Template-free 2D particle fusion in localization microscopy

Heydarian, Hamidreza; Schueder, Florian; Strauss, Maximilian T.; van Werkhoven, Ben; Fazel, Mohamadreza; Lidke, Keith A.; Jungmann, Ralf; Stallinga, Sjoerd; Rieger, Bernd

DOI
10.1038/s41592-018-0136-6

Publication date
2018

Document Version
Final published version

Published in
Nature Methods

Citation (APA)
Heydarian, H., Schueder, F., Strauss, M. T., van Werkhoven, B., Fazel, M., Lidke, K. A., Jungmann, R., Stallinga, S., & Rieger, B. (2018). Template-free 2D particle fusion in localization microscopy. *Nature Methods*, *15*, 781-784. https://doi.org/10.1038/s41592-018-0136-6

nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-018-0136-6

Template-free 2D particle fusion in localization microscopy

Hamidreza Heydarian[1], Florian Schueder[2,3], Maximilian T. Strauss[2,3], Ben van Werkhoven[4], Mohamadreza Fazel[5], Keith A. Lidke[5], Ralf Jungmann[2,3], Sjoerd Stallinga[1] and Bernd Rieger[1*]

Methods that fuse multiple localization microscopy images of a single structure can improve signal-to-noise ratio and resolution, but they generally suffer from template bias or sensitivity to registration errors. We present a template-free particle-fusion approach based on an all-to-all registration that provides robustness against individual misregistrations and underlabeling. We achieved 3.3-nm Fourier ring correlation (FRC) image resolution by fusing 383 DNA origami nanostructures with 80% labeling density, and 5.0-nm resolution for structures with 30% labeling density.

Single-molecule localization microscopy (SMLM) provides the ability to image well below the diffraction limit[1]. The resolution in the final reconstructed image is limited by localization uncertainty and emitter density to about 20 nm (ref. [2]). The fusion of multiple acquisitions into one hyper-resolved reconstruction can mitigate these limiting factors when many identical copies of the same structure (particle) can be imaged[3,4]. The final reconstruction has effectively many more localizations than each individual SMLM image, which results in a better signal-to-noise-ratio and thus better resolution. This approach is similar to single-particle analysis in cryo-electron microscopy (cryo-EM)[5,6]. A few studies have applied single-particle analysis to SMLM despite the fundamental differences in image formation between cryo-EM and SMLM[4,7–10]. Most important, fluorescent labeling is often incomplete, with only 30–70% density of labeling (DOL) typically achieved[11]. Statistical variations in localization uncertainty, false positive localizations[12,13] and repeated localizations of the same fluorophore are additional complications not encountered in cryo-EM. Some methods for SMLM data fusion use a template[3,4,8,14] for alignment, which carries the risk of generating a structure that is biased toward the template[15]. A template-free pyramid registration approach for SMLM datasets[14] registers N particles pairwise into $N-1$ reconstructions, after which a second set of pairwise registrations reduces the number of reconstructions to $N-2$, and so on. This method, like any iterative method of combining pairwise registrations, suffers from considerable sensitivity to registration errors in the initial step, which are propagated into subsequent phases of the procedure.

Here we present a particle-fusion approach that assumes no prior knowledge of the structure to be imaged (template-free), works directly on the localization data (including the uncertainties) and is robust against registration errors and underlabeling. The key idea is to use an all-to-all registration procedure in which each particle is registered to all the others, implying $N(N-1)/2$ pair registrations for N given particles. This generates the maximum amount of information that can be extracted from the alignment of N particles. Each pair registration results in an estimate of the relative orientation and position of the two particles. What is needed, however, are the N absolute orientations and positions of all particles. For this step we use a technique from the field of computer vision in which camera position and orientation are estimated from a sequence of images ('structure from motion')[16]. Lie-algebraic representations of the transformation parameters (rotations and translations) are averaged in this technique, which provides robustness to outlier pair registrations (details in the Supplementary Material and the Methods). In this way, we make optimal use of the very large redundancy in the $N(N-1)/2$ pair registration parameters and overcome the main flaw of any iterative registration method. A critical improvement over ref. [16] is the use of a sparsity-promoting L_1 norm (Methods). The performance is further improved via self-consistency. The found N absolute transformation parameters are used to retrodict the $N(N-1)/2$ relative transformation parameters, which can then be compared with the values found from the all-to-all registration. Registration pairs with a deviation in these relative transformation parameters that is too high (defined by a suitable threshold) are discarded before the second round of Lie-algebraic averaging. This removal of outlier registration pairs results in a reconstruction that is used in a final step as a data-driven model to bootstrap the registration process. This last step is especially effective for samples with low DOL.

The major drawback of all-to-all registration is the computational cost, which scales as N^2 instead of as N for a template-based registration. Therefore, we devised a computationally efficient implementation of the registration algorithm. The starting point is the Bhattacharya cost function[14]:

$$D=\sum_{i=1}^{K_t}\sum_{j=1}^{K_m}\exp\left(-(\vec{x}_{t,i}-M(\vec{x}_{m,j}))^{T}(\Sigma_{t,i}+R\Sigma_{m,j}R^{T})^{-1}(\vec{x}_{t,i}-M(\vec{x}_{m,j}))\right) \quad (1)$$

where the two particles t and m are represented by K_t and K_m localizations, $\vec{x}_t,\vec{x}_m\in R^n$ are the localization coordinates of the two particles (where n is the number of spatial dimensions), $\Sigma_t,\Sigma_m\in\mathbb{R}^{n\times n}$ are the corresponding uncertainties of the form $\Sigma=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$, and $M(\vec{x})=R\vec{x}+\vec{s}$ is the function that applies rotation (R) and translation ($\vec{s}$) to position $\vec{x}$. This cost function has the advantage that it works directly on localization data, thereby eliminating the need for a pixelated representation of the SMLM data. Furthermore,

