Detecting structural heterogeneity in single-molecule localization microscopy data

Teun A.P.M. Huijben, Hamidreza Heydarian, Alexander Auer, Florian Schueder, Ralf Jungmann, Sjoerd Stallinga & Bernd Rieger

Particle fusion for single molecule localization microscopy improves signal-to-noise ratio and overcomes underlabeling, but ignores structural heterogeneity or conformational variability. We present a-priori knowledge-free unsupervised classification of structurally different particles employing the Bhattacharya cost function as dissimilarity metric. We achieve 96% classification accuracy on mixtures of up to four different DNA-origami structures, detect rare classes of origami occurring at 2% rate, and capture variation in ellipticity of nuclear pore complexes.
Single-molecule localization microscopy (SMLM) enables imaging below the diffraction limit. The image quality can be improved further by fusing hundreds of super-resolution images of identical bio-molecular structures, further referred to as particles, into a single reconstruction. This approach overcomes the problem of incomplete labeling using the central assumption that all particles represent the same underlying structure. In reality, however, the sample might be heterogeneous in structure due to the biology itself, sample preparation, diseases or drug-induced variations. These potential variations between structures blur standard fusion and small subsets of structurally different particles remain undetected. Sporadic 9-fold symmetric nuclear pores, for example, cannot be detected in the reconstruction.

Image classification is commonly used in single-particle averaging (SPA) for cryo-electron microscopy (cryo-EM) to find the viewing direction of each particle from its projection. Using EM classification techniques as such to SMLM data does not employ the full potential of SMLM data, due to the different image formation process, in particular, the incomplete labeling, and the use of localization coordinates instead of pixelated images. Previous work to separate classes in SMLM uses a deep neural network for classification. Here, however, the different classes need to be known a-priori and imaged in separate experiments to form learning sets in order to train the neural network. This strong a-priori knowledge is not compatible with discovering unknown data variation and is, therefore, inapplicable to most cellular imaging applications.

Here, we present an unsupervised classification tool to cluster 2D/3D SMLM data into (rare) structural subclasses without assuming any prior knowledge about the different classes. The method is capable of detecting multiple repeated structures in one field of view or capture structural variations, which is expected for many biological structures due to biological dynamics, deformation of the cellular structure, or phenotypical variations, within one class of particles. We successfully classified experimental DNA-PAINT and (d)STORM datasets, containing either multiple structures, rare subclasses or continuous structural variation. On experimental data, our approach shows 96% classification performance on multi-class DNA-origami data, detects rare classes of mirrored origami structures occurring at a 2% rate, identifies elliptical shape variation in nuclear pore complexes, and reveals height variation in 3D origami tetrahedron structures.

### Results

#### Clustering pipeline

Our approach uses pairwise registration of all particles to obtain a dissimilarity metric. Subsequently, a feature space is obtained by performing multidimensional scaling (MDS) on the dissimilarity representation which is followed by k-means clustering and particle fusion per cluster (Fig. 1).

In order to compare every pair of particles, we employ the Bhattacharyya cost function, which we used earlier in the all-to-all registration of template-free particle fusion (Methods). In contrast to this earlier use, we use the optimum value of this cost function as similarity metric, we are not interested in their relative translations and rotations. The Bhattacharyya metric works directly on the localization data, takes (possibly anisotropic) localization uncertainties into account, and is robust against underlabeling. The pairwise registration of N particles results in \( N(N-1)/2 \) similarity values. After converting the similarity to dissimilarity values, we use MDS to translate the pairwise dissimilarities into spatial coordinates of the particles in a multi-dimensional space. This constellation preserves the pairwise dissimilarities by minimizing the so-called metric stress loss function (Methods).

We used k-means clustering in the multidimensional scaling space in order to group the particles into clusters. K-means clustering has several advantages over other existing approaches that can be used on the MDS maps. First of all, it has the best performance on spherical clusters which is the dominant cluster shape appearing in our MDS space (Supplementary Fig. 1). Secondly, it is fast (compared to for example mean-shift) and has only one tuning parameter, \( K \). Mean-shift clustering, which also only has one free parameter, can potentially be used as an alternative to k-means, however, the bandwidth size selection itself is challenging and in our case does not have a physical meaning as the distance between the points in the MDS maps is based on the dissimilarity values. Other complex clustering approaches like DBSCAN or OPTICS have similar difficulties in tuning even more free parameters which are prohibitive for non-expert end-users.

