





## TOOLS

# Parameter-free molecular super-structures quantification in single-molecule localization microscopy

Mattia Marena<sup>1,2</sup> , Elena Lazarova<sup>1</sup> , Sebastian van de Linde<sup>3</sup>, Nick Gilbert<sup>1</sup> , and Davide Michieletto<sup>1,2</sup> 

Understanding biological function requires the identification and characterization of complex patterns of molecules. Single-molecule localization microscopy (SMLM) can quantitatively measure molecular components and interactions at resolutions far beyond the diffraction limit, but this information is only useful if these patterns can be quantified and interpreted. We provide a new approach for the analysis of SMLM data that develops the concept of structures and super-structures formed by interconnected elements, such as smaller protein clusters. Using a formal framework and a parameter-free algorithm, (super-)structures formed from smaller components are found to be abundant in classes of nuclear proteins, such as heterogeneous nuclear ribonucleoprotein particles (hnRNPs), but are absent from ceramides located in the plasma membrane. We suggest that mesoscopic structures formed by interconnected protein clusters are common within the nucleus and have an important role in the organization and function of the genome. Our algorithm, SuperStructure, can be used to analyze and explore complex SMLM data and extract functionally relevant information.

## Introduction

Single-molecule localization microscopy (SMLM; van de Linde et al., 2011; Schermelleh et al., 2010; Henriques et al., 2011; Sauer and Heilemann, 2017) is now commonly employed for quantitative analysis of molecular structures and interactions in both cell-based (Cisse et al., 2013; Kapanidis et al., 2018; Chong et al., 2018) and in vitro experiments (Revyakin et al., 2006; Deniz et al., 2008). Unlike other light microscopy techniques, SMLM achieves resolutions far beyond the diffraction limit, and its typical output is a list of 3D coordinates (or localization events) that are naturally analyzed using efficient clustering algorithms borrowed from quantitative big-data analysis and even astronomy (Owen et al., 2010; Sengupta et al., 2011; Garcia-Parajo et al., 2014; Baumgart et al., 2016; Spahn et al., 2016; Griffié et al., 2016). However, traditional clustering algorithms rely on user-defined parameters that are intrinsically intertwined with the notion of similarity that is necessary to define a cluster. These parameters can be either hypothesized by physical intuition or inferred via preemptive analysis (Burgert et al., 2017; Williamson et al., 2020; Malkusch and Heilemann, 2016), yet their choice has a significant impact on the results, in turn hindering the portability of clustering algorithms and the comparison between different datasets.

At the same time, recent evidence suggest that assemblies of proteins (Brangwynne et al., 2015; Larson et al., 2017; Strom et al., 2017; Sabari et al., 2018; Cho et al., 2018; Maharana et al., 2018; Chong et al., 2018) and chromatin (Bintu et al., 2018; Boettiger et al., 2016; Frank and Rippe, 2020) form functional complex structures that are not fully captured by standard clustering algorithms. For example, the heterogeneous nuclear ribonucleoprotein U (hnRNP-U), also called scaffold attachment factor A (SAF-A), is suggested to form a dynamic and functional mesh-like structure while interacting with RNA to maintain transcriptionally active genomic loci in a decompacted configuration (Nozawa et al., 2017; Michieletto and Gilbert, 2019). Other examples include SC35, a nuclear protein involved in RNA splicing and chromatin elongation (Lin et al., 2008) that displays localized nuclear speckles (Xie et al., 2006; Jackson et al., 2000), or actin and microtubules, which form elongated and interconnected networks involved in cell motility and division, as well as in the synaptic plasticity of dendritic spines (Resch et al., 2002; Rogers et al., 2003; Izeddin et al., 2011). Additionally, recent super-resolution studies indicate that chromatin is also functionally organized in connected nano-scale compartments (Prakash et al., 2015; Szabo et al., 2018; Nir et al., 2018; Maiser

<sup>1</sup>Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK; <sup>2</sup>Scottish Universities Physics Alliance, School of Physics and Astronomy, University of Edinburgh, Edinburgh, UK; <sup>3</sup>Scottish Universities Physics Alliance, Department of Physics, University of Strathclyde, Glasgow, UK.

Correspondence to Mattia Marena: [mattia.marena@igmm.ed.ac.uk](mailto:mattia.marena@igmm.ed.ac.uk); Nick Gilbert: [nick.gilbert@ed.ac.uk](mailto:nick.gilbert@ed.ac.uk).

© 2021 Marena et al. This article is available under a Creative Commons License (Attribution 4.0 International, as described at <https://creativecommons.org/licenses/by/4.0/>).

et al., 2020). Rapidly evolving methods of chromatin tracing (Boettiger et al., 2016; Wang et al., 2016; Beliveau et al., 2015; Nir et al., 2018; Bintu et al., 2018) and super-resolved imaging of the accessible genome (Xie et al., 2020) require sophisticated algorithms to analyze the topology of the generated paths (Goundaroulis et al., 2020). To understand the relationship between these complex structures and the underlying biological mechanism and functions of the genome (Bronstein et al., 2015; Khanna et al., 2019; Leidescher et al., 2020 Preprint; Smeets et al., 2014), a more sophisticated and standardized analysis of SMLM data is urgently required.

It is clear that quantification of complex structures is a ubiquitous problem in molecular and cell biology, and it is intimately connected to cellular function. Motivated by this problem, here, we introduce a new algorithm termed SuperStructure, which extends in a novel and original way the popular density-based clustering algorithm DBSCAN. SuperStructure allows (1) a parameter-free detection and quantification of complex structures made of connected clusters in SMLM data and (2) a parameter-free quantification of the density of molecules within clusters.

Here, we demonstrate the capabilities of SuperStructure on simulated datasets and then use it to analyze two groups of experimental datasets: (1) nuclear proteins involved in RNA processing, namely SAF-A, hnRNP-C, and SC35; and (2) ceramide lipids involved in cellular trafficking at the membrane. We find that interconnections between clusters are abundant in classes of proteins in the hnRNP family and that they are surprisingly absent from ceramides, suggesting this feature is relevant for the biological function of SAF-A and hnRNP-C. Therefore, SuperStructure enables us to discover new facets of protein organization in human cells and provides a better understanding of the molecular mechanisms underlying the organization of subcellular (super-)structures.

Finally, since SuperStructure is parameter-free, it provides the community with a standardized tool for the discovery and quantification of complex patterns in SMLM data. Furthermore, beyond helping our understanding of complex biological structures, it might be used to assess the fluorophore blinking quality and thus offers versatility in assessing also technical imaging properties (van de Linde and Sauer, 2014; Hennig et al., 2015; Sieberg and Herten, 2011).

## Results

### SuperStructure algorithm

SuperStructure is best explained in relation to the well-known DBSCAN algorithm. DBSCAN detects clusters by grouping together high-density localizations and classifies as outliers low-density ones (Ester et al., 1996). In practice, DBSCAN determines that a localization is part of a cluster if more than  $N_{min}$  other localizations are found within a neighborhood distance  $\epsilon$  (or if it is part of the neighborhood of another localization with this property). Conversely, SuperStructure extracts connectivity information from the rate at which the number of detected clusters  $N_c$  changes with the neighborhood radius  $\epsilon$  for a fixed  $N_{min}$  (see Fig. 1). Indeed, the curves  $N_c(\epsilon)$  contain important

overlooked information about the structure of connections. To simplify the analysis, and without loss of generality, we set  $N_{min} = 0$ , which means that we do not require a minimum number of localizations within the neighborhood to define a cluster. As a consequence,  $N_c(\epsilon)$  is necessarily a monotonically decreasing function, as for  $\epsilon = 0$ , every localization is detected as a single cluster and increasing  $\epsilon$  yields fewer but larger clusters. Following on, the rate at which  $N_c$  decays with  $\epsilon$  is an indicator of how quickly localizations, and then clusters of localizations, coalesce, thus indicating how much localizations and clusters are connected.

The  $N_c(\epsilon)$  curves provided by SuperStructure identify different clustering regimes (Fig. 1). The first (small  $\epsilon$ ) regime describes the merging of localizations within clusters (intra-cluster regime), the second (intermediate  $\epsilon$ ) regime captures the growth of clusters into super-structures (first super-cluster regime), and the third (large  $\epsilon$ ) regime describes the merging of super-clusters into higher-order super-structures (second/third super-cluster regimes). The  $N_c(\epsilon)$  curve in the first regime typically follows a Poissonian function (Eq. 1), and its decay rate is related to the density of emitters  $\rho_{em}$  within the clusters (see Materials and methods and Figs. 1 and S1). The width of the Poisson function also sets the critical value of  $\epsilon$  at which this first regime is expected to end (Eq. 2). On the other hand, the decay in the second and third regimes follows an exponential decay with characteristic length-scale  $\lambda$  and are highly dependent on the connectivity between (super-)clusters, as well as on the density of (super-)clusters (Eq. 4).

The number of super-cluster regimes depends on the homogeneity of both cluster distribution and connections. In the two extreme cases of a completely connected or unconnected homogeneous distribution of clusters, we expect a single super-cluster regime. However, while in the former case this regime is exponential (because the clusters are connected), in the latter it assumes a Poissonian functional form (see respectively Eqs. 4 and 3). This is not surprising, as free (unconnected) clusters that are randomly distributed behave (on a larger scale) as single emitters inside clusters (see Materials and methods and Fig. S1). Also, in the case of clusters embedded in a random distribution of other localizations (such as noise), we obtain a Poissonian decay. Importantly, a random distribution of localizations (also at high density) is different from “connected” clusters, where nearby localizations are mostly distributed in between clusters. As a result, the curves generated by SuperStructure allow us to identify the presence/absence of connectivity by investigating the functional form of the curves, as well as to extract their decay rates.

In heterogeneous systems that display a mix of randomly dispersed localizations/clusters and connected ones over similar length-scales, we strongly recommend restricting the analysis with regions of interest (ROIs) over subregions that display qualitatively similar phenotypes. A good example of heterogeneous system is given by the nuclear protein SC35, which we analyze below. Restricting the analysis to ROIs is also recommended when quantifying nuclear or cellular substructures that display boundaries. Masking localizations falling outside these boundaries allows SuperStructure to generate cleaner curves that are easier to interpret.

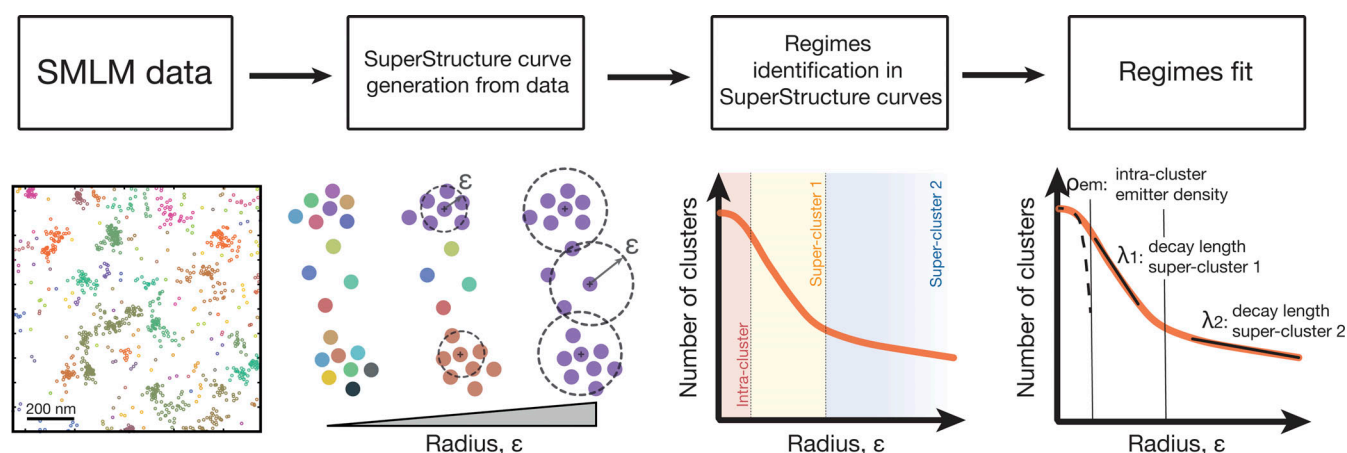


Figure 1. **Working principle of SuperStructure analysis.** Left: SMLM data are taken as input for the analysis. Center left: Cluster analysis is run using the DBSCAN algorithm with  $N_{min} = 0$  and  $\epsilon$  progressively increasing in an adequate range for the system. SuperStructure curves describing the number of detected clusters  $N_c$  as a function of  $\epsilon$  are generated. Center right: SuperStructure curves are plotted and inspected to identify super-cluster regimes representing the onset of connected structures. Right: Intra- and super-cluster regimes are fitted with our models (see Materials and methods) to quantify the emitter density inside clusters  $\rho_{em}$  and the connectivity among clusters (via the decay length  $\lambda_i$  for super-cluster regime  $i$ ).