[1]Department of Imaging Physics, Delft University of Technology, Delft, the Netherlands. [2]Department of Physics and Center for Nanoscience, Ludwig Maximilian University, Munich, Germany. [3]Max Planck Institute of Biochemistry, Martinsried, Germany. [4]Netherlands eScience Center, Amsterdam, the Netherlands. [5]Department of Physics and Astronomy, University of New Mexico, Albuquerque, NM, USA. *e-mail: b.rieger@tudelft.nl

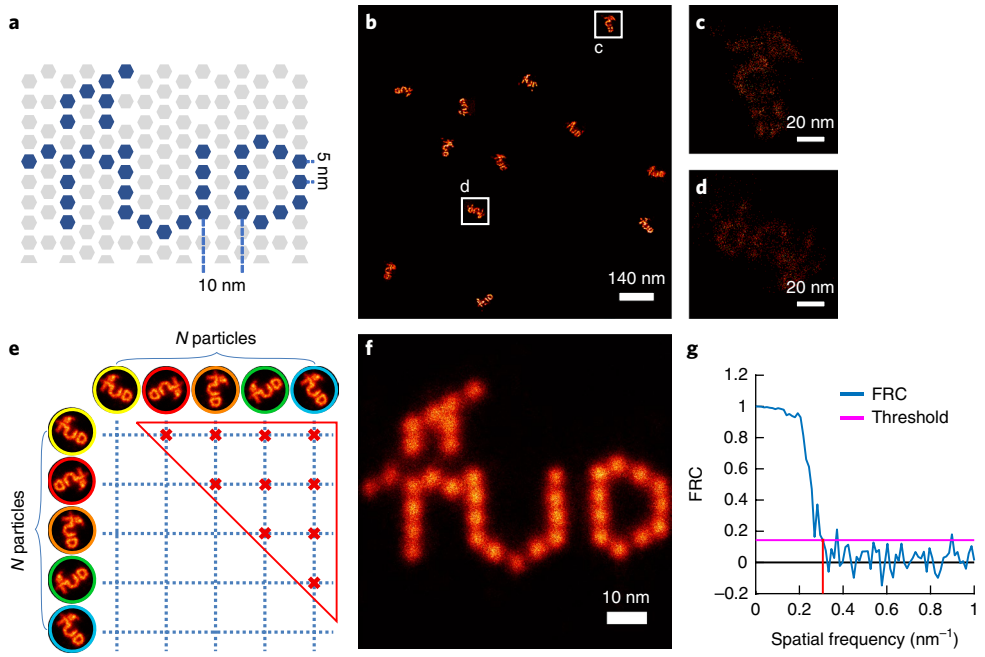

Fig. 1 | Template-free 2D particle averaging for localization microscopy. a, Schematic representation of the DNA origami for a grid structure designed to carry 37 DNA-PAINT docking sites, generating the 'TUD' logo. Strands are color-coded to denote strand extensions. **b**, Regions of interest in a DNA-PAINT SMLM image reconstructed from many particles. **c**,**d**, Magnified images of highlighted areas c and d in **b** depicting similar copies of the structure with different orientations. **e**, All-to-all registration schematic showing all $N(N-1)/2$ pairwise registrations (red crosses). The N absolute registration parameters (translation and rotation) are robustly obtained from the redundant relative registration parameters. **f**, The final reconstruction resulting from fusion of 383 individual particles. **g**, The average FRC curve for the final reconstruction in **f**, showing an image resolution of 3.3 ± 0.3 nm (the red line indicates the intersection of the FRC curve with the threshold).