In case of the absence of prior knowledge about the number of subclasses present in the data, we propose a two-step approach for determining the optimal number of clusters (\( K \)) for the k-means algorithm. Firstly, the user can inspect the silhouette cluster evaluation values (see Methods), for a range of \( K \) and choose for the one which gives the highest average score for all particles (Supplementary Fig. 2). Secondly, the user can inspect the scatter plot of the first three dimensions of the MDS space to determine the proper value of \( K \) (Supplementary Fig. 1). In certain situations, such as highly imbalanced subclasses, the silhouette graph would not have a predominant peak and visual inspection of the MDS maps may not be helpful to infer the number of underlying classes. Here, the general advice is to choose a large value of \( K \). This will result in many small classes which can subsequently be grouped further using the eigen image approach (Methods). Reconstruction per cluster is quick since the computationally most intensive procedure, the all-to-all registration, is already performed prior to the classification. This fast reconstruction makes it possible to investigate multiple values for the number of clusters, \( K \).

### Clustering of multi-class SMLM data imaged separately

We applied our classification algorithm to different multi-class
SMLM datasets in Figs. 2 and 3. We show the nanoTRON dataset\textsuperscript{15} (further referred to as digits), which consisted of four classes: the digits 1, 2, and 3, and a 3 × 4 grid (Fig. 2a). Each class comprises 2500 instances of each structure. The data were imaged separately per class. We randomly selected 5000 particles from all classes and fused them without classification. This resulted in a blurred reconstruction which resembled vaguely the digit 3 (Fig. 2b), caused by the different numbers of localizations per class (Supplementary Fig. 3a). Classification of the particles followed by fusion, however, resulted in four clearly distinct classes (Fig. 2c–f). The classification performance, which we define as the percentage of correctly classified particles, is 96.4%, and can be calculated since the ground truth class of each particle is known. The confusion matrix (Fig. 2g) gives further insight into the performance by indicating the false/true positives/negatives per class.

Clustering of multi-class SMLM data imaged together. To demonstrate that the classification is not selecting on different imaging conditions between the four classes, we repeated the same classification procedure on 5000 particles, 1250 per class, where all four designs were mixed in a single sample and acquired together. Here, the fusion without classification showed the digit 1, again explained by different numbers of localizations (Supplementary Fig. 3b). The fusions after classification in Fig. 2h–l show similar improvements as before, with an overall classification performance of 96.4% (Fig. 2m), using manual labeling\textsuperscript{15} as ground truth. The worse reconstruction of digit 1 (Fig. 2c) and digit 2 (Fig. 2j) compared to other classes, is due to lower quality of the individual particles in this experiment (Supplementary Fig. 4), not to suboptimal classification performance. The Fourier ring correlation (FRC) image resolution\textsuperscript{20} after reconstruction ranges from 3.7 nm to 5.7 nm per class (Supplementary Fig. 5).

Clustering of multi-class SMLM data wa...
scatter plots in Supplementary Fig. 1d–f and the silhouette plot in Supplementary Fig. 2c) resulted in clearly visible classes (Supplementary Fig. 6a–c) and a classification performance of 89% (Supplementary Fig. 6d). In contrast to the digits dataset, this dataset contained misfolded and unrecognizable particles (Supplementary Fig. 6h), because the simple design of only a few dots makes it hard for particle picking to exclude misfolded structures. Therefore, classification into four classes (Fig. 2p–s) was needed to capture the misfolded structures into a separate class (Fig. 2s). An average classification performance of 97.3 ± 1.9% was obtained for two independent repeats (Fig. 2t). The same three classes were visible after classification when all three designs were acquired from a single sample (Fig. 2u–w). Again, K = 3 (suggested by the analysis of MDS scatter plots and silhouette plots) was sufficient to correctly classify the structures (Supplementary Fig. 7a–c), but two additional misfold classes were needed to get correct reconstructions as the misfolds themselves vary (Supplementary Fig. 7). The digits dataset did not need a separate misfold class, as particle picking was successful in automatically removing all misfolded particles.