To quantify the intra-cluster density and (super-)cluster connectivities, one needs to define boundaries between regimes and to fit every regime with the corresponding function (see Eqs. 1, 3, and 4). Regime boundaries and fitting ranges can be selected either manually (where curves change their decay properties) or by rigorously running a preemptive goodness-of-fit test. For instance, once the rough regime range has been identified and fitted, one can modify the fit window to identify the boundaries of the regime outside which the fit is no longer acceptable. Arguably, the optimum regime is found by identifying the best goodness-of-fit window (e.g., the range with the minimum  $\chi^2$ ). It is also possible to define a single function fitting the entire curve by (1) defining a piecewise function where every “piece” is the fit of the corresponding regime or (2) adding together the contribution of the different regimes (appropriately weighted).

The workflow for the application of SuperStructure is shown in Fig. 1 and is described in detail in Materials and methods. Additionally, the codes and scripts are open source and available at git repository (see below).

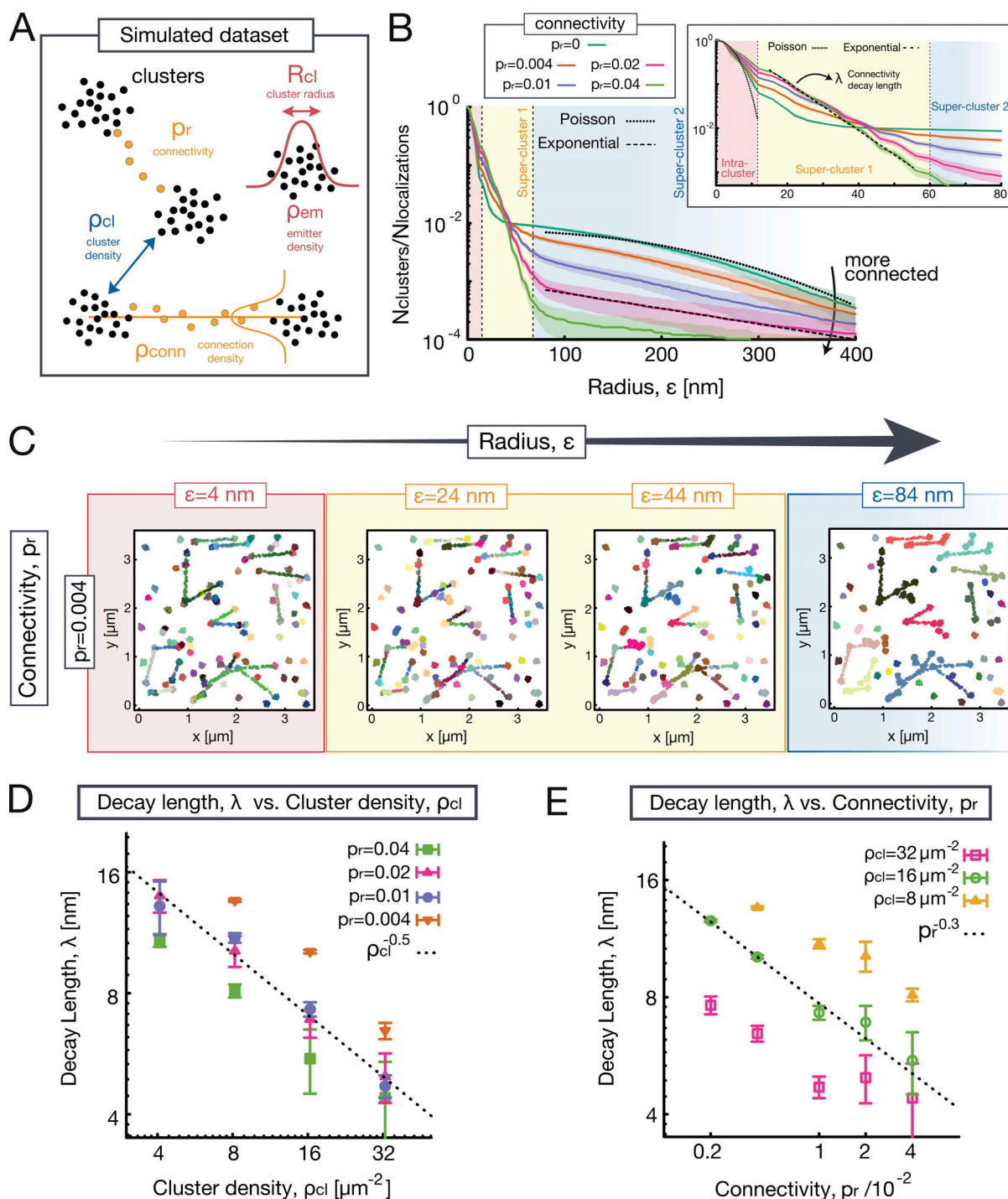
### Characterizing SuperStructure feature extraction from simulated SMLM data

To evaluate the performance of SuperStructure, we analyzed artificial datasets consisting of interconnected clusters of localizations on a 2D plane (see Fig. 2 A). Clusters are homogeneously and randomly positioned on the plane with a cluster density  $\rho_d = 8.2 \mu m^{-2}$  that is comparable to that of some nuclear proteins (see below). Every cluster has average radius  $R_d \sim 40 nm$  and an overall internal localization density  $\rho_{em} = N_{em}/\pi R_d^2 = 16,000 \mu m^{-2}$ , where  $N_{em}$  is the number of localizations per cluster. Pairs of clusters are connected with probability  $p_r$  by a sparse point distribution and only if the distance between the clusters is less than  $b = 1 \mu m$ . These choices allow us to readily tune the degree of “connectivity” in the system by varying a single parameter  $p_r$ . A second parameter,  $p_{r,conn}$ , is introduced to control the density of localizations within the connections  $\rho_{conn}$  (see Materials and methods for details).

The length-scales associated to density of emitters inside clusters  $\rho_{em}$  and inside connections  $\rho_{conn}$  define the boundaries among the three regimes of  $N_c(\epsilon)$  (Fig. 2 B): (1) for  $\epsilon \leq 12 nm$ , the intra-cluster regime follows a Poissonian decay (Eq. 1) with density parameter  $\rho_{em} = 16,000 \mu m^{-2}$  (as expected, since it was set by construction); (2) for intermediate values of  $\epsilon$ , the exponential super-cluster regime dominates (Eq. 4), and the fusion of connected clusters takes place (see inset of Fig. 2 B); (3) for  $\epsilon \geq 60 nm$ , we expect to observe the coalescence of super- and nonconnected clusters in a second super-cluster regime; this is captured by a second exponential for  $p_r \neq 0$  (Eq. 4). Conversely, for  $p_r = 0$ , we observe a single super-cluster regime that is well fitted by a Poissonian function with lower density (Eq. 3), as it corresponds to the density of clusters rather than emitters within clusters (see dark green curve in Fig. 2 B).

Examination of Fig. 2 B (inset) highlights the exponential behavior of the super-cluster regime (2) for different values of connectivity  $p_r$ . Importantly, a larger  $p_r$  results in an effectively shorter decay length (or larger spatial rate of merging) for the regime in which clusters merge into super-clusters. This strongly suggests that the effective decay length (or rate) mirrors the connectedness of the underlying super-structures (Fig. 2 C). In fact, these simulations reveal that the decay length represents the combined contribution of cluster density  $\rho_d$  and connectivity  $p_r$ . A larger density of clusters can impact the decay length as much as a larger connectivity, as shown by simulations at fixed  $p_r$  and different  $\rho_d$  (Fig. 2 D; and Fig. S2, A and B). In particular, we find that the functional form of the decay length is  $\lambda \sim \rho_d^{-1/2} p_r^{-0.3}$  (Fig. 2, D and E). The cluster density contribution is  $\sim \rho_d^{-1/2}$ , as it depends on the typical distance between clusters and is relevant when comparing datasets with different cluster density. By combining SuperStructure with a cluster analysis, one can estimate  $\rho_d$  and normalize  $\lambda$  to obtain the pure connectivity contribution in the decay length:  $\lambda^* = \lambda/\rho_d^{-1/2}$ .

Finally, in order to characterize the contribution to the  $N_c(\epsilon)$  curves coming from the density of localizations within the



**Figure 2. Evaluating SuperStructure on simulated datasets.** (A) Sketch representing the artificial dataset consisting of interconnected clusters of localizations on a 2D plane. Clusters are characterized by an internal density of localizations  $\rho_{em}$  and radius  $R_d$  and are randomly distributed on the plane at an average cluster density  $\rho_{cl}$ . Clusters can be connected by a sparse point distribution with probability  $p_r$ , and connections have a density of points  $\rho_{conn}$  (controlled by the  $p_{r,conn}$  parameter). (B) Average SuperStructure curves (zoomed in the inset) for simulated datasets with different connectivity  $p_r$ . Other parameters are kept fixed: average cluster radius  $R_d \approx 40$  nm, emitter density within clusters  $\rho_{em} = 16,000 \mu m^{-2}$ , cluster density  $\rho_{cl} = 8.2 \mu m^{-2}$ , and  $p_{r,conn} = 0.5$  (which fixes the density of emitters within connections  $\rho_{conn}$ ). The curves show the number of detected clusters normalized by the total number of localizations. Curves are the average of 20 independent simulated datasets. Shaded regions represent the standard deviation from the average. Three regimes can be distinguished: (1) the intra-cluster (red), (2) the first super-cluster (yellow), and (3) the second super-cluster (blue). The decay in the intra-cluster regime corresponds to a Poisson avoidance function with density parameter  $\rho_{em} = 16,000 \mu m^{-2}$  (Eq. 1, dotted line in the inset). The first super-cluster regime can be fitted by a single exponential (Eq. 4, dashed line in the inset) which returns an effective decay length  $\lambda$ . The second super-cluster regime can be fitted with another exponential if  $p_r \neq 0$  (Eq. 4, dashed line in the main figure). In case of  $p_r = 0$ , there is only one super-cluster regime, and it follows a Poisson function with density parameter  $\rho_{cl} = 8.2 \mu m^{-2}$  (Eq. 3, dotted line in the main figure). (C) Snapshots of detected clusters for an artificial dataset with connectivity  $p_r = 0.004$  and by progressively increasing the value of the radius  $\epsilon = 4, 24, 44, 84$  nm. (D) Decay length  $\lambda$  versus cluster density  $\rho_{cl}$  scales as  $\rho_{cl}^{-0.5}$  for any value of connectivity  $p_r$ . (E) Decay length  $\lambda$  versus connectivity  $p_r$  scales as  $p_r^{-0.3}$  for different values of  $\rho_{cl}$ . In D and E, 20 independent datasets were fitted with Eq. 4, and the resulting  $\lambda$  values were averaged. Vertical bars represent the standard deviation from the average.



connections, we further simulated SMLM datasets with a fixed, large connectivity  $p_r$  and varied the density of points in the connections by tuning  $p_{conn}$  (see simulated datasets in Figs. 2 A and S2 F). As expected, we observe a single super-cluster regime, and the denser the connections, the shorter the decay length. This indicates that our algorithm is able to describe not only how well clusters are connected (i.e., the number of connections per cluster) but also how strongly they are connected (i.e., how dense the connections are). These features are likely to be highly relevant for nuclear proteins.