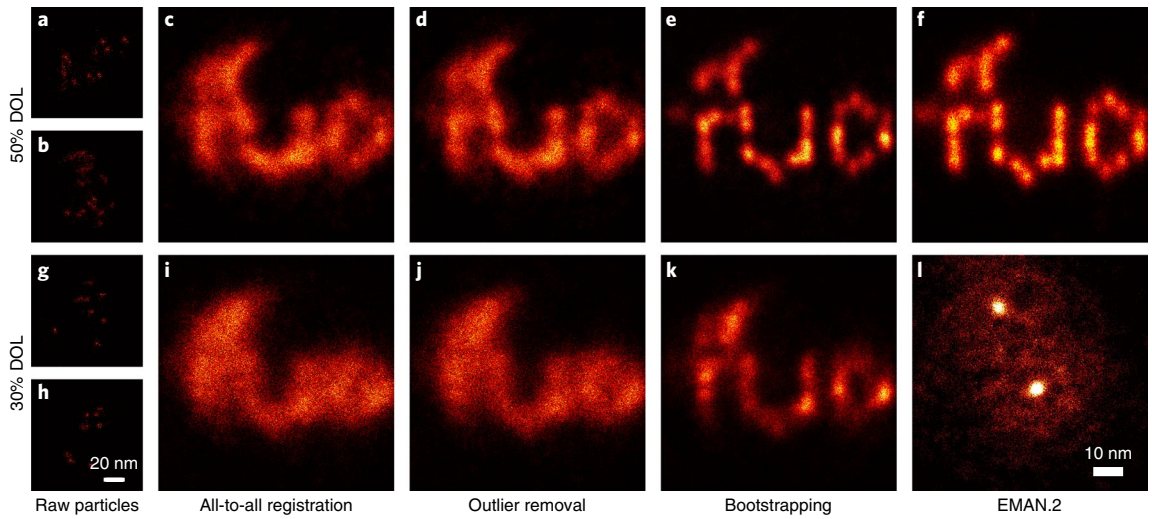

Fig. 2 | Particle fusion for underlabeled datasets. a,**b**, Two example raw particles with 50% DOL. **c**–**e**, The evolving super-particle for the dataset with 50% DOL. All-to-all registration and averaging of 442 logos with 50% DOL resulted in the blurry reconstruction shown in **c**. This was improved by the removal of outlier registrations (**d**), and the resulting image was subsequently used as the data-driven model in an all-to-template registration (bootstrapping). The final reconstruction in **e** illustrates the super-particle with around 280,000 localizations. **f**, Reconstruction generated by EMAN.2 software with 253 included particles and the minimum of three classes for the class averaging. **g**,**h**, Two example raw particles with 30% DOL. **i**–**k**, The evolving super-particle for the dataset with 30% DOL. Similar to **c**–**e**, each image depicts the output of a step in the particle-fusion pipeline. The final reconstruction in **k** is the result of fusion of 549 logos with around 250,000 localizations. **l**, Reconstruction generated by EMAN.2 software with 113 included particles and the minimum of three classes for the class averaging. EMAN.2 did not produce any meaningful reconstruction for 30% DOL. In all of the reconstructions in each row, the number of localizations is the same except for the EMAN.2 images, where classification excludes 42% and 79% of the data for 50% and 30% DOL, respectively. Scale bar in **h** applies to **a**,**b**,**g**; scale bar in **l** applies to **c**–**f** and **i**–**k.**

it can take into account (varying and anisotropic) localization uncertainties. To reduce the computational cost, we prealign the particles by using a fast Gaussian-mixture-model-based registration method[17] (details in the Methods section). We further speed up the computation by implementing both the Gaussian mixture model and the Bhattacharya cost function evaluation on a GPU.

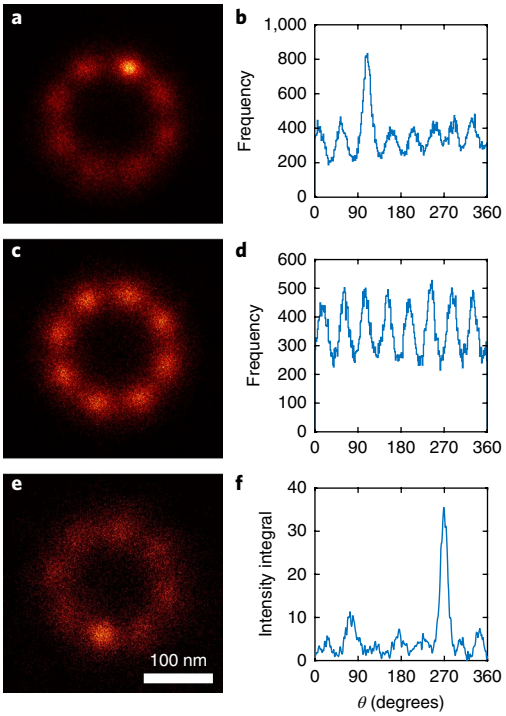

Fig. 3 | Particle fusion with (direct) STORM data for the NPC integral membrane protein gp210. a, Fusion of 304 NPCs by our method retrieved the eightfold symmetric ring structure without prior knowledge. **b**, Localization distribution over azimuthal angles of the reconstruction in **a**. **c**, Reconstruction after incorporation of the eightfold symmetry in the registration, which resolved the hot-spot artifact. **d**, Localization distribution over azimuthal angles of the reconstruction in **c**. **e**, Reconstruction by EMAN.2 resulted in 139 included particles (minimum of three classes for class averaging). **f**, Azimuthal intensity plot of **e**. In the EMAN.2 reconstruction, the hot-spot artifact is present, and the visibility is worse than that achieved with our method. Scale bar in **e** applies to **a**,**c**.

We tested our method on three different datasets with ~80%, ~50% and ~30% DOL, respectively. The datasets contained 2D DNA origami nanostructures with 37 designed binding sites on a hexagonal grid with 5-nm spacing, arranged to display the letters TUD (Fig. 1a). We imaged these nanostructures with DNA-PAINT[9] (Methods) to obtain an SMLM image (Fig. 1b–d shows part of the full field of view). Manual segmentation of the 80% DOL dataset resulted in 383 instances of the imaged logos, which we fed into the all-to-all registration pipeline (Fig. 1e). The average number of localizations per particle was around 2,060, and the mean localization uncertainty was 0.81 ± 0.26 nm. Figure 1f shows our final reconstruction, which has excellent correspondence to the designed origami (Supplementary Videos 1 and 2). The image resolution, quantified with the FRC value[2], was 3.3 nm for the reconstruction (Fig. 1g), a value close to $\lambda/175$ (with $\lambda = 580$ nm as the fluorescence emission wavelength). The reconstruction quality for the outside of the origami was consistent with the low incorporation efficiency of strands on the outside of origamis observed previously[18].

We further benchmarked the proposed particle-fusion algorithm with lower-DOL PAINT data. Figure 2 illustrates the evolution of the reconstruction through the different steps of the fusion pipeline. We obtained these images by fusing (manually segmented) 442 (Fig. 2a–c and Supplementary Videos 3 and 4) and 549 (Fig. 2d–f and Supplementary Videos 5 and 6) TUD logos for the 50% and 30% DOL datasets, respectively. With an average number of localizations per particle of 630 and 453 for 50% and 30% DOL, respectively, we were able to reconstruct the logo, which was unrecognizable in the raw data (Fig. 2c,f and Supplementary Videos 4 and 6). The final reconstructions for 50% and 30% DOL data consisted of around 280,000 and 250,000 localizations, respectively. The fusion of raw particles led to an FRC resolution of 3.5 nm and 5.0 nm for 50% and 30% DOL, respectively. We also compared our approach with the very popular cryo-EM software package EMAN.2[5]. Although the software-based reconstruction for 50% DOL was similar in visual appearance to that obtained via our approach, EMAN.2 did not produce any part of the logo for 30% DOL (Fig. 2f,l). The final reconstruction quality in our method seemed to be limited by residual drift on the order of 1–2 nm, as assessed by visual comparison with simulated data (Supplementary Fig. 1). This was further supported by simulation results without drift (Supplementary Note 1 and Supplementary Fig. 2). In additional simulations we observed that average localization uncertainties larger than the binding-site distance resulted in unresolvable binding sites in the final reconstruction, even with perfect registration (Supplementary Fig. 3).