**Fig. 3 Classification of flipped DNA Origami data and NPC data.** a Fusion result without classification of 456 DNA-origami structures of the TUD-logo with 80% DoL. b, c The two classes resulting from the classification of a, containing 446 (normal orientation) and 10 (flipped orientation) particles per class, respectively. d–f Same as a–c, but with 50% DoL and 381 and 8 particles per class, respectively. g, h The two classes resulting from the classification of i, containing 168 and 136 particles, respectively. We used K = 4 for the 80% DoL data and K = 40 for the 50% DoL data, followed by further grouping with the eigen image approach with C = 2 (Methods). i Fusion result without classification of 304 NPCs. j–l Same as g, h, but with classification into three classes, with 133, 90 and 81 particles per class, respectively. For g–l, the ellipticity values e are defined as the ratio of the major axis over the minor axis of the ellipse fitted to the localizations, and are indicated below each class. Scale bar of a applies to b–f. Scale bar of g applies to h–l.
Detection of rare subclasses. In addition to datasets that contain multiple, equally abundant classes, we tested the classification algorithm on experimental data with skewed distributions of the number of particles per class. The TUD-logo DNA-origami data contains a small fraction that is unintentionally imaged upside-down, resulting in mirrored logos. Normally, these mirrored images are manually removed from data before fusion. Both for 80% and 50% density of labeling (DoL), the mirrored particles were not visible when all of them are fused (Fig. 3a, d), due to their low abundance. However, classification found these classes containing, respectively, 10 and 8 mirrored particles (Fig. 3c, f), from a total of 456 and 381 particles. We used $K = 4$ for the 80% DoL data (guided by visual inspection of the MDS scatter plot in Supplementary Fig. 1g–i) and $K = 40$ and for the 50% DoL data, both followed by further grouping (Methods) into two final classes. For the 50% DoL data, neither the MDS scatter plots nor the silhouette plots (Supplementary Fig. 2f) could help in determining the optimal value for $K$. Following our rule of thumb to use a high value and investigate the outcome for multiple values, $K = 40$ appeared to work best. To test how low we can go in the detection of rare events, we also applied our method to simulated nuclear pore complexes (NPC), with a rare 9-fold symmetric class. Classification could find these rare cases, even when only 2% was 9-fold symmetric (Supplementary Fig. 8). For rare class detection, it is necessary to have a high total number of particles as the rare class needs at least approximately 10 particles for successful detection.

Detection of structural variations in 2D SMLM data. Next to DNA-PAINT data, we applied our method to data acquired with (d)STORM of the integral membrane protein gp210, part of the NPC. Previously, Heydarian et al. used prior knowledge about the 8-fold symmetry for particle fusion and obtained a circular reconstruction with a homogeneous distribution of localizations. We applied our method to the same data and found classes with a range of ellipses (Fig. 3g–i). The classification separated the particles on their ellipticity (Supplementary Fig. 9c, d). The major axes of the ellipses align in the field-of-view (Supplementary Fig. 9a, b), suggesting that the deformations were caused either by the sample preparation or by the projection of the tilted rings onto the image plane. Since a tilt of 25° is needed to create the detected elliptical deformation, we think that the sample preparation is the more likely root cause for the found ellipticity. We further characterized and quantified the classification performance of our method on simulated data with discrete variation, TUD-logos with and without flame, and continuous variation, NPCs with a varying diameter (Supplementary Fig. 10). In the latter case, the classification perfectly separates the NPCs on diameter.

Detection of structural variations in 3D SMLM data. Finally, we applied our classification pipeline to PAINT images of a tetrahedron-shaped 3D DNA-Origami nanostructures (Methods) with an edge length of 100 nm (height ~90 nm). Our analysis shows that there is a variation in the height distribution of the tetrahedrons. This is consistent with the fact that the structural stability of large DNA origami nanostructures is limited. Before tetrahedron-shaped 3D DNA-Origami nanostructures (Methods) have an axial localization uncertainty of 8 nm. Classification of the tetrahedrons and within-class fusion of them results in a more isotropic resolution of the reconstructions (Supplementary Fig. 11b, c). Here, we identified different heights with a mean height of 55 and 95 nm or 45, 65, and 95 nm depending on the number of quantization levels of the continuous height distribution. The data suggest a continuous distribution of heights, which by definition does not provide the user with a preferred number of clusters, allowing the user to divide the distribution into a desired number of clusters. This analysis shows that the elongation of the binding sites along the z-axis in the reconstruction of all 218 tetrahedrons is due to the structural heterogeneity of the DNA-origami structures and not due to the poor imaging conditions.

Discussion

The developed classification pipeline (Fig. 1) is based on the dissimilarities between all the particles. The clustering is performed by using MDS as a spatial embedding of the particles, followed by k-means clustering. An alternative would be to directly cluster based on the dissimilarity values using hierarchical clustering (HAC). We have implemented and optimized both strategies and tested them on multiple datasets. The performance of the first method, i.e., clustering based on a spatial embedding, performs significantly better than hierarchical clustering (Supplementary Note 1, Supplementary Fig. 12). Therefore, the first method, is used for all classifications in this paper. Sabrina et al. recently proposed a similar HAC approach, however, with a different dissimilarity metric. Easy-to-classify data can be classified correctly by both approaches, although HAC does require the delicate choice of the right parameters, like the distance criterion and the pruning threshold. Other advantages of our approach are that there is no choice for the linkage criterion and that no particles are lost in pruning the dendrogram, because of the creation of single-particle clusters. We can conclude that our MDS approach performs significantly better than HAC and has fewer tuning parameters, which supports the decision to use the MDS approach in this paper.

