set the number of candidates the emitter posterior outputs to $\bar{K} = HW/4$.

Table 3: Parameters of the two Photometrics cameras used in our experiments (Section 5.2).

Parameter	Symbol	Evolve 512 Delta	Prime 95B
Camera type	—	EMCCD	sCMOS
Quantum efficiency	q	0.9	0.95
Spurious charge (e^-/pixel/frame)	c	0.002	0.002
Electron multiplication gain	g_{EM}	300	1
Readout noise (e^-)	σ_R	74.4	1.535
Conversion factor (ADU/e^-)	g_{ADU}	$1/45$	$1/0.7471$
Baseline (ADU)	b_{ADU}	100	100
Saturation level (ADU)	y_{sat}	65535	65535

Shared trunk. A U-Net (Ronneberger et al., 2015) of depth 2 with $d = 48$ channels is applied to each frame of the block independently. Its F output feature maps, one per frame, are concatenated along channels and passed to a second U-Net of the same depth and width, which returns one feature map of d channels for the block. Each U-Net level uses one ResNet block, that is a sequence of 3×3 convolutions (without bias), followed by a layer normalization (Ba et al., 2016) and a SiLU activation (Hendrycks & Gimpel, 2016). Downsampling is a 2×2 max-pooling, and upsampling a transposed convolution with a skip connection.

Emitter head. A 2×2 max-pooling followed by a ResNet block brings the block feature map to $2d$ channels at half resolution, and a 1×1 convolution then gives an $H/2 \times W/2$ grid of $\bar{K}$ candidates with $7 + 2F$ channels each. Therefore, each candidate j is located on a slot of the half-resolution grid, whose centre we write $\boldsymbol{u}_j^{xy} \in \mathbb{R}^2$. Its $7 + 2F$ channels are respectively (1) a detection logit ℓ_j, (2-4) a mean $\boldsymbol{m}_j^{xyz}$ over its position, (5-7) a standard deviation $\boldsymbol{\sigma}_j^{xyz}$ over its position, (7-7 + F) a mean $\boldsymbol{m}_j^{\alpha}$ over its photon-count sequence, and (7 + F-7 + $2F$) a standard deviation $\boldsymbol{\sigma}_j^{\alpha}$ over its photon-count sequence. Its detection probability is $s_j = (1 + e^{-\ell_j})^{-1}$, and its lateral position is read off the two corresponding logit channels $\boldsymbol{o}_j$ by

$$\boldsymbol{m}_j^{xy} = \boldsymbol{u}_j^{xy} + 1.5 \tanh(\boldsymbol{o}_j). \tag{34}$$

The offset reaches 1.5 pixels, so that the reach of neighbouring slots overlaps and several emitters falling in the same slot can be predicted by its neighbours (Séailles et al., 2026). Emitters lying within 4.5 pixels of the border are excluded by construction, which matches the border rule imposed on the candidate grid (Appendix C).

Optimal-transport objective. Evaluating $-\log q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ requires matching the $\bar{K}$ candidates to the K target emitters, which Séailles et al. (2026) relax into an entropy-regularized optimal transport problem solved by the Sinkhorn algorithm (Cuturi, 2013). We keep that formulation with a small difference: we compute an uncertainty estimate for each detection, i.e. a per-detection $\boldsymbol{\sigma}_j$ instead of a global $\boldsymbol{\sigma}$. We run 200 Sinkhorn iterations in double precision, with an entropic regularization set to 10^{-2} times the median admissible cost of the batch (Flamary et al., 2021), and we treat the resulting plan as a constant of the loss. Unreachable pairs are given ten times the largest admissible cost of the batch rather than an infinite one, which keeps the Sinkhorn kernel finite.

PSF and background heads. The PSF posterior $q_\eta(\boldsymbol{c} \mid \boldsymbol{y})$ reads the per-frame features, then two downsampling blocks to $4d$ channels and a global average pooling give one vector per frame. The F vectors are averaged over the block, and a residual block (He et al., 2016) followed by a linear layer outputs the mean and the log standard deviation of a diagonal Gaussian over $\boldsymbol{c}$.