Although our particle-fusion method performed robustly even for low DOLs, it could not neutralize the effect of false positive localizations. We used conventional single-emitter fitting followed by localization filtering on the basis of the local sparsity of localization events (details in the Methods; compare Supplementary Videos 1 and 2, as well as Supplementary Videos 7 and 8). As an alternative, methods that better handle spatially proximate emitters[19] or that detect and remove false positives[13] can be used (Supplementary Fig. 4).

In addition to the above PAINT data, we applied our method to experimental and simulated localization data that included bleaching, and therefore had a different statistical distribution of localization events per fluorophore[20]. We analyzed stochastic optical reconstruction microscopy (STORM) images acquired from the integral membrane protein gp210 in the nuclear pore complex (NPC; data described previously[3]), from which we manually segmented 304 NPCs with an average of 313 localizations. Figure 3a shows our reconstruction, which reproduced the ring structure of the NPC without any prior assumptions. The 'hot spot' in Fig. 3a,b is a reconstruction artifact caused by the nonuniform distribution of localizations over the eight sites of each individual NPC. This statistical variation is enhanced during the registration step, as rings are most likely to be registered such that the sites with more than the average number of localizations become aligned. One can eliminate this artifact by taking into account the symmetry as prior knowledge (but not any other structural information). We randomly added multiples of $2\pi/8$ to the obtained absolute estimated rotation angles, which led to the uniform distribution depicted in Fig. 3c,d. Using EMAN.2, we obtained the averages shown in Fig. 3e,f, exhibiting a set of eight blobs with less visibility than achieved with our method, and showing the same hot spot artifact. We did not succeed in including the symmetry in EMAN.2, as there was no access to the estimated absolute angles and no way to explicitly impose the symmetry.

We also tested the applicability of our method to simulated STORM images (Supplementary Note 1). We generated TUD logos at 65% DOL with three different bleaching rates corresponding to average numbers of localizations per site of ~33, ~13 and ~7, respectively. Our method successfully reconstructed the logo for all three sets, whereas EMAN.2 succeeded only at the lowest bleaching rate (Supplementary Fig. 5). We compared STORM with PAINT-type data, keeping the labeling density and average number of localizations per particle the same (Supplementary Fig. 6), and found that the STORM images required a higher labeling density for successful reconstruction (~50% DOL, compared with ~30% for PAINT). We attribute this to bleaching effects, which skew the distribution of localizations per binding site and thus effectively decrease the fraction of sites with sufficiently high labeling density.

In summary, we have developed a template-free 2D particle-fusion algorithm for SMLM data that is robust to poor experimental conditions. We benchmarked the performance on PAINT data, and achieved a resolution of 3.3 nm for 80%, 3.5 nm for 50% and 5.0 nm for ~30% DOL. We successfully reconstructed an eight-fold symmetric ring structure from STORM data without a priori structural information. The framework can be generalized to treat 3D data, as individual subcomponents of the pipeline are not restricted to 2D.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41592-018-0136-6.

Received: 12 March 2018; Accepted: 23 July 2018;
Published online: 17 September 2018

References

1. Klein, T., Proppert, S. & Sauer, M. *Histochem. Cell Biol.* **141**, 561–575 (2014).
2. Nieuwenhuizen, R. P. J. et al. *Nat. Methods* **10**, 557–562 (2013).
3. Löschberger, A. et al. *J. Cell Sci.* **125**, 570–575 (2012).
4. Szymborska, A. et al. *Science* **341**, 655–658 (2013).
5. Tang, G. et al. *J. Struct. Biol.* **157**, 38–46 (2007).
6. Kudryashev, M., Castaño-Díez, D. & Stahlberg, H. *Comput. Struct. Biotechnol. J.* **1**, e201207002 (2012).
7. Van Engelenburg, S. B. et al. *Science* **343**, 653–656 (2014).
8. Gray, R. D. M., Mercer, J. & Henriques, R. *J. Vis. Exp.* **2017**, e55471 (2017).
9. Schnitzbauer, J., Strauss, M. T., Schlichthaerle, T., Schueder, F. & Jungmann, R. *Nat. Protoc.* **12**, 1198–1228 (2017).
10. Salas, D. et al. *Proc. Natl Acad. Sci. USA* **114**, 9273–9278 (2017).
11. Burgert, A., Letschert, S., Doose, S. & Sauer, M. *Histochem. Cell Biol.* **144**, 123–131 (2015).
12. van de Linde, S., Wolter, S., Heilemann, M. & Sauer, M. *J. Biotechnol.* **149**, 260–266 (2010).
13. Fox-Roberts, P. et al. *Nat. Commun.* **8**, 13558 (2017).
14. Broeken, J. et al. *Methods Appl. Fluoresc.* **3**, 014003 (2015).
15. Henderson, R. *Proc. Natl Acad. Sci. USA* **110**, 18037–18041 (2013).
16. Govindu, V. in *Proc. 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition: CVPR 2004* (eds Davis, L. et al.) 684–691 (IEEE, Piscataway, NJ, 2004).
17. Jian, B. & Vemuri, B. C. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 1633–1645 (2011).
18. Strauss, M. T., Schueder, F., Haas, D., Nickels, P. C. & Jungmann, R. *Nat. Commun.* **9**, 1600 (2018).
19. Huang, F., Schwartz, S. L., Byars, J. M. & Lidke, K. A. *Biomed. Opt. Express* **2**, 1377–1393 (2011).
20. Nieuwenhuizen, R. P. et al. *PLoS One* **10**, e0127989 (2015).