Before applying this methodology to experimental data, we also tested the effect of random noise in the system (i.e., unconnected isolated localizations from biological or technical sources). We observed that in presence of random noise the decay of SuperStructure curves becomes Poissonian for large  $\epsilon$  (see Fig. S2 C) with an effective density  $\rho$  larger than the cluster density (see Fig. S2 D). Decay lengths in the first super-cluster regime (yellow regime) are still distinguishable even in presence of noise at reasonable density (albeit smaller than the connection density), but their absolute values are altered, with weakly connected systems more severely affected (see Fig. S2 E). These observations suggest that, as in most analysis algorithms, large noise might obscure exponential decays of connected systems. In case a single Poissonian behavior or a combination of exponential and Poissonian decay is found in the SMLM dataset, it is therefore important to combine SuperStructure with an independent cluster analysis at different length scales (e.g., at three or four selected values of  $\epsilon$ ) and a direct observation of the dataset in order to exclude the presence of hidden connectivity.

### Quantification of super-structures in nuclear proteins

We now examine biological data and apply SuperStructure to dSTORM data acquired for three different nuclear proteins (Fig. 3, A and B): the serine/arginine-rich splicing factor SC35, hnRNP-C, and hnRNP-U (also known as SAF-A). These proteins are abundantly expressed in the nucleus of human cells and are involved with RNA processing at different stages. SC35 is necessary for RNA splicing, while hnRNPs are implicated not only in the regulation and maturation of mRNA but also in chromatin structure (Nozawa et al., 2017; Xiao et al., 2012; Caudron-Herger et al., 2011). In particular, SAF-A is thought to form a dynamic homogeneous mesh that regulates large-scale chromatin organization by keeping gene-rich loci in a decompacted state (Nozawa et al., 2017; Michieletto and Gilbert, 2019). Hence, capturing the organization of this protein beyond the traditional single-cluster analysis is an important step toward understanding how it regulates chromatin structure in different cell stages and conditions.

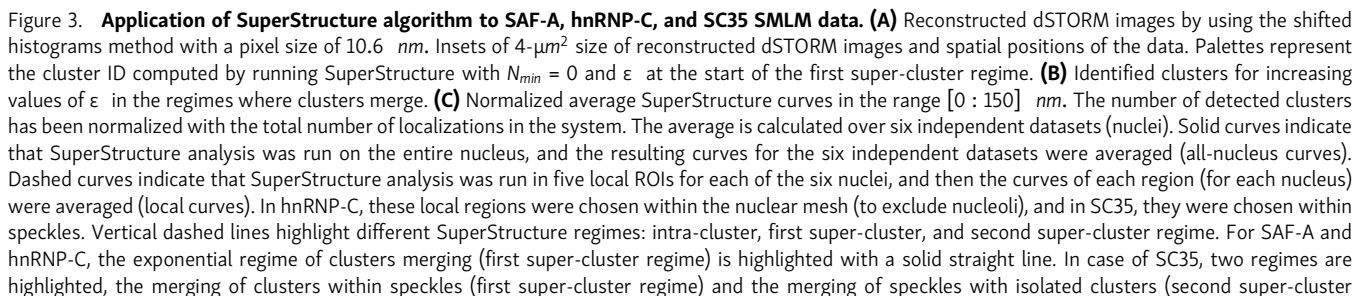
Curves obtained from SuperStructure analysis after masking signal in the nuclear region are shown in Fig. 3 C, where we highlighted the super-cluster regimes discussed above. Global nuclear analysis is represented by filled curves, while analysis on localized ROIs is represented by dashed ones (hnRNP-C nuclear mesh and SC35 speckles). Both hnRNPs display a first super-cluster regime for which the curves decay as exponentials, suggesting that within this range, distinct clusters are in reality connected. Interestingly, while SAF-A displays a unique long

super-cluster regime, hnRNP-C seems to also show a second exponential regime (filled curve). However, this regime appears at very large values of  $\epsilon$  and is due to sparse clusters of localizations in nucleoli. Running SuperStructure on ROIs with nucleoli masked out (dashed line) indeed generates a single exponential function, confirming that hnRNP-C clusters are fully connected. We can therefore conclude that both hnRNPs exhibit a single exponential regime, typical of fully connected meshes. On the other hand, SC35 displays exponentials with different characteristic decay rates in two distinct and significant super-cluster regimes (filled curve), one for intermediate  $\epsilon \in [10, 20]$  nm, when clusters inside speckles merge (first super-cluster regime), and another one for large  $\epsilon \in [40, 150]$  nm, indicating that speckles merge together and with isolated clusters (second super-cluster regime). The SC35 connectivity is further confirmed by running SuperStructure on ROIs masking the speckles, as we observed a clear single exponential decay (dashed line). These regimes are further confirmed by directly looking at the arrangement of identified clusters for certain values of  $\epsilon$  (see Fig. 3, A [inset] and B).

From the SuperStructure curves, we first obtained the density of intra-cluster emitters by fitting the intra-cluster regime with the Poisson function (Eq. 1). Interestingly, both SAF-A and SC35 form clusters with similar densities, while hnRNP-C clusters are less dense (see Fig. 3, D and E). Then, in order to have a quantitative description of the clusters/speckles connectivities, we fitted the curves in the exponential regimes (Eq. 4) to extract the decay length  $\lambda$ . However, a direct comparison is possible only by normalizing decay lengths by the cluster/speckle density (see Materials and methods for details and Fig. S3, A and B). Fig. 3 F highlights that while hnRNP-C has a short normalized decay length  $\lambda^*$  due to the highly connected clusters, SAF-A displays a weaker decay (larger  $\lambda^*$ ) due to sparser connections. Finally, SC35 displays a first (intra-speckle) very connected regime, even more than that of hnRNPs (smaller  $\lambda^*$ ). This is followed by a second (inter-speckle) regime that shows a cluster connectivity weaker than that of hnRNPs.

In summary, our analysis revealed that while different nuclear proteins may have similar cluster sizes or densities of emitters within clusters (e.g., SAF-A and SC35), they have distinct super-cluster arrangements and connectivities. For instance, we find that the super-structures inside nuclear speckles are more connected than those formed by hnRNPs and also very dense (see Fig. 3, E and F; and Table S1). We stress that these features, which we further verified not emerging from technical artifacts (see Fig. S3 C), cannot be quantified using standard clustering algorithms or pair-correlation functions. Additionally, the analysis in Fig. 3, E and F shows that our method is sensitive enough to distinguish connectivity features of two closely related wild-type hnRNPs in cell-based experiments.

The results presented in Fig. 3 give us confidence not only that SuperStructure can be applied to a variety of nuclear wild-type or mutated proteins in different cells, cell stages, and conditions, but also that it has the capability to extract unique features that may yield new mechanistic insights into the functioning of such proteins. For instance, the analysis of SC35 reveals that speckles are themselves made of clusters that are as



regime). **(D)** Normalized all-nucleus average SuperStructure curves in the range  $[0 : 200]$  nm for the three proteins. Average is computed over six nuclei. Shaded regions represent standard deviation from the average. Poisson fits (Eq. 1) for the intra-cluster regime at small  $\epsilon$  are shown in the inset. **(E)** Intra-cluster density of emitters  $\rho_{em}$  as parameter of Poisson fit for six independent nuclei (Eq. 1). **(F)** Normalized decay length  $\lambda^*$  for the super-cluster regimes highlighted in C for six independent nuclei. SuperStructure curves were fit with Eq. 4 to extract the decay length  $\lambda$ , and then the normalization  $\lambda^* = \lambda/\rho_{cl}^{-1/2}$  was performed (where  $\rho_{cl}$  is the detected cluster density at the beginning of each regime of interest). Details are explained in Materials and methods and Fig. S3. P values were calculated using a Student's t test: ns,  $P > 0.05$ ; \*\*,  $P < 0.01$ ; \*\*\*,  $P < 0.001$ .

heavily interconnected as the clusters formed by hnRNP proteins. Given the fact that all these proteins interact with RNA, our findings suggest that RNA binding may facilitate the formation of connections between clusters of proteins; in turn, this also points to a suspected structural role of noncoding RNAs in structuring the organization of the nuclear interior (Hall and Lawrence, 2016). Studying the effect of RNA depletion on the super-cluster connectivity is therefore a natural next step to perform in the future.

In general, while certain mutations or conditions may not alter the size of protein cluster itself, they may affect the connectivity between clusters. In these cases, the analysis provided by SuperStructure would be invaluable and indeed essential to reveal the underlying mechanisms that guide the formation of such protein assemblies.

#### Ceramide clusters at the plasma membrane are not connected

To test our algorithm on a different class of molecules, we applied SuperStructure on published dSTORM datasets (Burgert et al., 2017) taken on ceramides-membrane lipids involved in cellular trafficking (Fig. 4 A). The authors (Burgert et al., 2017) found that bacillus cereus sphingomyelinase (bSMase) treatment increases the size of ceramides clusters and the overall localization density. By applying SuperStructure analysis (Fig. 4 B), we confirmed these results and further detected that the difference in localization density persists inside clusters (see Fig. 4, C and D; and Fig. S4, C and D). Furthermore, we detected the absence of connectivity between clusters, as the large  $\epsilon$  regime is well captured by a Poisson function (Eq. 3) and not by an exponential (see Fig. 4, B and E). In other words, clusters of ceramides behave as unconnected, uniformly and randomly distributed emitters. The possibility of local connectivities at intermediate  $\epsilon$  has also been ruled out, as no merging of clusters was directly observed (see Fig. S4, A and B). The crossing of the curves at  $\epsilon \approx 25$  nm is a consequence of the difference in overall localization density (which in turn causes a horizontal shift between the curves; see Fig. 4, B [inset] and C), rather than a difference in local connectivities. The notable absence of connections between clusters of ceramides further supports that the ones detected in hnRNP-U/C and SC35 are significant.

#### Limitations and potential interpretation pitfalls

While we have provided evidence that SuperStructure can detect connected clusters and distinguish them from noise (at low density) or unconnected but dense clusters, in this section, we discuss potential pitfalls and interpretation issues.

First, as mentioned earlier, datasets should always be segmented in order to identify the main ROI. Spurious localizations outside the ROI (e.g., outside of the nucleus, if we are interested

in nuclear proteins) may affect the curves generated by SuperStructure and render their interpretation difficult. An analogous issue may arise if the localizations are embedded within heterogeneous structures, as in the case of SC35 proteins that form structures strongly connected within nuclear speckles and weakly connected outside speckles (see Fig. 3). Due to this mixed behavior over similar length-scales, it is recommended to restrict the analysis to regions that display similar structural phenotypes. Even better, and to be preferred when possible, is to label the region or structure of interest with orthogonal markers.

The key difference between connected and unconnected (albeit possibly more clustered) structures is the functional form of the SuperStructure curves. However, in some cases, Poisson curves may be difficult to distinguish from exponentials (especially over short intervals). In this case, the best way to identify connected clusters (and distinguish them from noisier or more clustered subregions) is to restrict the analysis over smaller ROIs to avoid potential contaminations and to perform goodness-of-fit tests on the curves. Additionally, in these complex cases we also suggest performing an independent cluster analysis over different length-scales and directly inspecting the results.