For the background, a depth-1 U-Net on the fused features outputs the mean and log standard deviation of a diagonal Gaussian over the reference background $\boldsymbol{\beta}$ of Appendix E, one pair per pixel.

The PSF head gives the second term of Eq. (3) as a Gaussian negative log-likelihood averaged over coordinates, and the background head adds a third term of the same form, averaged over pixels.

Detectability. We only seek *detectable* emitters: an emitter is included as a localization target only if its signal-to-noise ratio (SNR) under the imaging model exceeds a user-specified threshold $\mathrm{SNR}_{\min}$. This threshold defines the population the network learns to detect, and therefore affects

the precision–recall trade-off. Simulated emitters below this threshold still contribute to the image, but are excluded from the localization targets. The SNR calculation is given in Appendix E.

C.2 The fitted and reference distributions

Emitter count. A single learnable scalar gives a per-candidate activation probability p through a sigmoid. The number of emitters of a simulated block is drawn as $K \sim \mathrm{Bin}(\bar{K}, p)$. In practice, $\bar{K}$ is much greater than the effective number of active emitters in a block, hence the approximation is valid. We clamp p in $[0, 0.1]$. Furthermore, as p stays relatively low, the binomial distribution becomes a good approximation of a Poisson distribution for p low, which connects with our Poisson point process modeling. During training, the learnable weight of p is updated via the first divergence of Eq. (7), written between the expected count of the simulator and the expected count $\sum_j s_j$ inferred on real blocks, both read as Poisson rates.

The reference $p_0(K)$ is uniform over $\{0, \dots, \bar{K}\}$.

Spatial density. We model $q_\phi(\boldsymbol{r})$ by a flow on the box $[0, W] \times [0, H] \times [-1, 1]$, whose base distribution is uniform, and we fit it by conditional flow matching (Lipman et al., 2022). Its velocity field is a linear layer, four residual blocks of width 512, and a linear layer, applied to Fourier features of the position and of the time (Rahimi & Recht, 2007). The 1024 spatial features have random directions and frequencies log-uniform below a ceiling, and a low-pass gate selects those the field may use. We set the gate from the resolution the posterior asks for, namely an exponential moving average of the predicted $\boldsymbol{\sigma}_j^{xyz}$ weighted by the detection probabilities, which lets the density sharpen only as the localizations do.

To sample, we push a uniform draw through the field with a second-order Adams-Bashforth integrator in 10 steps (Hairer et al., 1993), and fold the result back into the box by mirroring. The flow-matching targets are drawn from the candidate Gaussians of real blocks, sampling slots with probability proportional to s_j. We couple them to their base points by an entropic optimal transport plan, which straightens the learnt paths (Peyré et al., 2019; Tong et al., 2023).

The reference $p_0(\boldsymbol{r})$ is uniform over $[0, W] \times [0, H] \times [-1, 1]$.

Photophysics. In $q_\phi(\boldsymbol{\alpha})$, the photophysics of an emitter factor into two elements: a photon flux, drawn once, and an ON burst with each frame. First, the photon flux is drawn from a 1D distribution, that is stored as its inverse cumulative distribution function on 32 knots uniformly spaced in probability. Second, the burst has an exponential length of learnable mean τ_{on} and a starting phase uniform over $[-(F-1), F)$, so that it may already be on when the block opens. We fit both against the photon-count marks of real blocks under a maximum mean discrepancy loss (Gretton et al., 2012).

The reference $p_0(\boldsymbol{\alpha})$ is uniform over the calibrated photon box $[-\kappa, \kappa]$, and truncated to positive values with a relu (see Appendix E for the definition of κ). Under the external calibration in Table 1, both $p_0(\boldsymbol{\alpha})$ and $q_\phi(\alpha)$ are the fixed three-state law of Sage et al. (2019), which is not learnt.