Acknowledgements

We thank A. Chatterjee for providing the code for transformation averaging. This work was supported by the European Research Council (Nano@cryo, grant no. 648580 to H.H. and B.R.; MolMap, grant no. 680241 to R.J.), the eScience Center (path finder grant 027016P04 to B.v.W. and B.R.), the NIH (grants 1R21EB019589 and P50GM085273 to K.A.L. and M.F.), the New Mexico Spatiotemporal Modeling Center (K.A.L. and M.F.), the International Max Planck Research School for Molecular and Cellular Life Sciences (IMPRS-LS; to M.T.S.), the Max Planck Society (R.J.), the Max Planck Foundation (R.J.), the DFG (Emmy Noether Program; DFG JU 2957/1-1 to R.J.), the SFB 1032 (Nanoagents for the spatiotemporal control of molecular and cellular reactions; to R.J.) and the Center for Nanoscience (CeNS; R.J.). K.A.L. and M.F. acknowledge the UNM Center for Advanced Research Computing, supported in part by the National Science Foundation, for providing high-performance computing resources.

Author contributions

S.S. and B.R. conceived the project. H.H. performed simulations. B.v.W. wrote GPU code. M.F. and K.A.L. provided multi-emitter fitting data. M.T.S., F.S. and R.J. designed DNA origami and acquired images. H.H., S.S. and B.R. wrote the paper, and all authors commented on the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0136-6.

Correspondence and requests for materials should be addressed to B.R.

Methods

Materials. Unmodified, dye-labeled and biotinylated DNA oligonucleotides were purchased from MWG Eurofins. Streptavidin was purchased from Invitrogen (S-888). BSA–biotin was obtained from Sigma-Aldrich (A8549). Coverslips were purchased from Marienfeld (18×18 mm, #1.5; catalog number 0107032). Microscopy slides were ordered from Thermo Fisher Scientific (10756991). Double-sided adhesive tape was purchased from Scotch (665D). Epoxy glue was ordered from Toolcraft (TC-EPO5-24). M13mp18 scaffold was obtained from New England Biolabs (N4040s). Freeze 'N Squeeze columns were ordered from Bio-Rad (7326165). Agarose was obtained from Biomol (01280.100). 50× TAE buffer was ordered from Fluka Analytical (67996-10L-F). SYBR Safe DNA gel stain was purchased from Invitrogen (SS33102). DNA gel loading dye was ordered from Thermo Fisher Scientific (R06111). Protocatechuate 3,4-dioxygenase pseudomonas (PCD) (P8279), 3,4-dihydroxybenzoic acid (PCA) (37580-25G-F) and (+−)-6-hydroxy-2,5,7,8-tetra-methylchromane-2-carboxzlic acid (Trolox) (238813-5G) were obtained from Sigma. 1 M Tris, pH 8.0 (AM9856), 1 M magnesium (AM9530G), 0.5 M EDTA, pH 8.0 (AM9261), and 5 M NaCl (AM9759) were obtained from Ambion. H_2O (10977-035) was ordered from Gibco. Tween 20 was ordered from Sigma-Aldrich (p2287).

Microscopy setup. DNA-PAINT experiments were carried out on an inverted Nikon Ti-Eclipse microscope (Nikon Instruments) with the Perfect Focus System. For the experiment, an oil-immersion objective (Plan Apo 100×/1.49-NA (numerical aperture); Nikon Instruments) was used. As the excitation laser, a 561-nm (200 mW nominal; Coherent) was used. Excitation light was filtered with a laser clean-up filter (zet561/10×; Chroma Technology Corp). As the dichroic, a laser dichroic mirror was used (zt561rdc; Chroma Technology Corp). Fluorescent light was spectrally filtered with an emission filter (et575lp, et600/50 m; Chroma Technology Corp) and imaged on a scientific complementary metal-oxide semiconductor (sCMOS) camera (Zyla 4.2; Andor Technologies).

DNA origami self-assembly. The DNA origami structures were formed in a one-pot reaction with a 50-µl total volume containing 10 nM scaffold strand (M13mp18), 100 nM core staples, 1 µM biotinylated staples and 1 µM staples extended with DNA-PAINT docking sites for the 20-nm grid and 10-nm grid drift markers. For samples with the TUD logo at 80% DOL, we used a 1 µM concentration of staples extended for DNA-PAINT. For 50% DOL of the TUD logo, we used a mixture of 0.8 µM extended staples and 0.2 µM not-extended staples. For 30% DOL, we used a mixture of 0.6 µM extended staples and 0.4 µM not-extended staples. The folding buffer was 1× TE buffer with 12.5 mM $MgCl_2$. The structures were annealed using a thermal ramp. We incubated samples first for 5 min at 80 °C, then from 65 °C to 4 °C over the course of 3 h. After self-assembly, the structures were mixed with 1× loading dye and then purified by agarose gel electrophoresis (1.5% agarose, 0.5× TAE, 10 mM $MgCl_2$, 1× SYBR Safe) at 3 V/cm for 3 h. Gel bands were cut, crushed and filled into a Freeze 'N Squeeze column and spun for 5 min at 1,000*g* at 4 °C. As the DNA-PAINT docking site, we used a TT spacer followed by a 9-nt 3′ extension (5′-staple-TT-ATACATCTA-3′). The imager was the 9-nt reverse complement of the docking site with a Cy3b fluorescent molecule attached at the 3′ end (5′-TAGATGTAT-dye-3′).