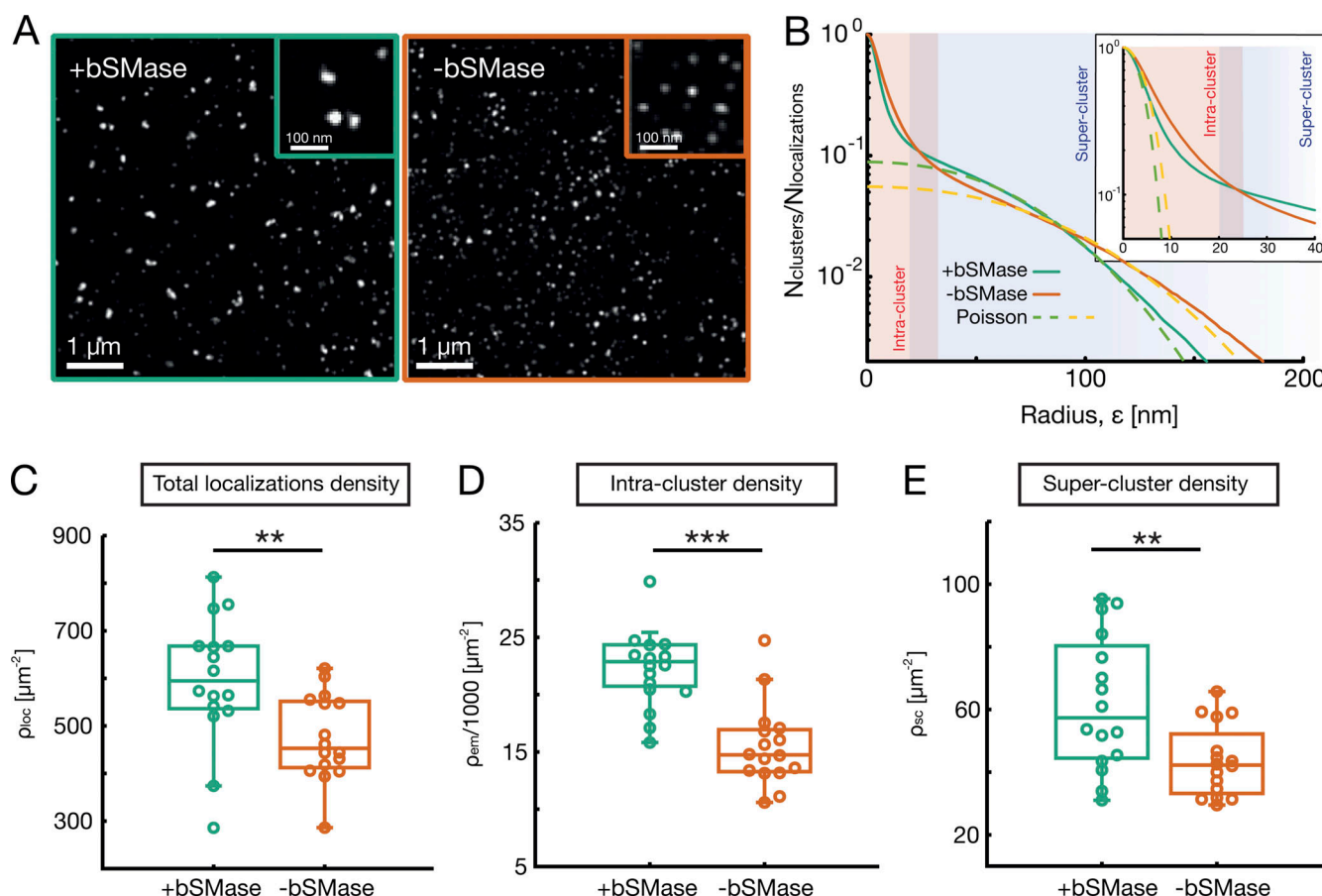
As with all computational algorithms, the danger of incorrect interpretation can be addressed with quality control. In the case of SuperStructure, this means directly monitoring the formation of connected clusters/structures while increasing  $\epsilon$ . Nonetheless, thanks to its parameter-free execution, SuperStructure may currently offer one of the safest ways to analyze SMLM data.

## Discussion

In this work, we have introduced a novel algorithm that extends the traditional idea of cluster analysis of SMLM data and that can quantify both the connections between clusters and the density of emitters within clusters. SuperStructure introduces for the first time the concept of connectivity between clusters, which is different from a random distribution of points at high density. In this concept, connection points are preferentially found in between clusters and this feature manifests itself in SuperStructure curves behaving as single exponentials rather than Poissonian. Because SuperStructure is parameter-free, it does not require any prior knowledge of the sample and it thus takes a crucial step toward a more standardized, portable, and democratic quantification of complex patterns and super-structures in SMLM data.

Here, we have tested the capabilities of SuperStructure first on simulated datasets, where we observed that it could capture not only the degree of connectivity between clusters but also the





**Figure 4. Application of SuperStructure algorithm to ceramide data.** Analysis was performed on published data (Burgert et al., 2017). **(A)** dSTORM reconstruction of ceramides dataset using the shifted histogram method. The left panel represents signal from cells treated with bSMase; the right panel is a control without treatment. **(B)** SuperStructure curves of the two conditions for the entire dataset. Curves show the number of detected clusters normalized by the total number of localizations. The red region highlights the intra-cluster regime, while the blue region highlights the Poissonian unconnected super-cluster regime. The shaded purple region highlights the horizontal shift between the two curves. Dashed lines represent Poisson fits at low and high  $\epsilon$ . **(C–E)** Average density of total localizations (C), intra-cluster density extracted as parameter from Poisson fit (Eq. 3; D), and overall density in the super-cluster regime extracted as parameter from Poisson fit (Eq. 3; E) for +bSMase and –bSMase treatment datasets. Calculations and fits were performed on data and SuperStructure curves from 16 independent circular regions of radius  $r = 1.5 \mu\text{m}$  within the original dataset. P values were calculated using a Student's *t* test: \*\*,  $P < 0.01$ ; \*\*\*,  $P < 0.001$ .

strength of the connections, and then on biological dSTORM data from nuclear proteins and membrane lipids. SuperStructure allowed us to discover that the speckles formed by the splicing factor SC35 are made of connected clusters. Further, that the density of emitters in those clusters is high and the connectivity between clusters even higher than that of hnRNP proteins. We argue that this may reflect the RNA-binding feature that characterizes both hnRNPs and SC35 and that may be driving the formation of interconnected nuclear super-structures. We highlight that this discovery could not be made simply by looking at clustering with traditional algorithms, as both proteins display clusters of similar size at small/intermediate  $\epsilon$ .

We further stress that SuperStructure is perfectly suited to compare different datasets without a priori assumptions (albeit, as discussed before, segmentation to ROIs is recommended for strongly heterogeneous structures). The datasets of nuclear proteins we chose to analyze are an example of this. SAF-A, hnRNP-C, and SC-35 are three nuclear proteins involved in the metabolism of RNA at different stages, and they display three

different connectivity phenotypes, which point to three different nuclear functions. In particular, SAF-A, which also plays a major role in maintaining the chromatin active loci in a de-compacted state, is detected as a fully connected mesh. This finding is in agreement with a previous study that hypothesized the formation of a dynamic and RNA-interacting nuclear mesh made by SAF-A (Nozawa et al., 2017). We thus argue that SuperStructure is a useful tool for studying the structural and functional properties of this nuclear mesh. For instance, we expect that in absence of RNA, the SAF-A mesh would be disrupted and its connectivity strongly weakened (not necessarily affecting the protein clusters, which may be formed via an RNA-independent mechanism, such as phase separation by weak unspecific interactions of SAF-A's intrinsically disordered domain). In turn, the application of SuperStructure would in this case be indispensable for understanding the link between the spatial arrangement, mechanics, and function of this nuclear protein. A similar example is given by the V(D)J locus, whereby interacting segments appear to be trapped by a protein or



chromatin network whose (super-)structure is still poorly understood (Khanna et al., 2019). We argue that SuperStructure can shed light also on this problem.

In addition to all this, super-resolved chromatin tracing (Boettiger et al., 2016; Bintu et al., 2018) and super-resolved imaging of the accessible genome (Xie et al., 2020) generate complex datasets that will benefit from “beyond-traditional-clustering” algorithms. Connections between nanodomains and chromatin paths do not resemble the structure of isolated clusters but rather that of a mesh of clusters, which would be perfectly suited for quantification via the SuperStructure algorithm.

The use of SuperStructure is not limited to biological applications, and we propose it can be used as a standardized and parameter-free tool for assessing imaging technical aspects (van de Linde and Sauer, 2014; Hennig et al., 2015). One of the main issues in SMLM data, especially in dSTORM, is the evaluation of fluorophore blinking quality, as it strongly affects the localization accuracy in the analysis process. For example, an elevated blinking frequency would result in a high emitter density (per frame) and therefore in a high localization inaccuracy due to overlapping emissions. A similar detrimental effect could also be due to a poor blinking signal (few emitted photons per blinking event). As a consequence, lower localization precision of emitters may create pseudo-clusters, as well as pseudo-connections. We envisage that SuperStructure would be well suited to evaluate the blinking quality of fluorophores, for instance by measuring the emerging pseudo-connectivity in a controlled setup, such as fluorophores attached to a grid.

As discussed above, SuperStructure has been developed with the aim of going beyond “simple clustering” and in particular to measure connectivity between clusters. However, our method might be used in combination with other pairwise distance and clustering methods. For instance, one can compute Ripley’s (pairwise distance) functions to preliminarily detect if localizations are uniform or clustered and, in case, what is the average cluster radius. Yet, Ripley’s functions cannot identify single clusters or more complex structures. Thus, one could use SuperStructure to determine whether the system under investigation displays connected or isolated clusters. At the same time, by computing SuperStructure curves, one can have a firm ground to decide the value of  $\epsilon$  that can be used as input in DBSCAN for cluster analysis. This second approach can be used, for example, to measure the size or shape of local super-structures. Indeed, one can fix  $\epsilon$  at the value that identifies super-structures, perform a cluster analysis, and calculate the gyration tensor of the identified clusters.

We tested the segmentation capabilities of the latter approach by estimating the radius and circularity of SC35 speckles; we observed that it yields similar results as the well-known SR-Tesseler software (Levet et al., 2015; see Fig. S5). Although SuperStructure lacks a graphical user interface, it has several advantages. First, it can be run on any operating system and can be easily automatized to run on a large number of cells. Second, since it is based on DBSCAN, the algorithm scales as  $n_\epsilon N^2$  in its simplest implementation (where  $n_\epsilon$  is the number of  $\epsilon$  values used in the analysis and  $N$  is the total number of localizations).

The calculations on different  $\epsilon$  are independent, so SuperStructure scales extremely well with the number of central processing units available. For instance, the analysis of  $n_\epsilon = 100$  values and  $10^5$  localizations can be done on a six-core machine in  $\sim 19$  min. Third, since our algorithm is aimed at extracting beyond-simple-clustering information, it is flexible and intended to be used in combination with other pair-correlation or segmentation methods that are extensively employed for single-clustering analysis.

We conclude by highlighting that SuperStructure provides an unbiased and parameter-free estimation of (1) the density of localizations within single clusters and (2) the formation of super-structures made of connected clusters. Here, we tested SuperStructure both in simulated and cell-based SMLM datasets. Importantly, we revealed previously undocumented system-spanning structures made of connected clusters of nuclear proteins that we argue may have a functional role in shaping genome organization. The use of SuperStructure on cells under different conditions or with protein mutations is thus an exciting direction to uncover the biological significance of these newly discovered nuclear structures.

## Materials and methods

### SuperStructure algorithm

SuperStructure is an algorithm that detects and quantifies super-structures formed by interconnected clusters on SMLM datasets. Additionally, it can also evaluate the density of emitters inside clusters.

SuperStructure is mainly based on DBSCAN, a density-based algorithm to detect clusters of points in arbitrary dimensional space. The key concept underlying DBSCAN scheme is that it groups together points at high density, while it marks as outliers points in low-density regions. After defining a neighborhood size  $\epsilon$ , a point  $x$  can be part of a cluster if the number of points  $N(\epsilon, x)$  within a circular region  $\Omega(\epsilon, x)$  of size  $\epsilon$  centered in  $x$ , exceeds some threshold  $N_{min}$  (or is within the region  $\Omega(\epsilon, y)$  of another point  $y$  satisfying this condition).