PSF and background. The learnt $q_\psi(\boldsymbol{c})$ is a Gaussian with learnable mean and diagonal covariance. Fitting it to the logarithmic pool of Eq. (6) amounts to minimizing the expected energy of $q_\psi(\boldsymbol{c})$ under each block's predicted Gaussian, minus the same energy under q_ψ itself held fixed, which divides out the focusing proposal as in Greenberg et al. (2019), plus $D_{\mathrm{KL}}(q_\psi \parallel p_0)$ divided by the number of blocks.

Its reference $p_0(\boldsymbol{c})$ is the diagonal Gaussian of Eq. (28).

The background is fitted in the same way, against the per-pixel Gaussians of its own head, with both of its terms averaged over pixels rather than summed. Its reference $p_0(\boldsymbol{\beta})$ carries the measured background map under a random symmetry of the frame.

Componentwise mixture. In practice, to increase samples diversity, we replace the mixture of point processes $\tilde{p}^\lambda(\boldsymbol{x})$ in Eq. (2) by a point process whose components are themselves mixtures with their respective learnt and reference distributions, using the same λ: thus, one configuration $\boldsymbol{x}$ can combine fitted and reference characteristics. The choice is made per emitter for positions and photon counts, and per block for the PSF parameters and the background.

D TRAINING DETAILS

D.1 LOSS FUNCTIONS

In $\mathcal{L}_0$, following Séailles et al. (2026), each of the $\bar{K}$ candidates predicted by the neural posterior $q_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ has a detection probability s_j and a diagonal Gaussian distribution over its position and photon-count sequence, with mean $\boldsymbol{m}_j$ and standard deviations $\boldsymbol{\sigma}_j$. The emitter term in $\mathcal{L}_0$ is

$$\ell_{0,\boldsymbol{x}}(\theta; \boldsymbol{x}, \boldsymbol{y}) = \langle \Gamma^*, C_\theta \rangle, \quad \text{where} \quad \Gamma^* = \underset{\Gamma \in \mathbb{B}}{\arg\min} \, \langle \Gamma, C_\theta \rangle - \epsilon H(\Gamma), \tag{35}$$

$\mathbb{B}$ is the Birkhoff polytope, and $C_\theta \in \mathbb{R}^{\bar{K} \times \bar{K}}$ contains the costs of assigning each candidate j to each emitter $k \leq K$, or to a dustbin for $k' > K$:

$$[C_\theta]_{j,k} = -\log s_j - \log \mathcal{N}\big((\boldsymbol{r}_k, \boldsymbol{\alpha}_k); \, \boldsymbol{m}_j, \boldsymbol{\sigma}_j\big), \quad \text{and } [C_\theta]_{j,k'} = -\log(1 - s_j). \tag{36}$$

The PSF term $\ell_{0,\boldsymbol{c}}$ is the Gaussian negative log-likelihood of $\boldsymbol{c}$ under $q_\eta(\boldsymbol{c} \mid \boldsymbol{y})$.

The loss function $\mathcal{L}_{1,\boldsymbol{x}}$ which fits $q_\phi(\boldsymbol{x})$, splits into one term per component of the point process. First, for $q_\phi(\boldsymbol{r})$, we draw $\bar{K}$ samples $(\hat{\boldsymbol{r}}_i, \hat{\boldsymbol{\alpha}}_i)_{1 \leq i \leq \bar{K}}$ from the aggregated posterior $\frac{1}{B}\sum_{i=1}^{B} q_\theta(\boldsymbol{x} \mid \boldsymbol{y}_i)$ on a minibatch of real blocks, by picking candidates with probabilities s_j and sampling from their Gaussians. We fit $q_\phi(\boldsymbol{r})$ to these samples by conditional flow matching (Lipman et al., 2022), with a velocity field v_ϕ and a uniform base distribution over Ω:

$$\mathcal{L}_{1,\boldsymbol{r}}(\phi) = \mathbb{E}_{t \sim \mathcal{U}([0,1]),\, \boldsymbol{r}_0 \sim \mathcal{U}(\Omega),\, i} \big\| v_\phi\big((1-t)\boldsymbol{r}_0 + t\hat{\boldsymbol{r}}_i,\, t\big) - (\hat{\boldsymbol{r}}_i - \boldsymbol{r}_0) \big\|^2. \tag{37}$$