Super-resolution DNA-PAINT imaging with DNA origami. For chamber preparation, a piece of coverslip (no. 1.5, 18×18 mm^2, ~0.17 mm thick) and a glass slide (3×1 $inch^2$, 1 mm thick) were held together by two strips of double-sided tape to form a flow chamber with an inner volume of ~20 µl. First, 20 µl of biotin-labeled bovine albumin (1 mg/ml, dissolved in buffer A (10 mM Tris-HCl, pH 7.5, 100 mM NaCl, 0.05% Tween 20, pH 7.5)) was flowed into the chamber and incubated for 2 min. Then the chamber was washed with 40 µl of buffer A. Second, 20 µl of streptavidin (0.5 mg/ml, dissolved in buffer A) was flowed through the chamber and incubated for 2 min. Next, the chamber was washed with 40 µl of buffer A and subsequently with 40 µl of buffer B (5 mM Tris-HCl, pH 8, 10 mM $MgCl_2$, 1 mM EDTA, 0.05% Tween 20, pH 8). Then ~100 pM TUD DNA origami structures, ~100 pM 10-nm grid DNA origami structures and ~200 pM 20-nm DNA origami structures were flowed into the chamber and allowed to bind for 2 min. Afterward the chamber was washed with 40 µl of buffer B. Finally, the imaging buffer with buffer B and 1× Trolox, 1× PCA and 1× PCD[9] with the Cy3b-labeled imager strand was flowed into the chamber. The chamber was sealed with epoxy before subsequent imaging. For the 30% and 80% DOL experiments, we used an imager concentration of 1 nM. For the 50% DOL experiment, we used an imager concentration of 2 nM. At the end, we determined the actual labeling densities by counting the number of occupied sites on each particle and comparing it with the number of designed sites on the logo (37).

For all three experiments (80%, 50% and 30% DOL), an Andor Zyla 4.2 with a readout bandwidth of 200 MHz at 16 bit was used. We applied 2×2 pixel binning, which resulted in an effective pixel size of 130 nm (taking the 100× magnification of the microscope into account). The recorded field of view was 512×512 pixels (66.5×66.5 µm). Acquisition frame rates of 2.86 Hz for the 50% and 30% DOL cases and 3.33 Hz for the 80% DOL case were used over the course of 100,000 frames. The excitation intensity was ~1.86 kW/cm^2 at 561 nm at the sample plane.

Single-emitter and multi-emitter fitting of experimental data. Single-molecule reconstruction and drift correction were performed as described[9]. The average numbers of photons per localization event for 80%, 50% and 30% DOL were 7.0×10^4, 4.9×10^4 and 5.3×10^4, and the background photon counts per frame per pixel were 1.1×10^3 for 80% and 50% and 0.9×10^3 for 30%. Average uncertainties were 0.96 nm, 1.33 nm and 1.28 nm for the 80%, 50% and 30% DOL datasets, respectively. The uncertainties were estimated per localization from the data as previously reported[21]. We post-processed data by omitting localizations with localization uncertainties of more than 2 nm. To reduce the effect of false positive localization in single-emitter-fitted datasets, we filtered 80% and 50% DOL datasets before fusion. In each segmented particle, localizations were discarded if there were fewer than ten localizations in a circular neighborhood of radius $r=0.015$ pixels around the localization of interest (Supplementary Fig. 7 describes the effects of filtering on the final reconstruction). The effect of false positives on the 30% DOL data was less serious because the probability of overlapping emission patterns was low in that dataset. Therefore, for single-emitter-fitted 30% DOL data, we did not filter the raw particles.

Multi-emitter fitting was performed in the following way. Sub-regions identified as containing TUD logos were selected from the raw data for multi-emitter fitting. Each time frame in each TUD-containing sub-region was analyzed independently. We carried out multi-emitter fitting by finding the posterior probability distribution of the parameters $\theta=\{x_1, y_1, I_1, \ldots, x_N, y_N, I_N, \alpha, \beta, \gamma\}$ using Markov chain Monte Carlo (MCMC), where x_n, y_n and I_n correspond to the location and intensity of the nth emitter, and α, β and γ parameterize a tilted-plane background model. The mean and s.d. of x_n, y_n were used for further analysis. We initialized the MCMC chain by first using a reversible-jump MCMC[22] procedure to find the most probable number of emitters and their locations. We created the point spread function (PSF) model used in the fitting by localizing, shifting and averaging together more than 100 high-signal single-emitter events from the raw data. We created a 4× sub-sampled PSF by padding the Fourier transform. The model of each single emitter was created by linear interpolation of the sub-sampled PSF and scaling by I. The x_n, y_n values were connected across time frames, and only binding events that spanned two or more frames were retained. False positive and large uncertainty localizations were removed from the data. A large uncertainty was defined as an s.d. larger than 0.0075 pixels. We removed false positives by keeping localizations if they had N_{min} localizations within a distance D. For each dataset, we found N_{min} and D by taking D as the median localization uncertainty (before thresholding) and N_{min} as the median number of localizations within a distance D. On the basis of visual inspection of several origami structures, we then adjusted N_{min} and D to minimize false localizations between docking stands while retaining as many localizations as possible. For 80%, 50% and 30% DOL labeling, the values used for N_{min} and D were 10, 10 and 9 localizations within 0.0075, 0.008 and 0.0095 pixels, respectively (pixel size: 130 nm).