The concept of clusters is subject to the choice of  $\epsilon$  and  $N_{min}$  and therefore to some sort of likeness or proximity. Furthermore, the change in number of clusters detected by DBSCAN when varying  $\epsilon$  contains some information of the underlying distribution of points that has been overlooked.

SuperStructure progressively runs DBSCAN to detect the number of clusters  $N_c$  within a broad range of the neighborhood parameter  $\epsilon$ , while  $N_{min}$  is kept fixed. The resulting  $N_c(\epsilon)$  curves, and in particular the change  $dN_c(\epsilon, N_{min})$  due to a small change in neighborhood parameter  $d\epsilon$ , contain fundamental information about the formation and organization of super-structures and connected clusters.

As we aim for a parameter-free algorithm, without losing generality, we fix  $N_{min} = 0$ , which means no minimum number of other emitters necessary in the neighborhood to define a localization as part of a cluster. For  $\epsilon = 0$ , any point is found to be a cluster by itself. Then, points merge upon increasing  $\epsilon \rightarrow \epsilon + d\epsilon$ , resulting in  $dN_c/d\epsilon \leq 0 \forall \epsilon$ . Additionally, the larger  $|dN_c/d\epsilon|$ , the more identified clusters are coalescing together for a certain  $\epsilon$ .

At  $\epsilon$  smaller than the typical (true, rather than the one detected by DBSCAN) cluster size, the decay of  $dN_c/d\epsilon$  is determined by the intra-cluster density of points  $\rho_{em}$  (intra-cluster regime), as they are the points at the highest density. The decay of this regime is Gaussian and it is described by the Poisson function:

$$N_c(\epsilon) = \sum_{k=0}^m c_k \frac{(\pi \rho_{em} \epsilon^2)^k}{k!} e^{-\pi \rho_{em} \epsilon^2}. \quad (1)$$

To understand the origin of this functional form, let us imagine to apply the SuperStructure algorithm by setting  $N_{min} = 0$  and increasing the radius  $\epsilon$ . For sufficiently small  $\epsilon$ , every point is considered as a single cluster itself, as no other points are detected in its neighborhood. However, by increasing  $\epsilon$ , the probability of finding another point in the neighborhood increases, implying that points start to merge in bigger clusters for small  $\epsilon$ . It is then legitimate to argue that the number of detected clusters  $N_c$  decreases (with  $\epsilon$ ) as the probability of not finding any other emitter in the neighborhood. This is the so-called Poisson avoidance function  $N_c(\epsilon) = P(n(\epsilon) = 0) = e^{-\pi \rho_{em} \epsilon^2}$ , and it is a good approximation for very small  $\epsilon$ , where the contribution of clusters formed by two emitters dominates over clusters formed by three or more points. For larger  $\epsilon$ , this function underestimates the number of detected clusters. The number of detected clusters can therefore be described by the probability of not finding more than  $m$  particles in the circle of radius  $\epsilon$ . The function we are seeking is the linear combination of the probabilities of not finding any other point in the neighborhood and finding one or more other points (up to  $m - 1$ ). Being the probability of finding  $k$  particles  $P(n(\epsilon) = k) = \frac{(\pi \rho_{em} \epsilon^2)^k}{k!} e^{-\pi \rho_{em} \epsilon^2}$ , it is then straightforward to get the functional form of Eq. 1.

Note that  $c_k = 1/(k + 1)$  in Eq. 1 is to avoid overcounting clusters. In fact, if we consider two points within distance  $\epsilon$  from each other (and hence in the same cluster), both points will count toward  $P(n(\epsilon) = 1)$ , so this contribution must be divided by 2, etc. Importantly, Eq. 1 displays a natural length-scale  $\kappa_0 = (\pi \rho_{em})^{-1/2}$  that is intrinsically determined by the internal density of emitters  $\rho_{em}$ . Therefore,  $\rho_{em}$  is a parameter that can be quantified by fitting the  $N_c(\epsilon)$  curve, and it can also be used to quantify the approximate upper limit of this regime (with 99% confidence level):

$$\epsilon^* \simeq 3\kappa_0 = 3/\sqrt{\pi \rho_{em}} = 3R_{cl}/\sqrt{N_{em}}, \quad (2)$$

where  $R_{cl}$  is the average cluster radius and  $N_{em}$  is the average number of localizations within a single cluster. We successfully tested that SuperStructure curves are well fitted by Eq. 1 up to  $m = 2$  using a system where we simulated localization of points inside a single cluster (see Fig. S1).

At  $\epsilon$  of the order than the typical (true) cluster size, the decay is determined by the rate at which distinct clusters merge upon  $\epsilon \rightarrow \epsilon + d\epsilon$  (first super-cluster regime). This merging can be due to either (1) distinct clusters starting to overlap as their distance is smaller than  $\epsilon$  or (2) the presence of points, which we call connections, bridging two clusters. In case of total absence of connectivity and a homogeneous cluster distribution, the merging is only due to the random positioning of clusters, and therefore, it also follows a Poisson function:

$$N_c(\epsilon) = f \sum_{k=0}^m c_k \frac{(\pi \rho_{cl} \epsilon^2)^k}{k!} e^{-\pi \rho_{cl} \epsilon^2}, \quad (3)$$

where  $f$  is a normalization factor and  $\rho_{cl}$  the density of clusters. We observed that SuperStructure curves of simulated systems are well fitted by using  $m = 1$ . This equation holds also in presence of noise, but in that case,  $\rho_{cl} \rightarrow \rho_{cl} + \rho_{noise}$  (see Fig. S2). The decay is different in presence of connections between clusters; connected clusters will merge at smaller  $\epsilon$  than unconnected ones (assuming same distance between the centers of clusters). In particular, the larger the number of connections or of the local density of connection points  $\rho_{conn}$  (i.e., thicker connections), the faster the merging of bridged clusters as a function of  $\epsilon$  and thus the larger  $|dN_c/d\epsilon|$ . The functional form of this second regime is exponential in presence of connections:

$$N_c(\epsilon) = g \cdot e^{-\epsilon/\lambda}, \quad (4)$$

where  $g$  is a normalization factor and  $\lambda$  the decay length quantifying the rate of decay and therefore the connectivity. This decay length can be used to discern systems that exhibit either different grades of connectivity or homogeneous meshes at different densities. Note  $\lambda$  purely quantifies the connectivity only when the cluster density  $\rho_{cl}$  is small and homogeneous, as we could have underlying highly dense clusters overlapping and therefore merging. We showed that  $\lambda \sim \rho_{cl}^{-1/2}$  and therefore the pure connectivity decay length can be further evaluated if the density of clusters is known:  $\lambda^* \sim \lambda/\rho_{cl}^{-1/2}$ .

We need to stress that by choosing  $N_{min} = 0$ , connections will also be considered as points to be merged. However, it is important that we identify connection points as having a lower local density  $\rho_{conn}$  than the groups of points that are bridged by them (clusters). In this way, they will merge in this second regime to form super-structures. The limiting case in which the local density of connection points is the same as the one in the clusters at the two ends of the connections is indistinguishable from the case of one elongated cluster. A special case is that in which both clusters and connections have the same density of points but the connections are slightly detached from the clusters, thus forming three independent clusters at intermediate  $\epsilon$ , which may then merge (we assume this to be a rare event). The above reasoning can be extended to multiply connected clusters via the analysis of pairwise connections.

At larger  $\epsilon$ , we could have additional super-cluster regimes if the system is heterogeneous. Most common cases showing two (or more) super-cluster regimes are the following: (1) inhomogeneous system displaying different connectivities at different length-scales, (2) connected clusters embedded in a noisy environment (in this case we observe an exponential followed by a Poissonian decay), and (3) unconnected clusters within a random noise and/or unconnected clusters at different densities (in this case, we observe two or more Poissonian decays).

### SuperStructure pipeline

To apply SuperStructure, we adopt the following steps.

(1) Generation of SuperStructure curves. We run SuperStructure on a SMLM dataset by first masking our data in the ROI, such as the nucleus for nuclear proteins as mentioned in the

section below. Then, we choose a  $\epsilon$  range to analyze. For example, in SMLM datasets of nuclear proteins a typical choice is  $\epsilon \in [0 : 200] \text{ nm}$  with  $d\epsilon = 2 \text{ nm}$ . One should notice that lower  $d\epsilon$  may be necessary for fitting the intra-cluster regime. SuperStructure curves are generated by progressively running DBSCAN clustering algorithm on the SMLM dataset in the chosen  $\epsilon$  range (and  $N_{\min} = 0$ ). The DBSCAN software we use is from <https://github.com/gyaikhom/dbscan>, and the progressive run is performed with bash scripts available in the repository. SuperStructure output curves are saved in a three-column file ( $\epsilon$ ,  $N_d$ ,  $N_d/N_{\text{loc}}$ ), where  $N_d$  is the number of detected clusters for the corresponding  $\epsilon$  and  $N_{\text{loc}}$  the number of total localizations. Additionally, the classification of localizations in clusters is saved on a separate file for every  $\epsilon$ .

(2) Evaluation of SuperStructure regimes. As a second step, we evaluate regimes by plotting and investigating SuperStructure curves (we adopt a log scale in the y axis). This step includes a preliminary check for the number of regimes and their decay behavior (exponential versus Poissonian). In the case we observe a single Poissonian behavior, we can state that the dataset does not show any, or very limited, connectivity, and therefore, we are in presence of homogeneous isolated clusters (and eventually noise). Limited connectivity needs to be checked with a cluster analysis and direct dataset observation in case noise has obscured an exponential decay. On the other hand, if we observe a single exponential regime (a straight line in a log-linear plot), we conclude that the system is made of fully connected clusters. If SuperStructure curves show multiple super-cluster regimes, it is likely that the system is heterogeneous. Indeed, multiple exponential regimes may reflect heterogeneous/multiscale connectivities combined with heterogeneous distributions of clusters. Alternatively, we may find also a combination of exponential and Poissonian regimes, and in this case, the system may be made of connected clusters embedded in a noisy region. Other more complex combinations may be possible; however, one should notice that in heterogeneous systems, it might be difficult to recognize and fit super-cluster regimes. To clarify these contributions, it is useful to combine the analysis of SuperStructure curves with a direct observation of the dataset and identified structures and to run SuperStructure on smaller ROIs to analyze different regions of the sample with similar structural phenotypes. Nonetheless, SuperStructure will be able to unambiguously detect differences in connectivity and behaviors in, for example, samples that have been subjected to different conditions or expressing mutated proteins.

(3) Fit of SuperStructure regimes. Once regimes have been identified, one needs to define the boundaries where regimes crossover from one to another. This can be either done manually or by using a preemptive goodness-of-fit test (this procedure would also define fitting ranges). The intra-cluster regime is typically fitted with a Poisson equation (Eq. 1) to evaluate the density of emitters inside clusters as well as obtain an estimation of the upper limit of the intra-cluster regime (using Eq. 2). For super-cluster regimes, we use Eq. 3 if they show a Poissonian decay (curved on a log-linear plot) or Eq. 4 if they otherwise appear straight on a log-linear plot; from the latter, we quantify the connectivity parameter  $\lambda$ . We can then additionally calculate

the cluster density  $\rho_{cl}$  to extract the pure connectivity part  $\lambda^* = \lambda/\rho_{cl}^{-1/2}$ . The cluster density  $\rho_{cl}$  can be computed by performing a cluster analysis with DBSCAN on local circular regions representative of that decay regime and by fixing  $\epsilon$  at the start of that regime (e.g., by counting the number of clusters one obtains by fixing  $\epsilon$  at the beginning of the yellow area in Fig. 3). In the section below and in Fig. S3, we describe in detail the procedure for  $\lambda$  normalization for the nuclear protein datasets. Finally, and optionally, it is also possible to define a single function fitting the entire curve by either (1) defining a piecewise function where every piece is the fit of the corresponding regime or (2) adding together the contribution of the different regimes (appropriately weighted). We performed fits with a combination of bash and gnuplot scripts available in the repository.