Second, for $q_\phi(\boldsymbol{\alpha})$, we use a maximum mean discrepancy (Gretton et al., 2012), computed with random Fourier features φ (Rahimi & Recht, 2007) so that its cost is linear in the number of samples:

$$\mathcal{L}_{1,\boldsymbol{\alpha}}(\phi) = \Big\| \frac{1}{\bar{K}} \sum_{i=1}^{\bar{K}} \varphi(\hat{\boldsymbol{\alpha}}_i) - \frac{1}{\bar{K}} \sum_{i=1}^{\bar{K}} \varphi(\boldsymbol{\alpha}_i) \Big\|^2, \quad \text{with } \boldsymbol{\alpha}_i \sim q_\phi(\boldsymbol{\alpha}). \tag{38}$$

Third and last, we fit $q_\phi(K)$ by the KL divergence between two Poisson distributions,

$$\mathcal{L}_{1,K}(\phi) = D_{\mathrm{KL}}\big(\mathcal{P}(\hat{\mu}) \parallel \mathcal{P}(\mu_\phi)\big) = \hat{\mu} \log \frac{\hat{\mu}}{\mu_\phi} - \hat{\mu} + \mu_\phi, \tag{39}$$

where $\hat{\mu}$ is the average of $\sum_j s_j$ over the minibatch, and μ_ϕ the expected number of detectable emitters under $q_\phi(K)$. As the learnt activation probability stays low in $q_\phi(K)$, the binomial is well approximated by a Poisson distribution. Hence, $\mathcal{L}_{1,\boldsymbol{x}} = \mathcal{L}_{1,K} + \mathcal{L}_{1,r} + \mathcal{L}_{1,\boldsymbol{\alpha}}$.

Finally, the contribution of the minibatch to the second divergence in Eq. (5) is

$$\frac{1}{N} D_{\mathrm{KL}}\big(q_\psi \parallel p_0\big) - \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}_{\boldsymbol{c} \sim q_\psi(\boldsymbol{c})} \Big[\log \frac{q_\eta(\boldsymbol{c} \mid \boldsymbol{y}_b)}{\tilde{p}^\lambda(\boldsymbol{c})} \Big]. \tag{40}$$

As $\tilde{p}^\lambda(\boldsymbol{c})$ is a mixture, this expectation has no closed form. Hence, in $\mathcal{L}_{1,\boldsymbol{c}}$, we replace the proposal by $q_{\bar{\psi}}(\boldsymbol{c})$, a copy of $q_\psi(\boldsymbol{c})$ that is not differentiated through, so that every term is analytic. This approximation becomes nearly exact as training progresses, since the proposal is then dominated by $q_\psi(\boldsymbol{c})$ (Eq. (41)).

D.2 TRAINING.

Following SHOT (Séailles et al., 2026), we train for $M = 100{,}000$ iterations with a batch size of $B = 128$, using the AdamW optimizer (Loshchilov & Hutter, 2019) and a cosine annealing learning rate schedule. We use an accelerated, capped cosine schedule for the mixture weight:

$$\forall n \in \{1, \ldots, M\}, \; \lambda_n = \frac{0.95}{2} \left[1 - \cos\left(\pi \min\left(\frac{2n}{M}, 1 \right) \right) \right]. \tag{41}$$

The weight of the learnt component reaches 0.95 halfway through training and remains constant thereafter. Each proposal therefore retains a contribution of at least 5% from its reference prior, which we find to bring better results.

E CALIBRATION

Before training starts, three passes over the acquisition set the reference priors and the fixed constants of the imaging model. They estimate the background, the width of the PSF, and the photon scale of the emitters. These three estimates, together with the camera parameters of Appendix B.2 supplied by the user, are the only quantities fixed before training; everything else is learnt.