The final distribution of localizations per particle (Supplementary Fig. 8) agreed qualitatively with the assessed DOL for the three datasets. The width of the distributions was an indication that the overall distribution was a convolution of the Poisson distribution of the number of localizations per binding site and the distribution of active binding sites according to the average DOL.

All-to-all registration. The developed particle-fusion algorithm consists of four main building blocks: (1) computation of the upper triangular matrix A that contains all relative registrations (Fig. 1e), which we call the all-to-all registration matrix; (2) calculation of the absolute orientations from these relative elements; (3) registration outlier removal; and (4) bootstrapping of the registrations.

Relative registrations. Each element of the matrix A is obtained through optimization of equation (1) in a coarse-to-fine manner. We aligned each pair of particles by using the Gaussian mixture model (GMM) registration method[17] with multiple initial angles. This provided us with a set of transformation parameters (rotation angle, translation vector), out of which we selected the set that maximized the Bhattacharya cost function as the final value for the set of relative transformation parameters for that pair. The GMM registration method minimizes the special case of the Bhattacharya cost function in which all localization uncertainties are equal:

$$D=\sum_{i=1}^{K_t}\sum_{j=1}^{K_m}\exp(-\|\vec{x}_{t,i}-M(\vec{x}_{m,j})\|^2/2\sigma^2) \quad (2)$$

For this case, there exists an analytical approximation to the problem with only linear computational complexity[17]. Here, σ is a tuning parameter that is dataset specific and which we set empirically to $0.01\times l$, where l is the camera pixel size in nanometers, for 80% and 50% DOL, and to $0.1\times l$ for 30% DOL experimental data. We optimized the GMM cost function by using the interior-point algorithm for multiple initial angles ranging from $-\pi$ to π evenly spaced by $\pi/4$. The Bhattacharya cost function equation (1) is evaluated for each of the found local optima of the GMM cost function, and the set of transformation parameters with the optimum Bhattacharya cost function is finally selected. This procedure results in an all-to-all registration matrix A with $N(N-1)/2$ relative registration parameters. Each

element a_{ij}, $i,j \in \{1,\ldots,N\}$, $\forall j > i$, of this matrix is the set of estimated relative rigid transformation parameters M_{ij} that aligns particle i to particle j:

$$M_{ij} = \begin{bmatrix} R_{ij} & t_{ij} \\ 0 & 1 \end{bmatrix} \quad (3)$$

with $R_{ij} \in SO(n)$ and $t_{ij} \in \mathbb{R}^{n\times 1}$ being the relative rotation matrix and the translation vector, respectively, in n spatial dimensions.

From relative to absolute transformation parameters. To properly align all particles without bias toward the final reconstruction, we need to estimate the absolute transformation parameters M_i for $i = 1,\ldots, N$. The consistency equation that relates the relative parameters to absolute parameters is

$$M_{ij} = M_j M_i^{-1}, \forall j > i \quad (4)$$

Direct linear numerical solution of equation (4) is difficult because the M_i contain a rotation matrix that is modulo 2π. To handle this problem, we use the (smooth) Lie-algebraic representation of the transformation[16,23], which solves the following optimization problem instead:

$$\underset{\{M_1,\ldots,M_N\}}{\operatorname{argmin}} \sum_{\substack{i,j \\ i>j}} \rho\left(d\left(M_{ij}, M_j M_i^{-1}\right)\right) \quad (5)$$

with distance function $d(X, Y) = ||\log(YX^{-1})||_F$, with F denoting the Frobenius norm (square root of the sum of absolute squares of the elements of the matrix) and $\rho(x) = |x|$ the L_1 loss function. The use of this norm makes the procedure robust to outliers[24]. This is important, as the registration can be trapped in a local minimum for nearly symmetric particles at 180° rotations. We solve the optimization problem in equation (5) by using an iterative gradient descent approach[16,25]. For our 2D geometry, we use as input a matrix with $4\times 4\times N(N-1)/2$ elements that holds the $N(N-1)/2$ stack of all the relative matrices M_{ij} of size 4×4 together with the indicator matrix of size $2\times N(N-1)/2$, which stores the correspondence indices of $N(N-1)/2$ particle pairs.

Removal of outlier registrations via self-consistency. After the first round of Lie-algebraic averaging, N sets of absolute transformation parameters are obtained. By plugging them into equation (4), one obtains $N(N-1)/2$ relative transformation parameters $\hat{M}_{ij}$, which can differ from the parameters M_{ij} estimated from the all-to-all registration. It appears that these differences are mostly in the rotation rather than the translation; thus, we do this consistency check on the basis of $\hat{R}_{ij}$ only (or equivalently in 2D the angle $\hat{\alpha}_{ij}$). Supplementary Fig. 9 shows the histograms of the quantity $|\hat{\alpha}_{ij} - \alpha_{ij}|$ for different DOL. Next to the correct pair registrations (the central peak) there are clearly outliers. Furthermore, the outlier fraction increases when the DOL decreases. Removal of these outlier registrations is therefore mandatory. We implemented this by excluding registration pairs with $|\hat{\alpha}_{ij} - \alpha_{ij}| > \varepsilon$, where ε is a threshold parameter, for a second round of Lie-algebraic averaging. Assuming that the all-to-all registration matrix A is a graph in which each matrix element A_{ij} is an edge that connects node (particle) i to node j, we can perform the optimization as long as this graph is connected, that is, for every pair there is at least one path connecting them. Intuitively, the outlier-removal step is equivalent to optimization of this graph for the most consistent path through all the nodes. For very low DOL, most of these paths are inconsistent. Therefore, a smaller threshold results in a better reconstruction as long as the number of remaining new relative parameters is greater than $N-1$. In this work, we set the threshold parameter $\varepsilon = 5°$ because that is the smallest angle that still keeps the graph of all-to-all registration for the worst dataset (30% DOL) connected. This step does not remove particles from the fusion; it only filters the redundant set of registration parameters. The large fraction of outlier pair registrations is the reason that the redundancy of the all-to-all registration is needed to achieve a robust fusion pipeline, and why a pyramid registration would lead to an inferior reconstruction (because of the error propagation from the randomly picked wrong pair registrations at the bottom layer of the pyramid). It is worth mentioning that this step is actually a simplified variant of the RANSAC-based motion averaging scheme as detailed in ref. [25]. We obtain, however, the absolute angles in step 2 by using all available relative angles instead of computing the initial absolute angles based on the minimum number of relative angles, that is, $N-1$. This is especially important for very low DOL datasets, as the result of averaging only $N-1$ random relative angles will be too corrupted to provide a good initial start for RANSAC.