In summary, we have developed an a-priori knowledge-free classification tool to identify (rare) structural subclasses in SMLM data. Once data heterogeneity is found, particle fusion algorithms can be employed. We successfully classified experimental DNA-PAINT and (d)STORM datasets, containing either multiple structures, rare subclasses or continuous structural variation.

Methods

DNA origami design and self-assembly of letters dataset. The DNA origami structures were designed with the software package Picasso using the Design module. The DNA origami were synthesized in a one-pot reaction with 50 μL total volume containing 10 nM scaffold strand, 100 nM core staples, 1 μM biotinylated staples and 1 μM DNA-PAINT docking sites staples. The DNA sequences are listed in Supplementary Tables 1–4. The folding buffer was 1× TE buffer with 12.5 mM MgCl₂. The DNA origami were annealed using a thermal ramp. First, incubating for 5 min at 80 °C, then going from 65 °C to 4 °C over the course of 3 h. DNA origami structures were purified via two rounds of PEG precipitation by adding the same volume of PEG-buffer, centrifuging at 14,000 × g at 4 °C for 30 min, removing the supernatant and resuspending in folding buffer. The 2× PEG buffer stock was 1× TE buffer with 500 mM NaCl and 15% PEG-8000 (v/v).

Optical setup for the acquisition of letters dataset. Fluorescence imaging was carried out on an inverted microscope (Nikon Instruments, Eclipse Ti2) with the Perfect Focus System, applying a custom-built objective-type TIRF configuration with an oil-immersion objective (Nikon Instruments, SR HP Apo TIRF 100Xc, numerical aperture 1.49, OIL). A 560 nm and a 642 nm laser (both MPB Communications Inc., 500 mW, DPSS-system) were used for excitation. The coaxially aligned laser beams were passed through a quad-band cleanup filter (Chroma Technology, ZET405/488/561/640xv2) and coupled into the microscope objective using a quad-band beam splitter (Chroma Technology, ZT405/488/561/640rpcv2). Fluorescence light was spectrally filtered with a quad-band emission filter (Chroma Technology, ZET405/488/561/640trf) and imaged on a sCMOS camera (Andor, Zyla 4.2 Plus) without further magnification, resulting in an effective pixel size of 130 nm (after 2 × 2 binning).
NAKANISHI et al.

DNA-PAINT sample preparation of letter dataset. For chamber preparation, a piece of coverslip (no. 1.5, 18 × 18 mm, ~0.17 mm thickness) and a glass slide (76 × 26 mm, 1 mm thick) were sandwiched together by two strips of double-sided tape to form a flow chamber with inner volume of ~20 μl. First, 20 μl of biotin-labeled boilium avinum (1 mg/ml, dissolved in buffer A) was flown into the chamber and incubated for 2 min. Then the chamber was washed using 40 μl of buffer A, second, 20 μl of streptavidin (0.5 mg/ml, dissolved in buffer B) was flown through the chamber and incubated for 2 min. Next, the chamber was washed with 20 μl of buffer A and subsequently with 20 μl of buffer B. Then ~500 pM of the DNA origami structures were flown into the chamber and allowed to attach to the surface for 2 min. Finally, the imaging buffer with dye-labeled imager strands was flown into the chamber and sealed with silicon. DNA-origami sequences are stated in Supplementary Table 5. Imaging conditions are listed in Supplementary Tables 6–9.

Two buffers were used for sample preparation and imaging: buffer A (10 mM Tris-HCl pH 7.5, 100 mM NaCl, 0.05% Tween 20, pH 7.5); buffer B (5 mM Tris-HCl pH 8, 10 mM MgCl2, 1 mM EDTA, 0.05% Tween 20, pH 8). Imaging Buffer B was supplemented with: 1× Trolox, 1× PCA, and 1× PCD. Trolox, PCA and PCD stocks were as follows: 100× Trolox: 100 μg Trolox, 430 μl 100% methanol, 345 μl 1 M NaOH in 3.2 ml H2O. 40× PCA: 154 mg PCA, 10 ml water and NaOH were mixed and the pH was adjusted to 9.0. 100× PCD: 93 mg PCD, 133 μl of buffer was used (100 mM Tris-HCl pH 8, 50 mM KCl, 1 mM EDTA, 50% glycerol).