### Simulated dataset generation and SuperStructure analysis

The simulated dataset consists of spatially homogeneous and interconnected clusters randomly distributed on a plane. We set to work with clusters made by taking random clusters centers on the plane and by sampling  $N_{em} = 80$  emitters within a Gaussian of standard deviation  $\sigma_{em} = 20 \text{ nm}$ , thereby setting the cluster radius to  $R_{cl} = 2 \sigma_{em} = 40 \text{ nm}$  with a 95% confidence and the intra-cluster emitters density at  $\rho_{em} = 16,000 \mu\text{m}^{-2}$ . The clusters are positioned in a  $L = 3.5 \mu\text{m}$  large area, and their number  $N_{cl}$  is varied in order to consider different clusters densities. In the example shown in the main text, we fixed  $N_{cl} = 100$ , thus fixing a cluster density to approximately  $\rho_{cl} = 8.2 \mu\text{m}^{-2}$ , roughly similar to the values found in experiments for some nuclear proteins. Pairs of clusters are connected with probability  $p_r$  if they are positioned closer than a distance  $b = 1 \mu\text{m}$ . The value of  $p_r$  is calculated as the ratio between the actual drawn connections and  $N_{cl}(N_{cl} - 1)/2$ , which is the maximum possible connections (i.e., when every cluster is connected with every other cluster). To generate a single connection, we considered the vector joining the centers of two clusters and sampled one emitter with probability  $p_{conn}$  every 10 nm. Emitters are sampled from a 2D Gaussian centered on the vector connecting the two clusters centers and with a width  $\sigma_{conn} = 10 \text{ nm}$ . In the main text, we fixed  $p_{conn} = 0.5$ . Note that  $p_r$  controls the number of connections, while  $p_{conn}$  controls their density,  $\rho_{conn}$ . We generated at least 20 independent replicas for each simulated dataset using a combination of bash and python scripts, and then we ran SuperStructure analysis in the range  $\epsilon \in [0 : 400] \text{ nm}$  with a change  $d\epsilon = 2 \text{ nm}$ . If not differently specified, the first super-cluster regime was fitted with Eq. 4 for  $\epsilon \in [15 : 60]$ , while the second super-cluster regime was fitted with either Eq. 3 (unconnected systems) or Eq. 4 (connected systems) for  $\epsilon \in [70 : 300]$ .

### Cell preparation for dSTORM imaging of nuclear proteins

hTERT-RPE1 cells (catalog no. ATCC-CRL-4000; American Type Culture Collection) were grown overnight in an eight-well Lab-Tek II Chambered Coverglass-1.5 borosilicate glass (Thermo Fisher Scientific) at 37°C at initial concentration of  $10^5 \text{ cells/ml}$  in 400  $\mu\text{l}$  (~40% confluency). We fixed the cells with 4% PFA (Sigma-Aldrich) for 10 min, washed three times in PBS, permeabilized with 0.2% Triton X-100 (Sigma-Aldrich) for 10 min,



washed three times in PBS, and blocked with 1% BSA (Sigma-Aldrich) for 10 min.

Immunofluorescence labeling was done by exposing the cells for 2 h to (1) hnRNP-U polyclonal rabbit antibody (A300-690A; Bethyl Laboratories) at 10  $\mu\text{g/ml}$ , (2) hnRNP-C1/C2 (4F4) mouse monoclonal antibody (sc-32308; Santa Cruz Biotechnology) at 0.2  $\mu\text{g/ml}$ , or (3) SC-35 mouse monoclonal antibody (abl1826; Abcam) at 2  $\mu\text{g/ml}$  and then three washes. Cells were then exposed for 1 h to secondary antibody. The secondary antibody was made by AffiniPure  $F(ab')_2$  fragment donkey anti-rabbit or donkey anti-mouse IgG (H+L; 711-006-152 and 715-007-003, Jackson ImmunoResearch Europe Ltd.) conjugated to the organic fluorophore CF647 (92238A-IVL; Sigma-Aldrich) at a stoichiometric ratio of  $\sim 1$ . After that, cells were washed three times in PBS.

Oxygen scavenger imaging buffer based on the glucose oxidase enzymatic system (GLOX) for dSTORM was prepared fresh. The recipe employed was similar to that used previously (McSwiggen et al., 2019). We mixed (1) 5.3 ml of 200 mM Tris and 50 mM NaCl solution with (2) 2 ml of 40% glucose solution, (3) 200  $\mu\text{l}$  GLOX, (4) 1.32 ml of 1 M 2-mercaptoethanol (Sigma-Aldrich), and (5) 100  $\mu\text{l}$  of 50  $\mu\text{g/ml}$  DAPI solution (Sigma-Aldrich). The GLOX solution was made by mixing 160  $\mu\text{l}$  of 200 mM Tris and 50 mM NaCl with 40  $\mu\text{l}$  catalase from bovine liver (Sigma-Aldrich) and 18 mg glucose oxidase (Sigma-Aldrich).

The 8.9-ml final solution was enough to fill the chambers of the eight-well dish; a coverglass was sealed at the top of the dish to prevent inflow of oxygen.

### dSTORM acquisition of nuclear proteins

We performed 3D-STORM acquisitions using a Nikon N-STORM total internal reflection fluorescence system (TIRF) with Eclipse Ti-E inverted microscope and laser TIRFilluminator (Nikon). We equipped the microscope with a CFI SR HP Apo TIRF 100 $\times$  objective lens (N.A. 1.49) and applied a 1.5 $\times$  additional optical zoom. We also used a cylindrical astigmatic lens to obtain elliptical shapes for emitters that reflect their z-position (Huang et al., 2008). Laser light was provided via a Nikon LU-NV laser bed with 405-, 488-, 561-, and 640-nm laser lines. In particular, CF647 fluorophores were stochastically excited using the 640-nm laser beam with an additional 405-nm weak pulse. Images were acquired with an Andor iXon 897 EMCCD camera (Andor Technologies). The z-position was stabilized during the entire acquisition by the integrated perfect focus system. Acquisition were performed at room temperature.

For every nucleus, we acquired a stack of 20,000 frames at 19-ms exposure time by using the Nikon NIS-Element software. Acquired images have a 256  $\times$  256 pixel resolution with pixel size equal to 106 nm. For every condition (SAF-A, hnRNP-C, and SC35), we acquired six nuclei (i.e., six independent datasets).

### Raw images and post-processing analysis for nuclear protein data

The raw stack of frames was initially segmented based on a DAPI marker to carefully mask out the extranuclear signal. Then, frames were analyzed using FIJI (Schindelin et al., 2012) and in

particular the Thunderstorm plugin (Ovesný et al., 2014). First, we filtered them by using Wavelet functions to separate signal from noise. The B-Spline order was set to 3 and the B-Spline scale to 2.0 as suggested previously (Ovesný et al., 2014) for localizations of  $\sim 5$  pixels. To localize the emitters centroids, we thresholded filtered images (threshold value was set 1.2 times the standard deviation of the first Wavelet function) and calculated the local maximum relative to the eight nearest neighbors. Finally, we fitted the emitters signal distribution with elliptical gaussians (ellipses are necessary for z-position reconstruction) using the weighted least-square method and by setting 3 pixels as fitting radius and 1.6 pixels as initial sigma.

Localized data were then postprocessed using the same plugin. We corrected the xy drift using a pair-correlation analysis, filtered data with a position uncertainty  $< 40$  nm, restricted the z-position to the interval  $[-100 : 100]$  nm, and projected the data in a 2D plane, as the z-axis precision is  $\sim 100$  nm.

Reconstructed images shown in the main text were created by using the average shifted histograms method of the same plugin with 10 $\times$  magnification (final resolution set to 10.6 nm/pixel).

### SuperStructure analysis for nuclear protein data

SuperStructure analysis was run on the entire nuclear region by setting  $N_{\min} = 0$  and by increasing  $\epsilon$  in the range  $[0 : 200]$  nm, and “all-nucleus” curves were generated for six independent nuclei. We set the change rate  $d\epsilon = 0.25$  nm for  $\epsilon \in [0 : 10]$  nm and  $d\epsilon = 10$  nm for  $\epsilon \in [10 : 200]$  nm. This choice was due to the higher resolution necessary to extract intra-cluster information at small  $\epsilon$ . As shown in Fig. 3, SuperStructure all-nucleus curves show that SAF-A has a single exponential super-cluster regime, while hnRNP-C and SC35 have two regimes. In the case of hnRNP-C, the second regime is due to weakly connected and sparse clusters in nucleoli, while in SC35 it is due to the cluster/connectivity heterogeneity in the system (i.e., speckles). Therefore, we additionally run SuperStructure analysis on local ROIs for hnRNP-C and SC35 to obtain the isolated contribution for the first super-cluster regime. In particular, for hnRNP-C, we considered five independent circular ROIs per nucleus with radius  $r = 1.5$   $\mu\text{m}$  within the nuclear mesh; for SC35, we considered five independent circular ROIs per nucleus with radius  $r = 0.5$   $\mu\text{m}$  within speckles. We ran the analysis on these ROIs and generated SuperStructure “local” curves (five for each nucleus).

The values of the intra-cluster density  $\rho_{em}$  were extracted by fitting with Eq. 1 the intra-cluster regime in the all-nucleus curves in the range  $\epsilon \in [0, 3]$  nm. Resulting average values are  $\rho_{em}^{hnRNP-C} = 7,973 \pm 1,732$   $\mu\text{m}^{-2}$ ,  $\rho_{em}^{SAF-A} = 16,998 \pm 2,444$   $\mu\text{m}^{-2}$ , and  $\rho_{em}^{SC35} = 18,680 \pm 1,520$   $\mu\text{m}^{-2}$ .

Then, we identified the super-cluster regimes of interest: the first super-cluster regimes of SAF-A and hnRNP-C and both super-cluster regimes of SC35 (SC35-1 and SC35-2). For SAF-A and SC35-2, the decay length  $\lambda$  was obtained by fitting all-nucleus curves with Eq. 4. For hnRNP-C and SC35-1, we fitted the local curves (five curves per nucleus) and then averaged  $\lambda$  values obtained from different local curves in the same nucleus. Fit ranges are  $\epsilon \in [16, 100]$  nm for SAF-A,  $\epsilon \in [14, 70]$  nm for

hnRNP-C,  $\epsilon \in [8, 20]$  nm for SC35-1, and  $\epsilon \in [40, 150]$  nm for SC35-2.

Finally, the values of  $\lambda$  for SAF-A, hnRNP-C, SC35-1, and SC35-2 were normalized by the cluster density:  $\lambda^* = \lambda / \rho_{cl}^{-1/2}$ . In the case of SAF-A and SC35-2, the normalization was performed for  $\lambda$  for every nucleus by using the average cluster density  $\rho_{cl}$  of that nucleus. In particular,  $\rho_{cl}$  was calculated as the average of the cluster density in five independent circular regions of radius  $r$  in the same nucleus as shown in the example of Fig. S3 A. In the case of hnRNP-C and SC35-1, where  $\lambda$  values were obtained from local curves, the normalization of  $\lambda$  was performed using the cluster density of the same local region; then,  $\lambda^*$  values obtained from different regions in the same nucleus were averaged (see Table S1). The number of clusters estimation (to calculate the cluster density) was made with DBSCAN by setting  $N_{min} = 0$  and  $\epsilon$  close to the beginning of the exponential regime of interest, as shown in Fig. S3 B, and by keeping only clusters with at least 30 particles. To compute the cluster density, for SAF-A and hnRNP-C, we set local circular regions of radius  $r = 1.5$   $\mu\text{m}$  and fixed  $\epsilon = 20$  nm for cluster analysis (for hnRNP-C, we used the same local regions as defined above). For SC35, we considered two sets of local regions: (1) inside speckles to normalize the shorter decay length, where we used ROIs with radius  $r = 500$  nm and fixed  $\epsilon = 10$  nm for cluster analysis (same regions as above); and (2) outside speckles to normalize the longer decay length, where we used ROIs with radius  $r = 1.5$   $\mu\text{m}$  and  $\epsilon = 40$  nm for cluster analysis. Average nuclear values of  $\lambda$ ,  $\rho_{cl}$ , and  $\lambda^*$  are shown in Table S1.