Background. We accumulate 15 per-pixel quantiles of these frames, at levels regularly spaced in $[0.01, 0.2]$. Each level is converted back to a photon count by inverting the camera model of Appendix B.2 at that quantile, which gives 15 candidate background maps $\boldsymbol{b}_1, \dots, \boldsymbol{b}_{15}$. We keep their pixelwise minimum $\boldsymbol{b}_0$: indeed, two biases at play push in the same direction: emitters still present in the higher levels raise the candidate backgrounds, and the skewed tail of the camera noise also raises the lower ones. Hence, the smallest of the 15 estimates is the safest estimate.

The background is not held fixed afterwards. It is a latent variable $\boldsymbol{\beta}$ with $\boldsymbol{b} = \boldsymbol{b}_0 \odot e^{\boldsymbol{\beta}}$, inferred by its own posterior head and fitted exactly like the PSF parameters $\boldsymbol{c}$ (Appendix C). Its reference prior is a per-pixel Gaussian, with mean $\boldsymbol{b}_0$ and standard deviation equal to the spread of the 15 estimates.

PSF width. On the background-subtracted frames, we fit the radially averaged power spectrum by a Gaussian on a flat noise floor,

$$P(f) = A \exp(-4\pi^2 \sigma^2 f^2) + C, \tag{42}$$

where f is the radial spatial frequency, σ the width of the average emitter image in pixels, and C is estimated as the median of the highest quarter of the frequencies. Fitting the floor explicitly matters: a pure Gaussian fit follows the spectrum into the noise and underestimates σ. We then select σ on a grid of 256 values over $[0.3, 5]$ pixels by least squares in log space, and the resulting $\mathrm{FWHM} = 2\sqrt{2\log 2}\,\sigma$ sets the reference lateral scale ℓ_{ref} of Eq. (24), so that the latent ℓ only has to carry the deviation from the measured spot size.

Photon scale. We convolve the background-subtracted frames with a matched filter for a Gaussian blob of that width: a Gaussian kernel of radius $\lceil 3\sigma \rceil$, of null total energy, which makes it blind to a constant background, and normalized so that a one-photon blob gives a response of one. Its local maxima over 3×3 neighbourhoods whose SNR is above a threshold are detected, and we sketch the quantiles of their responses. Their median α_{ref} is treated as the photon unit of the whole model: every photon count in $p_0(\boldsymbol{\alpha})$, in $q_\phi(\boldsymbol{\alpha})$ and in the forward model Eq. (1) is expressed in units of α_{ref}. Their 0.99 quantile, divided by α_{ref}, gives the maximal accepted photon value.

Detectability. An emitter is a localization target only if its signal-to-noise ratio exceeds a user-specified $\mathrm{SNR}_{\min}$, which is proposed by the user. For an emitter at $\boldsymbol{r}$ emitting α photons on a frame, we define it as

$$\mathrm{SNR} = \frac{\alpha_{\mathrm{ref}}\, \alpha\, \|\boldsymbol{h}_{\boldsymbol{c}}(\boldsymbol{r})\|_2}{\sqrt{\mathrm{Var}[\boldsymbol{b}_0(\boldsymbol{r})]}}, \tag{43}$$

where $\|\boldsymbol{h}_{\boldsymbol{c}}(\boldsymbol{r})\|_2$ is the Euclidean norm of the PSF patch, and $\mathrm{Var}[\boldsymbol{b}_0(\boldsymbol{r})]$ the camera noise variance at the calibrated background under the emitter.

F EVALUATION METRICS

F.1 METRICS FOR SYNTHETIC DATA

Bold values Results are mean $\pm$ standard deviation over multiple training seeds. For each setting, the best mean and results not significantly different from it are in bold. We use a two-sided Welch's t-test (unadjusted $p \geq 0.05$).