DNA-origami design and self-assembly of tetrahedron dataset. The tetrahedron DNA-origami structures were formed in a one-pot reaction with a DNA origami total volume containing 10 mM sodium chloride (pH 8064), 100 mM core staples, 100 mM connector staples, 100 mM vector staples, 100 mM biotin handles, 100 mM DNA-PAINT docking site staples, and 1400 mM biotin anti-handles in folding buffer (1× TE 5 mM Tris, 1 mM EDTA) buffer with 10 mM MgCl2). The solution was annealed using a thermal ramp from 80 to 4 °C over the next 15 min. 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 MgCl2, 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 10000 x g at 4 °C.

Optical setup for the acquisition of tetrahedron dataset. Fluorescence imaging was carried out on an inverted microscope (Nikon Instruments, Eclipse Ti) with the Perfect Focus System, applying an objective-type TIRF configuration with a 3D oil-immersion objective (Nikon Instruments, Apo SR TIRF ×100, numerical aperture 1.49, Oil). A 561 nm (200 mW, Coherent Sapphire) laser was used for excitation. The laser beam was passed through cleanup filters (Chroma Technology, ZET561/10) and coupled into the microscope objective using a beam splitter (Chroma Technology, ZT561rxc). Fluorescence light was spectrally filtered with an emission filter (Chroma Technology, ET600/50m, and ET753ip) and imaged on a sCMOS camera (Andor, Zyla 4.2 Plus) without further magnification, resulting in an effective pixel size of 130 nm (after 2 x 2 binning).

DNA-PAINT sample preparation of tetrahedron structures. For chamber preparation, a piece of coverslip (no. 1.5, 18 × 18 mm, ~0.17 mm thickness) and a glass slide (3 x 1 inch 2 mm thick) were sandwiched together by two strips of double-sided tape to form a flow chamber with inner volume of ~20 μl. First, 20 μl of biotin-labeled boilium avinum (1 mg/ml, dissolved in buffer A) was flown into the chamber and incubated for 2 min. Then the chamber was washed using 40 μl of buffer A, second, 20 μl of streptavidin (0.5 mg/ml, dissolved in buffer B) was flown through the chamber and incubated for 2 min. Next, the chamber was washed with 20 μl of buffer A and subsequently with 20 μl of buffer B. Then ~500 pM of the DNA origami structures were flown into the chamber and allowed to attach to the surface for 2 min. Finally, the imaging buffer with dye-labeled imager strands was flown into the chamber and sealed with silicon. DNA-origami sequences are stated in Supplementary Table 5. Imaging conditions are listed in Supplementary Tables 6–9.

Two buffers were used for sample preparation and imaging: buffer A (10 mM Tris-HCl pH 7.5, 100 mM NaCl, 0.05% Tween 20, pH 7.5); buffer B (5 mM Tris-HCl pH 8, 10 mM MgCl2, 1 mM EDTA, 0.05% Tween 20, pH 8). Imaging Buffer B was supplemented with: 1× Trolox, 1× PCA and 1× PCD. Trolox, PCA and PCD stocks were as follows: 100× Trolox: 100 μg Trolox, 430 μl 100% methanol, 345 μl 1 M NaOH in 3.2 ml H2O. 40× PCA: 154 mg PCA, 10 ml water and NaOH were mixed and the pH was adjusted to 9.0. 100× PCD: 93 mg PCD, 133 μl of buffer was used (100 mM Tris-HCl pH 8, 50 mM KCl, 1 mM EDTA, 50% glycerol).

DNA-PAINT super-resolution microscopy of tetrahedron dataset. Images were acquired with an imager strand concentration of 3 nM (P1-Cy3B, 9nt) in imaging buffer. Here, 30,000 frames were acquired at 300 ms exposure time. The readout bandwidth was set to 200 MHz. Laser power (at 561 nm) was set to 110 mW (measured at the BFP of the objective). This power corresponds to an intensity of ~990 W cm-2 at the sample plane.

Super-resolution reconstruction. Raw fluorescence data were subjected to spot-finding and subsequent super-resolution reconstruction with the Picasso software package using the Localize module. The drift correction was performed with

Picosso Render using a redundant cross-correlation (RCC) with segmentation set to 1000 frames. The remaining drift was corrected with the Render’s function Undrift from picked with all picked DNA-origami structures. The DNA-origami were picked using Render Pick similar.