### SuperStructure analysis of ceramide data

SuperStructure analysis was run on the two ceramide datasets provided by the authors from Burgert et al. (2017), namely +bSMase and -bSMase, by setting  $N_{min} = 0$  and  $\epsilon \in [0 : 200]$ . We set  $d\epsilon = 0.5$  nm for  $\epsilon \in [0 : 10]$  nm and  $d\epsilon = 2$  nm for  $\epsilon \in [10 : 200]$  nm. This choice was due to the higher resolution necessary to extract intra-cluster information at small  $\epsilon$ . From the curves in Fig. 4 B, it is clear that there is no strong connectivity (we observe a Poissonian decay). Therefore, we identified free unclustered emitters as noise. We additionally ran SuperStructure in 16 independent local circular regions of radius  $r = 1.5$   $\mu\text{m}$  to extract the quantities of interest. In particular, we measured the average densities of total localizations,  $\rho_{loc}^+ = 595 \pm 130$   $\mu\text{m}^{-2}$  and  $\rho_{loc}^- = 475 \pm 87$   $\mu\text{m}^{-2}$ , respectively, for +bSMase and -bSMase treatment. This is in accordance with results in the original paper. Then, we fitted local SuperStructure curves in the intra-cluster regime with Eq. 1 for  $\epsilon \in [0 : 3]$  nm:  $\rho_{em}^+ = 22,391 \pm 3,306$   $\mu\text{m}^{-2}$  and  $\rho_{em}^- = 15,505 \pm 3,470$   $\mu\text{m}^{-2}$ , respectively, for +bSMase and -bSMase treatments. Finally, we fitted local SuperStructure curves in the super-cluster regime with Eq. 3 in the range  $\epsilon \in [50 : 200]$  nm for +bSMase and  $\epsilon \in [60 : 200]$  nm for -bSMase (the difference in fit starting value is explained by a horizontal shift between the two curves):  $\rho_{sc}^+ = 62.01 \pm 20.76$   $\mu\text{m}^{-2}$  and  $\rho_{sc}^- = 43.56 \pm 11.05$   $\mu\text{m}^{-2}$ . These two values are in accordance with the sum of cluster density and noise at the  $\epsilon$  value where the fit starts. We additionally performed a cluster analysis with DBSCAN, and results are in agreement with the original paper (see Fig. S4 for details). To

verify that there is no limited connectivity hidden by noise, we performed a cluster analysis at two different values of  $\epsilon$  and monitored the change in density of clusters and density of free emitters (see Fig. S4 for details).

### Data availability

The simulated and experimental datasets that support the findings of this study are available from the corresponding authors upon request.

### Code availability

The code for the generation of SuperStructure curves is available at <https://git.ecdf.ed.ac.uk/dmichiel/superstructure>.

### Online supplemental material

Fig. S1 shows a simulated distribution of points inside a single cluster and how it is well represented by Eq. 1 in Materials and methods. Fig. S2 shows SuperStructure curves for simulated datasets of connected clusters in different conditions, including systems with different cluster densities, systems embedded in a noisy environment, and fully connected meshes. Fig. S3 shows how the normalization of  $\lambda$  was performed in nuclear protein data (exhaustively explained in Materials and methods) and that nuclear proteins connectivity is not a technical artifact. Fig. S4 shows that there is no local connectivity in ceramide data and confirms the original paper's results on ceramide cluster size. Fig. S5 shows SuperStructure + DBSCAN segmentation capabilities by estimating the radius and circularity of SC35 speckles alongside SR-Tesseler software. Table S1 recapitulates values for  $\lambda$ ,  $\rho_{cl}$ , and  $\lambda^*$  in nuclear protein data.

### Acknowledgments

The authors thank the Edinburgh Super-Resolution Imaging Consortium (Institute of Genetics and Molecular Medicine section), in particular Matthew Pearson and Ann Wheeler, for help and support. The authors are grateful to Markus Sauer for providing the ceramides data. M. Marendza and D. Michieletto also thank Ibrahim Cissé for an igniting discussion and Davide Marenduzzo's group for discussions.

M. Marendza is a cross-disciplinary postdoctoral fellow supported by funding from the University of Edinburgh and the Medical Research Council (core grant MC\_UU\_00009/2 to the Medical Research Council Institute of Genetics and Molecular Medicine). S. van de Linde is supported by the Academy of Medical Sciences, the British Heart Foundation, the Government Department of Business, Energy and Industrial Strategy, and the Wellcome Trust Springboard Award (SBF003/1163). N. Gilbert is funded by the UK Medical Research Council (grant MC\_UU\_00007/13). D. Michieletto is a Royal Society University Research Fellow and was supported by the Leverhulme Trust (grant ECF-2019-088) and European Research Council Starting Grant (Topologically Active Polymers [TAP] grant 947918). The authors thank the Scottish University Life Science Alliance for support through a technology seed grant (Worktribe Project ID 8824507).

The authors declare no competing financial interests.

Author contributions: M. Marena, D. Michieletto, and N. Gilbert conceived the project. M. Marena and D. Michieletto analyzed both simulated and experimental datasets. M. Marena, S. van de Linde, and D. Michieletto generated the simulated dataset. M. Marena, E. Lazarova, and D. Michieletto performed super-resolution experiments and localization analysis. M. Marena, D. Michieletto, S. van de Linde, and N. Gilbert wrote the manuscript, with input from all authors.

Submitted: 2 October 2020

Revised: 6 January 2021

Accepted: 23 February 2021

## References

- Baumgart, F., A.M. Arnold, K. Leskova, K. Staszek, M. Fölser, J. Weghuber, H. Stockinger, and G.J. Schütz. 2016. Varying label density allows artifact-free analysis of membrane-protein nanoclusters. *Nat. Methods*. 13:661–664. <https://doi.org/10.1038/nmeth.3897>
- Beliveau, B.J., A.N. Boettiger, M.S. Avendaño, R. Jungmann, R.B. McCole, E.F. Joyce, C. Kim-Kiselak, F. Bantignies, C.Y. Fonseca, J. Erceg, et al. 2015. Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nat. Commun.* 6: 7147. <https://doi.org/10.1038/ncomms8147>
- Bintu, B., L.J. Mateo, J.H. Su, N.A. Sinnott-Armstrong, M. Parker, S. Kinrot, K. Yamaya, A.N. Boettiger, and X. Zhuang. 2018. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*. 362:eaau1783. <https://doi.org/10.1126/science.aau1783>
- Boettiger, A.N., B. Bintu, J.R. Moffitt, S. Wang, B.J. Beliveau, G. Fudenberg, M. Imakaev, L.A. Mirny, C.T. Wu, and X. Zhuang. 2016. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*. 529:418–422. <https://doi.org/10.1038/nature16496>
- Brangwynne, C.P., P. Tompa, and R.V. Pappu. 2015. Polymer physics of intracellular phase transitions. *Nat. Phys.* 11:899–904. <https://doi.org/10.1038/nphys3532>
- Bronshtein, I., E. Kepten, I. Kanter, S. Berezin, M. Lindner, A.B. Redwood, S. Mai, S. Gonzalo, R. Foisner, Y. Shav-Tal, and Y. Garini. 2015. Loss of lamin A function increases chromatin dynamics in the nuclear interior. *Nat. Commun.* 6:8044. <https://doi.org/10.1038/ncomms9044>
- Burgert, A., J. Schlegel, J. Bécam, S. Doose, E. Bieberich, A. Schubert-Unkmeir, and M. Sauer. 2017. Characterization of Plasma Membrane Ceramides by Super-Resolution Microscopy. *Angew. Chem. Int. Ed. Engl.* 56: 6131–6135. <https://doi.org/10.1002/anie.201700570>
- Caudron-Herger, M., K. Müller-Ott, J.P. Mallm, C. Marth, U. Schmidt, K. Fejes-Tóth, and K. Rippe. 2011. Coding RNAs with a non-coding function: maintenance of open chromatin structure. *Nucleus*. 2:410–424. <https://doi.org/10.4161/nucl.2.5.17736>
- Cho, W.-K., J.-H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, and I.I. Cisse. 2018. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*. 361:412–415. <https://doi.org/10.1126/science.aar4199>
- Chong, S., C. Dugast-Darzacq, Z. Liu, P. Dong, G.M. Dailey, C. Cattoglio, A. Heckert, S. Banala, L. Lavis, X. Darzacq, and R. Tjian. 2018. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science*. 361:eaar2555. <https://doi.org/10.1126/science.aar2555>
- Cisse, I.I., I. Izeddin, S.Z. Causse, L. Boudarene, A. Senecal, L. Muresan, C. Dugast-Darzacq, B. Hajj, M. Dahan, and X. Darzacq. 2013. Real-time dynamics of RNA polymerase II clustering in live human cells. *Science*. 341:664–667. <https://doi.org/10.1126/science.1239053>
- Deniz, A.A., S. Mukhopadhyay, and E.A. Lemke. 2008. Single-molecule biophysics: at the interface of biology, physics and chemistry. *J. R. Soc. Interface*. 5:15–45. <https://doi.org/10.1098/rsif.2007.1021>
- Ester, M., H.-P. Kriegel, J. Sander, and X. Xu. 1996. 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*. AAAI Press. 226–231.
- Frank, L., and K. Rippe. 2020. Repetitive RNAs as Regulators of Chromatin-Associated Subcompartment Formation by Phase Separation. *J. Mol. Biol.* 432:4270–4286. <https://doi.org/10.1016/j.jmb.2020.04.015>
- Garcia-Parajo, M.F., A. Cambi, J.A. Torreno-Pina, N. Thompson, and K. Jacobson. 2014. Nanoclustering as a dominant feature of plasma membrane organization. *J. Cell Sci.* 127:4995–5005. <https://doi.org/10.1242/jcs.146340>
- Goundaroulis, D., E. Lieberman Aiden, and A. Stasiak. 2020. Chromatin Is Frequently Unknotted at the Megabase Scale. *Biophys. J.* 118:2268–2279. <https://doi.org/10.1016/j.bpj.2019.11.002>
- Griffié, J., M. Shannon, C.L. Bromley, L. Boelen, G.L. Burn, D.J. Williamson, N.A. Heard, A.P. Cope, D.M. Owen, and P. Rubin-Delanchy. 2016. A Bayesian cluster analysis method for single-molecule localization microscopy data. *Nat. Protoc.* 11:2499–2514. <https://doi.org/10.1038/nprot.2016.149>
- Hall, L.L., and J.B. Lawrence. 2016. RNA as a fundamental component of interphase chromosomes: could repeats prove key? *Curr. Opin. Genet. Dev.* 37:137–147. <https://doi.org/10.1016/j.gde.2016.04.005>
- Hennig, S., S. van de Linde, S. Bergmann, T. Huser, and M. Sauer. 2015. Quantitative Super-Resolution Microscopy of Nanopipette-Deposited Fluorescent Patterns. *ACS Nano*. 9:8122–8130. <https://doi.org/10.1021/acsnano.5b02220>
- Henriques, R., C. Griffiths, E. Hesper Rego, and M.M. Mhlana. 2011. PALM and STORM: unlocking live-cell super-resolution. *Biopolymers*. 95: 322–331. <https://doi.org/10.1002/bip.21586>
- Huang, B., W. Wang, M. Bates, and X. Zhuang. 2008. Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy. *Science*. 319:810–813. <https://doi.org/10.1126/science.1153529>
- Izeddin, I., C.G. Specht, M. Lelek, X. Darzacq, A. Triller, C. Zimmer, and M. Dahan. 2011. Super-resolution dynamic imaging of dendritic spines using a low-affinity photoconvertible actin probe. *PLoS One*. 6:e15611. <https://doi.org/10.1371/journal.pone.0015611>
- Jackson, D.A., A. Pombo, and F. Iborra. 2000. The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells. *FASEB J.* 14:242–254. <https://doi.org/10.1096/fasebj.14.2.242>
- Kapanidis, A.N., A. Lepore, and M. El Karoui. 2018. Rediscovering Bacteria through Single-Molecule Imaging in Living Cells. *Biophys. J.* 115: 190–202. <https://doi.org/10.1016/j.bpj.2018.03.028>
- Khanna, N., Y. Zhang, J.S. Lucas, O.K. Dudko, and C. Murre. 2019. Chromosome dynamics near the sol-gel phase transition dictate the timing of remote genomic interactions. *Nat. Commun.* 10:2771. <https://doi.org/10.1038/s41467-019-10628-9>
- Larson, A.G., D. Elnatan, M.M. Keenen, M.J. Trnka, J.B. Johnston, A.L. Burlingame, D.A. Agard, S. Redding, and G.J. Narlikar. 2017. Liquid droplet formation by HP1a suggests a role for phase separation in heterochromatin. *Nature*. 547:236–240. <https://doi.org/10.1038/nature22822>
- Leidescher, S., J. Nuebler, Y. Feodorova, E. Hildebrand, S. Ullrich, S. Bultmann, S. Link, K. Thanisch, J. Dekker, H. Leonhardt, et al. 2020. Spatial Organization of Transcribed Eukaryotic Genes. *bioRxiv*. doi: (Preprint posted May 21, 2020). <https://doi.org/10.1101/2020.05.20.106591>
- Levet, F., E. Hosy, A. Kechkar, C. Butler, A. Beghin, D. Choquet, and J.-B. Sibarita. 2015. SR-Tesseler: a method to segment and quantify localization-based super-resolution microscopy data. *Nat. Methods*. 12: 1065–1071. <https://doi.org/10.1038/nmeth.3579>
- Lin, S., G. Coutinho-Mansfield, D. Wang, S. Pandit, and X.-D. Fu. 2008. The splicing factor SC35 has an active role in transcriptional elongation. *Nat. Struct. Mol. Biol.* 15:819–826. <https://doi.org/10.1038/nsmb.1461>
- Maharana, S., J. Wang, D.K. Papadopoulos, D. Richter, A. Pozniakovskiy, I. Poser, M. Bickle, S. Rizk, J. Guillén-Boixet, T.M. Franzmann, et al. 2018. RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science*. 360:918–921. <https://doi.org/10.1126/science.aar7366>
- Maisner, A., S. Dillinger, G. Längst, L. Schermelleh, H. Leonhardt, and A. Németh. 2020. Super-resolution in situ analysis of active ribosomal DNA chromatin organization in the nucleolus. *Sci. Rep.* 10:7462. <https://doi.org/10.1038/s41598-020-64589-x>
- Malkusch, S., and M. Heilemann. 2016. Extracting quantitative information from single-molecule super-resolution imaging data with LAMA - Localization Microscopy Analyzer. *Sci. Rep.* 6:34486. <https://doi.org/10.1038/srep34486>
- McSwiggen, D.T., A.S. Hansen, S.S. Teves, H. Marie-Nelly, Y. Hao, A.B. Heckert, K.K. Umamoto, C. Dugast-Darzacq, R. Tjian, and X. Darzacq. 2019. Evidence for DNA-mediated nuclear compartmentalization distinct from phase separation. *eLife*. 8:e47098. <https://doi.org/10.7554/eLife.47098>
- Michieletto, D., and N. Gilbert. 2019. Role of nuclear RNA in regulating chromatin structure and transcription. *Curr. Opin. Cell Biol.* 58:120–125. <https://doi.org/10.1016/j.ceb.2019.03.007>
- Nir, G., I. Farabella, C. Pérez Estrada, C.G. Ebeling, B.J. Beliveau, H.M. Sasaki, S.D. Lee, S.C. Nguyen, R.B. McCole, S. Chatteraj, et al. 2018. Walking