Detection. For synthetic data, ground-truth emitter configurations are available. In each frame, we match the predicted localizations to the ground-truth emitters through a Hungarian assignment based on squared Euclidean distances. We retain pairs whose lateral distance is at most $250\,\mathrm{nm}$ and axial distance at most $500\,\mathrm{nm}$. After matching, ground-truth emitters considered non-significant by the

challenge's rules are excluded, together with their associated predictions if they got matched. Non significant emitters include the 25% low quantile in terms of photon count, and the ones in a 450nm range of the border of the image. This procedure is the exact pipeline of Sage et al. (2019). Among the remaining localizations, matched pairs are *true positives* (TP), while unmatched predictions and ground-truth emitters are *false positives* (FP) and *false negatives* (FN), respectively. Then, we compute

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{Jaccard} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{44}$$

Precision measures the fraction of predictions that are correct, recall the fraction of target emitters that are detected, and the Jaccard index accounts for both missed emitters and false detections.

Localization. Localization errors are evaluated only on true positives. Let $(\Delta x_k, \Delta y_k, \Delta z_k)$ be the difference between the predicted and true position of the k-th matched emitter. Following Sage et al. (2019), the lateral and axial root mean square errors are

$$\text{RMSE}_{\text{lat}} = \sqrt{\frac{1}{\text{TP}} \sum_{k=1}^{\text{TP}} (\Delta x_k^2 + \Delta y_k^2)}, \qquad \text{RMSE}_{\text{ax}} = \sqrt{\frac{1}{\text{TP}} \sum_{k=1}^{\text{TP}} \Delta z_k^2}. \tag{45}$$

Both errors are expressed in nanometers. Lower values indicate more accurate localizations.

3D efficiency. The efficiency of Sage et al. (2019) combines detection and localization performance into a single score. We normalize it to have a maximum value of one:

$$\begin{aligned} \text{E}_{\text{lat}} &= 1 - \sqrt{(1 - \text{Jaccard})^2 + \alpha_{\text{lat}}^2 \text{RMSE}_{\text{lat}}^2}, \\ \text{E}_{\text{ax}} &= 1 - \sqrt{(1 - \text{Jaccard})^2 + \alpha_{\text{ax}}^2 \text{RMSE}_{\text{ax}}^2}, \\ \text{E}_{\text{3D}} &= \frac{\text{E}_{\text{lat}} + \text{E}_{\text{ax}}}{2}, \end{aligned} \tag{46}$$

where $\alpha_{\text{lat}} = 0.01\,\text{nm}^{-1}$ and $\alpha_{\text{ax}} = 0.005\,\text{nm}^{-1}$. These weights follow the challenge definition. All the metrics of this section are computed frame by frame and then averaged over the acquisition.

F.2 Metrics for real acquisition data

Image resolution. For real acquisitions, ground-truth emitters are unavailable. We estimate the lateral resolution of the reconstructed images using *Fourier ring correlation* (FRC) (Banterle et al., 2013). Our implementation follows the one of Ries (2020). We divide the acquisition into blocks of 100 consecutive frames, assign alternating blocks to two subsets, and render one image from the 2D localizations of each subset. This block split limits correlations due to emitters remaining active across neighboring frames. FRC measures the normalized correlation between the Fourier transforms of these images over rings of equal spatial frequency. We report the inverse of the first frequency at which the smoothed FRC curve crosses below $1/7$. Lower values indicate finer resolved details.

Image fidelity. We use the *resolution-scaled Pearson coefficient* (RSP) of SQUIRREL (Culley et al., 2018) to measure agreement with a reference image of the specimen. As discussed by Culley et al. (2018), using the mean of the acquisition as the reference can be unreliable for super-resolution methods that show out-of-focus emitters, such as SMLM. Hence, we use the pixel-wise temporal standard deviation of the raw acquisition as a reference, which is less affected by this phenomenon. On our data, it reveals the specimen more clearly than the mean image, which is dominated by background. We align the reconstruction with this reference, convolve it with a Gaussian resolution-scaling function, and average it onto the reference pixel grid. The Gaussian width is fitted to maximize the Pearson correlation, which gives RSP. Values close to one indicate strong agreement with this reference at its spatial scale.