Materials. Unmodified DNA oligonucleotides, fluorescently modified DNA oligonucleotides, and biotinylated DNA oligonucleotides were purchased from MWG Eurofins. M13mp18 scaffold was obtained from Tülbom. BSA-Biotin was obtained from Sigma (cat: S-888). Tris 1 M pH 8.0 (cat: AM9856), EDTA 0.5 M pH 8.0 (cat: AM9261). Magnesium 1 M (cat: AM95350) and Sodium Chloride 5 M (cat: AM9759) were obtained from Ambion. Ultrapure water (cat: 10977-035) was purchased from Gibco. Polyethylene glycol (PEG)–8000 (catalog no. 6510-1KG) was purchased from Merck. Tween-20 (P9416-50ML), glycerol (65516-500 ml), methanol (32213-25L), protocatechuate 3,4-dioxynegasemonas (PCD; P8279), 3,4-dihydroxybenzoic acid (PCA; 37580-25G-F) and (±)-6-hydroxy-2,5,7,8-tetramethyl-chromene-2-carboxylic acid (trolox; 238813-5-G) were ordered from Sigma-Aldrich. Glass slides (cat: 48811-703) were obtained from VWR. Coverslips were purchased from Mardenfeld (cat: 0107032). Silicon (cat:1300 1000) was ordered from picoind. Double-sided tape (cat:6653) was ordered from 3M.
a way, that the Euclidean distances between pairs of particles are approximately equal to their dissimilarity measure. In this way, particles with a high dissimilarity will be placed far apart in the new space, and particles with a low dissimilarity will end up close to each other. Basically, this is the inverse process to determining distances. Instead of having a spatial embedding and measuring the distances between all pairs of particles, we start with the distances (dissimilarities) and try to find their spatial embedding. We use nonclassical MDS which iteratively updates the MDS coordinates by minimizing the metric stress loss function, defined as:

\[
\text{Stress} = \frac{1}{2} \sum_{i,j} \left( d_{ij} - |x_i - x_j| \right)^2, 
\]

where \( d_{ij} \) is the dissimilarity between particles \( i \) and \( j \), and \( x_i \) the MDS coordinates of particle \( i \). This stress loss function assumes that the dissimilarities are preserved as the inter-particle distances in the multidimensional space. We choose to embed the particles into 30 dimensions, since we empirically determined that any value between above 15 is sufficient (Supplementary Fig. 13) and we take as rule of thumb twice this value.

K-means clustering. K-means clustering in the multidimensional scaling space results in \( K \) clusters per particle. To prevent finding suboptimal clusters due to the random initialization of the k-means algorithm, the clustering procedure is repeated 1000 times. Using a high number of repeats is especially important in cases where a group containing a small number of particles has to be found. Here, we advise taking the number of repeats equal to the number of total particles, to assure that every possible cluster is on average at least once selected as initial seed for the k-means algorithm. From all initializations, the clustering with the lowest within-cluster sums of points-to-centroidal distances is chosen. To determine the value of \( K \), the scatter plot containing the first 3 dimensions of the MDS space can be inspected visually (Supplementary Fig. 1). Reviewing only the first three dimensions is sufficient, since they contain the most important variation within the MDS space and reviewing the order in a higher-dimensional space is visually challenging. Additionally, determining the optimal number of clusters can be guided by using the silhouette measure as cluster evaluation method (Supplementary Fig. 2). If the first dimensions show separated clusters, or the silhouette plot shows a clear peak, \( K \) can be chosen to match the number of clusters. However, when the variation between the different classes of particles is small, the MDS scatter plot will be a single point cloud, and does not give a clear indication for the choice of \( K \). In this case, as always, selecting the optimal \( K \) requires some tweaking. The user can start with an educated guess of \( K \) and then vary \( K \) to manually inspect what number of clusters is preferred. This can be done quickly as the computationally most time-consuming part of the method is the pairwise registration, which only has to be done once. The fusing of clusters is significantly faster and can, therefore, be repeated for multiple values of \( K \). If the goal is to find a small subgroup with an estimated occurrence of 2%, for example, it is advised to use a high value for \( K \) (as for the mirrored TUD-logo’s with 50% DoL, Fig. 3d-f) or when the variation is a continuous spectrum, a low value for \( K \) in the range of 2-3 will suffice (as for the elliptical NPCa, Fig. 3g-1 and 3D tetrahedrons Supplementary Fig. 11).

Particle fusion per cluster. Each of the clusters found in the previous step is reconstructed according to the particle fusion pipeline of Heydaran et al.2. The reconstruction of the grid structure in the digits dataset is done differently. The data consists of 12 different classes at 20 spatial resolutions at 4 x 4 pattern. Due to the symmetric nature of this structure in combination with an unbalanced number of localizations per binding site, reconstruction with the above-mentioned algorithm is suboptimal. Therefore, the localizations are structurally different particles, for example low-abundant mirrored DNA-origami and 3D tetrahedrons Supplementary Fig. 11).