Bootstrap registrations. The above steps are enough to provide a good reconstruction. It is, however, beneficial to use the outcome as a data-driven model/template to realign all individual particles to this model, that is, bootstrap the registrations[25]. The all-to-template registration, in this step, uses the registration of every single particle to a resampled version of the super-particle from step 3. The resampling is crucial, as the goal is to do the realignment on the basis of the most consistent localizations in the dense areas of the previous reconstruction. We randomly draw samples from the total set of localizations with a probability density function proportional to the density of localizations. We set the number of resampled localizations to 5,000 for the experimental data, as it represents about 2 particles for 80% DOL and 5–10 particles for 50% and 30% DOL, which already gives a good overall shape. This value must not be too large, as high values can lead to overfitting to noise (false positives in the data). For the 50% DOL dataset, we obtained a better reconstruction when we repeated the bootstrapping twice (Supplementary Fig. 10). It seems that after two iterations, the registration has converged. For 80% DOL, the reconstruction already converged at the first iteration, whereas for 30% DOL, the image got worse because of the two very dense areas. Finally, we registered each of the raw particles to the data-generated template using the routine as described in step 1. We benchmarked the performance of our proposed registration method on simulated 100% DOL data for which we had the ground-truth registration parameters. Supplementary Fig. 11 shows the histogram of the overall error between the final estimated angles and the ground truth from simulation. The histogram fits a normal distribution with an s.d. of 0.9°. Considering the dimensions of the origami, this error will result in a displacement of ±0.55 nm at the edges of the logo in the worst case, indeed smaller than the minimum binding-site distance and approximately equal to the average localization uncertainty. The corresponding reconstruction in Supplementary Fig. 2g also indicates a perfect match with the origami design.

Computational complexity. We achieved a complete all-to-all registration of 383 structures (resulting in ~73,000 pair registrations) of on average 2,060 localizations per origami (80% DOL) on four K40c Tesla cards on a 40 core server (Xeon E5-2670v3) in 2 h, and in 40 h on a cheap GPU in a regular desktop PC (Quadro K620, 2 GB RAM, 8 core Xeon E5-1660v3). These run times correspond to the most computationally expensive dataset. However, in practice, for highly labeled data, averaging a small subset of registrations can yield a fast but high-quality reconstruction (Supplementary Video 9). Supplementary Fig. 12 shows the reconstruction of 80% DOL data with only ~2% of the registration pairs with a computational time of ~10 min.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The software is available as Supplementary Software. Updated versions are free for download from https://github.com/imphys/smlm_datafusion2d. The computational functions on the GPU were implemented and optimized with Kernel Tuner (https://zenodo.org/record/1220114).

Data availability

Localization data are available at https://doi.org/10.4121/uuid:0d42a28f-f625-41a3-ba77-25e397685466.

References

21. Smith, C. S., Joseph, N., Rieger, B. & Lidke, K. A. *Nat. Methods* **7**, 373–375 (2010).
22. Richardson, S. & Green, P. J. *J. R. Stat. Soc. Series B Stat. Methodol.* **59**, 731–792 (1997).
23. Ye, K. & Lim, L.-H. *SIAM J. Appl. Algebra Geometry* **1**, 507–535 (2017).
24. Hartley, R., Aftab, K. & Trumpf, J. in *Proc. IEEE Conference on Computer Vision and Pattern Recognition 2011* 3041–3048 (IEEE, Piscataway, NJ, 2011).
25. Govindu, V. M. in *Computer Vision—ACCV 2006* (eds Narayanan, P. J., Nayar, S. K. & Shum, H. Y.) 457–466 (Springer, Berlin, 2006).

nature research

Corresponding author(s): Bernd Rieger

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	DNA-origami nanostructures were collected, analyzed and reconstructed using the Picasso software package which is publicly available at https://github.com/jungmannlab/picasso
Data analysis	The data was analyzed with MATLAB 2017a, CUDA 8.0 and EMAN2.12. The computational functions on the GPU have been implemented and optimized using Kernel Tuner which is available at doi:10.5281/zenodo.1220113. The code implementation (on CPU and GPU) is a major part of the research as it concerns an algorithm. The source code is available under our public Github repository at https://github.com/imphys/smlm_datafusion2d.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Localization data is available at https://doi.org/10.4121/uuid:0d42a28f-f625-41a3-ba77-25e397685466.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Not applicable.
Data exclusions	There was no exclusion of raw data. The images are being processed as described in detail in the online methods section.
Replication	All experiments on DNA-origami were consistent with previous work (ref. [12]).
Randomization	Randomization is not applicable.
Blinding	Blinding is not applicable.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	Unmodified, dye-labeled and biotinylated DNA oligonucleotides were purchased from MWG Eurofins. Streptavidin was purchased from Invitrogen (catalog number: S-888). BSA-Biotin was obtained from Sigma-Aldrich (catalog number: A8549).