The decomposition of \( \mathbf{X} \) results in:

\[
\mathbf{X}' \mathbf{u} = \mathbf{s} \mathbf{v},
\]

where \( \mathbf{u} \) and \( \mathbf{v} \) are the left- and right-singular vectors, respectively, with associated singular values. Since the matrix \( \mathbf{XX}' \) is normal (it commutes with its conjugate transpose), the left- and right-singular vectors are identical, and the SVD algorithm is effectively the same as eigen decomposition. The resulting singular vectors, \( \mathbf{u} \), represent the eigen images and the associated singular value, \( \mathbf{s} \), indicates how much variation of the data is explained by its respective singular vector. Instead of computing the SVD on the immense covariance matrix \( \mathbf{M} \text{(sizeN}^2 \times N^2) \), applying SVD on \( \mathbf{XX}' \text{(size}N^2 \times K^2) \) is computationally less expensive:

\[
\mathbf{X}' \mathbf{XX}' = \mathbf{s} \mathbf{a} \mathbf{u},
\]

After left-multiplying both sides of the equation with \( \mathbf{X}' \):}

\[
\mathbf{X}' \mathbf{XX}' \mathbf{X} = \mathbf{s} \mathbf{a} \mathbf{u} \mathbf{X},
\]

we see this resembles the decomposition of \( \mathbf{XX}' \) as shown before (Eq. 4) and the sought-after singular vector, \( \mathbf{u} \), can be calculated as \( \mathbf{u} = \mathbf{Xa} \). By reshaping and normalizing the singular vectors, the eigen images of size \( N \times N \) are formed. Since the eigenvectors are sorted based on their singular values, the first one represents most of the variation between the different images. All \( K \) images are projected onto the first eigen image and the resulting weights are clustered into \( K \) classes by hierarchical agglomerative clustering with an average-linkage criterion25. The value of \( K \) represents the final number of classes of the classification. For the low-abundant mirrored DNA-origami structures (Fig. 3a-f), as well as for the 9-fold symmetric nuclear pores (Supplementary Fig. 8) we used \( K = 2 \). The localizations of the \( K \) clusters are fused per class (using the particle fusion pipeline of Heydaran et al.) to form the final \( C \) output classes.

Silhouette cluster evaluation. The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette value ranges from \(-1 \) to \( 1 \), where a high value indicates a big separation between the clusters mutually. The silhouette value for a point \( i \) is defined as \( \text{bi} - \text{ai} / \max(\text{ai}, \text{bi}) \), where \( ai \) is the average distance of this point to all other points in the same cluster, and \( bi \) the minimum average distance to points in a different cluster, minimized over the other clusters. The silhouette values shown in Supplementary Fig. 2 are based on the first three dimensions of the MDS space and represent the average silhouette value over all particles.

Fitting ellipses to nuclear pore data. The ellipticity is defined as the ratio of the length of the major axis over the length of the minor axis of the ellipse that is fitted to the data. To make the elliptical fit less sensitive to the imbalanced distribution of localizations over the 8 binding sites, we fit the ellipse to the medians of the 8 blobs. After grouping the localizations into 8 clusters by k-means clustering, the median position if calculated for all the localizations of each cluster and fitted with an ellipse.

Order parameter. The order parameter is calculated for the orientation of the elliptical fits of the experimental nuclear pore data (Supplementary Fig. 9) and is defined as \( \text{cos(2} \theta - \theta_0) \), where \( \theta \) represents the orientation of the ellipse, \( \theta_0 \) the director which is the average angle of all ellipses, and \( \theta_2 \) the population average over all ellipses of the respective class. By definition, a value for the order parameter of \( 0 \) indicates complete alignment of all ellipses and a value of \( 1 \) indicates a homogeneous random orientation.

Data availability

Single molecule localization data for the Digits and Letters DNA-origami nanostructures are publically available via 4TU.Research dataset repository26 (https://doi.org/10.4121/14074919.v1). Source data are provided with this paper.

Code availability

The software is available for download under the terms of the Apache2.0 license from Github and 4TU Research Data repository27 (https://github.com/impwys/smlm_classification2d (https://doi.org/10.4121/14135849.v1).
Received: 18 November 2020; Accepted: 2 June 2021; Published online: 18 June 2021

References
1. Klein, T., Proppert, S. & Sauer, M. Eight years of single-molecule localization microscopy. Histochem. Cell Biol. 141, 561–575 (2014).
2. Szymborska, A. et al. J. Nuclear pore scaffold structure analyzed by super-resolution microscopy and particle averaging. Science 341, 655–658 (2013).
3. Löschberger, A. et al. Super-resolution imaging visualizes the eightfold rotational symmetry of gp210 proteins around the nuclear pore complex and resolves the central channel with nanometer resolution. J. Cell Sci. 125, 570–575 (2012).
4. Heydarian, H. et al. Template-free 2D particle fusion in localization microscopy. Nat. Methods 15, 781–784 (2018).
5. Jimenez, A., Friedl, K. & Leterrier, C. About samples, giving examples: optimized single molecule localization microscopy. Methods 174, 100–114 (2020).
6. Hinshaw, J. E. & Milligan, R. A. Nuclear pore complexes exceeding eightfold rotational symmetry. J. Struct. Biol. 141, 259–268 (2003).
7. Burgert, A., Letschert, S., Doose, S. & Manley, S. Multicolor single-particle reconstruction of protein complexes. Nat. Methods 15, 777–780 (2018).
8. Huang, Z., & Zheng, L. Correlative single-molecule localization microscopy. Methods 174, 100–114 (2020).
9. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD’96). 226–231 (AAAI Press, 1996).
10. Sokal, R. R., Michener C. D. A Statistical Method for Evaluating Systematic Relationships (University of Kansas, USA, 1958).
11. Huijbens, T. A. P. M., Heydarian, H., Rieger, B. & Stallinga, S. Software for detecting structural heterogeneity in single-molecule localization microscopy. Bioinformatics 36, 3620–3622 (2020).
12. Mead, A. Review of the development of multidimensional scaling methods. J. R. Stat. Soc. 41, 27–39 (1992).
13. Jain, A. K., Murty, M. N. & Flynn, P. J. Data clustering: a review. ACM Comput. Surv. 31, 264–323 (1999).
14. Han, J., Kamber, M., & Pei, J. Data Mining: Concepts and Techniques (Waltham, MA: Morgan Kaufman Publishers, 10, 978-1, 2012).
15. Kaufman, L. & Roabbeew, P. J. Finding Groups in Data: An Introduction to Cluster Analysis (Wiley-Interscience, Hoboken, NJ, 2005).
16. Nieuwenhuizen, R. P. et al. Measuring image resolution in optical nanoscopy. Nat. Methods 15, 557–562 (2013).
17. Sablinina, V., & et al. 3D super-resolution fluorescence microscopy maps the variable molecular architecture of the Nuclear Pore Complex. Preprint at https://www.biorxiv.org/content/10.1101/2020.11.27.386599v1 (2020).
18. Schnitzbauer, J., Strauss, M. T., Schlichteaele, T., Schueder, F. & Jungmann, R. Super-resolution microscopy with DNA-PAINT. Nat. Protoc. 12, 1198 (2017).
19. Jian, B. & Vemuri, B. C. Robust point set registration using gaussian mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 33, 1633–1645 (2010).
20. Stallinga, S., Burgert, A., Letschert, S., Doose, S. & Manley, S. Multicolor single-particle reconstruction of protein complexes. Nat. Methods 15, 777–780 (2018).
21. Auer, A., Strauss, M. T., Strauss, S. & Jungmann, R. nanoTRON: a Picasso module for MLP-based classification of super-resolution data. Bioinformatics 36, 3620–3622 (2020).
22. Mead, A. Review of the development of multidimensional scaling methods. J. R. Stat. Soc. 41, 27–39 (1992).
23. Jain, A. K., Murty, M. N. & Flynn, P. J. Data clustering: a review. ACM Comput. Surv. 31, 264–323 (1999).
24. Han, J., Kamber, M., & Pei, J. Data Mining: Concepts and Techniques (Waltham, MA: Morgan Kaufman Publishers, 10, 978-1, 2012).
25. Kaufman, L. & Rouxbeew, P. J. Finding Groups in Data: An Introduction to Cluster Analysis (Wiley-Interscience, Hoboken, NJ, 2005).
26. Nieuwenhuizen, R. P. et al. Measuring image resolution in optical nanoscopy. Nat. Methods 15, 557–562 (2013).
27. Sablinina, V., & et al. 3D super-resolution fluorescence microscopy maps the variable molecular architecture of the Nuclear Pore Complex. Preprint at https://www.biorxiv.org/content/10.1101/2020.11.27.386599v1 (2020).
28. Schnitzbauer, J., Strauss, M. T., Schlichteaele, T., Schueder, F. & Jungmann, R. Super-resolution microscopy with DNA-PAINT. Nat. Protoc. 12, 1198 (2017).
29. Jian, B. & Vemuri, B. C. Robust point set registration using gaussian mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 33, 1633–1645 (2010).