- along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet.* 14:e1007872. <https://doi.org/10.1371/journal.pgen.1007872>
- Nozawa, R.-S., L. Boteva, D.C. Soares, C. Naughton, A.R. Dun, A. Buckle, B. Ramsahoye, P.C. Bruton, R.S. Saleeb, M. Arnedo, et al. 2017. SAF-A Regulates Interphase Chromosome Structure through Oligomerization with Chromatin-Associated RNAs. *Cell.* 169:1214–1227.e18. <https://doi.org/10.1016/j.cell.2017.05.029>
- Ovesný, M., P. Krížek, J. Borkovec, Z. Švindrych, and G.M. Hagen. 2014. ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics.* 30: 2389–2390. <https://doi.org/10.1093/bioinformatics/btu202>
- Owen, D.M., C. Rentero, J. Rossy, A. Magenau, D. Williamson, M. Rodriguez, and K. Gaus. 2010. PALM imaging and cluster analysis of protein heterogeneity at the cell surface. *J. Biophotonics.* 3:446–454. <https://doi.org/10.1002/jbio.200900089>
- Prakash, K., D. Fournier, S. Redl, G. Best, M. Borsos, V.K. Tiwari, K. Tachibana-Konwalski, R.F. Ketting, S.H. Parekh, C. Cremer, and U.J. Birk. 2015. Superresolution imaging reveals structurally distinct periodic patterns of chromatin along pachytene chromosomes. *Proc. Natl. Acad. Sci. USA.* 112:14635–14640. <https://doi.org/10.1073/pnas.1516928112>
- Resch, G.P., K.N. Goldie, A. Krebs, A. Hoenger, and J.V. Small. 2002. Visualisation of the actin cytoskeleton by cryo-electron microscopy. *J. Cell Sci.* 115:1877–1882.
- Revyakin, A., C. Liu, R.H. Ebricht, and T.R. Strick. 2006. Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science.* 314:1139–1143. <https://doi.org/10.1126/science.1131398>
- Rogers, S.L., U. Wiedemann, N. Stuurman, and R.D. Vale. 2003. Molecular requirements for actin-based lamella formation in *Drosophila* S2 cells. *J. Cell Biol.* 162:1079–1088. <https://doi.org/10.1083/jcb.200303023>
- Sabari, B.R.A., A. Dall'Agnese, A. Boija, I.A. Klein, E.L. Coffey, K. Shrinivas, B.J. Abraham, N.M. Hannett, A.V. Zamudio, J.C. Manteiga, et al. 2018. Co-activator condensation at super-enhancers links phase separation and gene control. *Science.* 361:eaar3958. <https://doi.org/10.1126/science.aar3958>
- Sauer, M., and M. Heilemann. 2017. Single-Molecule Localization Microscopy in Eukaryotes. *Chem. Rev.* 117:7478–7509. <https://doi.org/10.1021/acs.chemrev.6b00667>
- Schermelleh, L., R. Heintzmann, and H. Leonhardt. 2010. A guide to super-resolution fluorescence microscopy. *J. Cell Biol.* 190:165–175. <https://doi.org/10.1083/jcb.201002018>
- Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, et al. 2012. Fiji: an open-source platform for biological-image analysis. *Nat. Methods.* 9: 676–682. <https://doi.org/10.1038/nmeth.2019>
- Sengupta, P., T. Jovanovic-Talman, D. Skoko, M. Renz, S.L. Veatch, and J. Lippincott-Schwartz. 2011. Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. *Nat. Methods.* 8: 969–975. <https://doi.org/10.1038/nmeth.1704>
- Sieberg, D., and D.-P. Herten. 2011. Fluorescence Quenching of Quantum Dots by DNA Nucleotides and Amino Acids. *Aust. J. Chem.* 64:512–516. <https://doi.org/10.1071/CH10293>
- Smeets, D., Y. Markaki, V.J. Schmid, F. Kraus, A. Tattermusch, A. Cerase, M. Sterr, S. Fiedler, J. Demmerle, J. Popken, et al. 2014. Three-dimensional super-resolution microscopy of the inactive X chromosome territory reveals a collapse of its active nuclear compartment harboring distinct Xist RNA foci. *Epigenetics Chromatin.* 7:8. <https://doi.org/10.1186/1756-8935-7-8>
- Spahn, C., F. Herrmannsdörfer, T. Kuner, and M. Heilemann. 2016. Temporal accumulation analysis provides simplified artifact-free analysis of membrane-protein nanoclusters. *Nat. Methods.* 13:963–964. <https://doi.org/10.1038/nmeth.4065>
- Strom, A.R., A.V. Emelyanov, M. Mir, D.V. Fyodorov, X. Darzacq, and G.H. Karpen. 2017. Phase separation drives heterochromatin domain formation. *Nature.* 547:241–245. <https://doi.org/10.1038/nature22989>
- Szabo, Q., D. Jost, J.M. Chang, D.I. Cattoni, G.L. Papadopoulos, B. Bonev, T. Sexton, J. Gurgo, C. Jacquier, M. Nollmann, et al. 2018. TADs are 3D structural units of higher-order chromosome organization in *Drosophila*. *Sci. Adv.* 4:eaar8082. <https://doi.org/10.1126/sciadv.aar8082>
- van de Linde, S., and M. Sauer. 2014. How to switch a fluorophore: from undesired blinking to controlled photoswitching. *Chem. Soc. Rev.* 43: 1076–1087. <https://doi.org/10.1039/C3CS60195A>
- van de Linde, S., A. Löschberger, T. Klein, M. Heidbreder, S. Wolter, M. Heilemann, and M. Sauer. 2011. Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nat. Protoc.* 6:991–1009. <https://doi.org/10.1038/nprot.2011.336>
- Wang, S., J.H. Su, B.J. Beliveau, B. Bintu, J.R. Moffitt, C.T. Wu, and X. Zhuang. 2016. Spatial organization of chromatin domains and compartments in single chromosomes. *Science.* 353:598–602. <https://doi.org/10.1126/science.aaf8084>
- Williamson, D.J., G.L. Burn, S. Simoncelli, J. Griffié, R. Peters, D.M. Davis, and D.M. Owen. 2020. Machine learning for cluster analysis of localization microscopy data. *Nat. Commun.* 11:1493. <https://doi.org/10.1038/s41467-020-15293-x>
- Xiao, R., P. Tang, B. Yang, J. Huang, Y. Zhou, C. Shao, H. Li, H. Sun, Y. Zhang, and X.-D. Fu. 2012. Nuclear matrix factor hnRNP U/SAF-A exerts a global control of alternative splicing by regulating U2 snRNP maturation. *Mol. Cell.* 45:656–668. <https://doi.org/10.1016/j.molcel.2012.01.009>
- Xie, S.Q., S. Martin, P.V. Guillot, D.L. Bentley, and A. Pombo. 2006. Splicing speckles are not reservoirs of RNA polymerase II, but contain an inactive form, phosphorylated on serine2 residues of the C-terminal domain. *Mol. Biol. Cell.* 17:1723–1733. <https://doi.org/10.1091/mbc.e05-08-0726>
- Xie, L., P. Dong, X. Chen, T.S. Hsieh, S. Banala, M. De Marzio, B.P. English, Y. Qi, S.K. Jung, K.-R. Kieffer-Kwon, et al. 2020. 3D ATAC-PALM: super-resolution imaging of the accessible genome. *Nat. Methods.* 17:430–436. <https://doi.org/10.1038/s41592-020-0775-2>

## Supplemental material

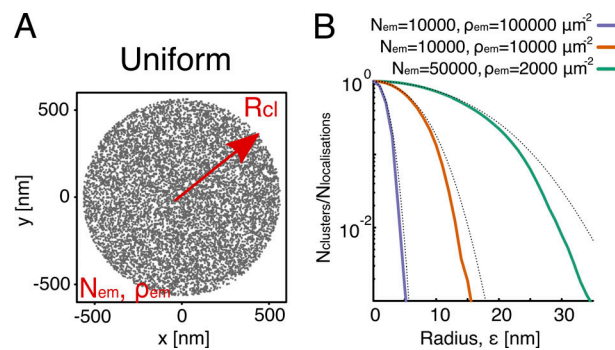
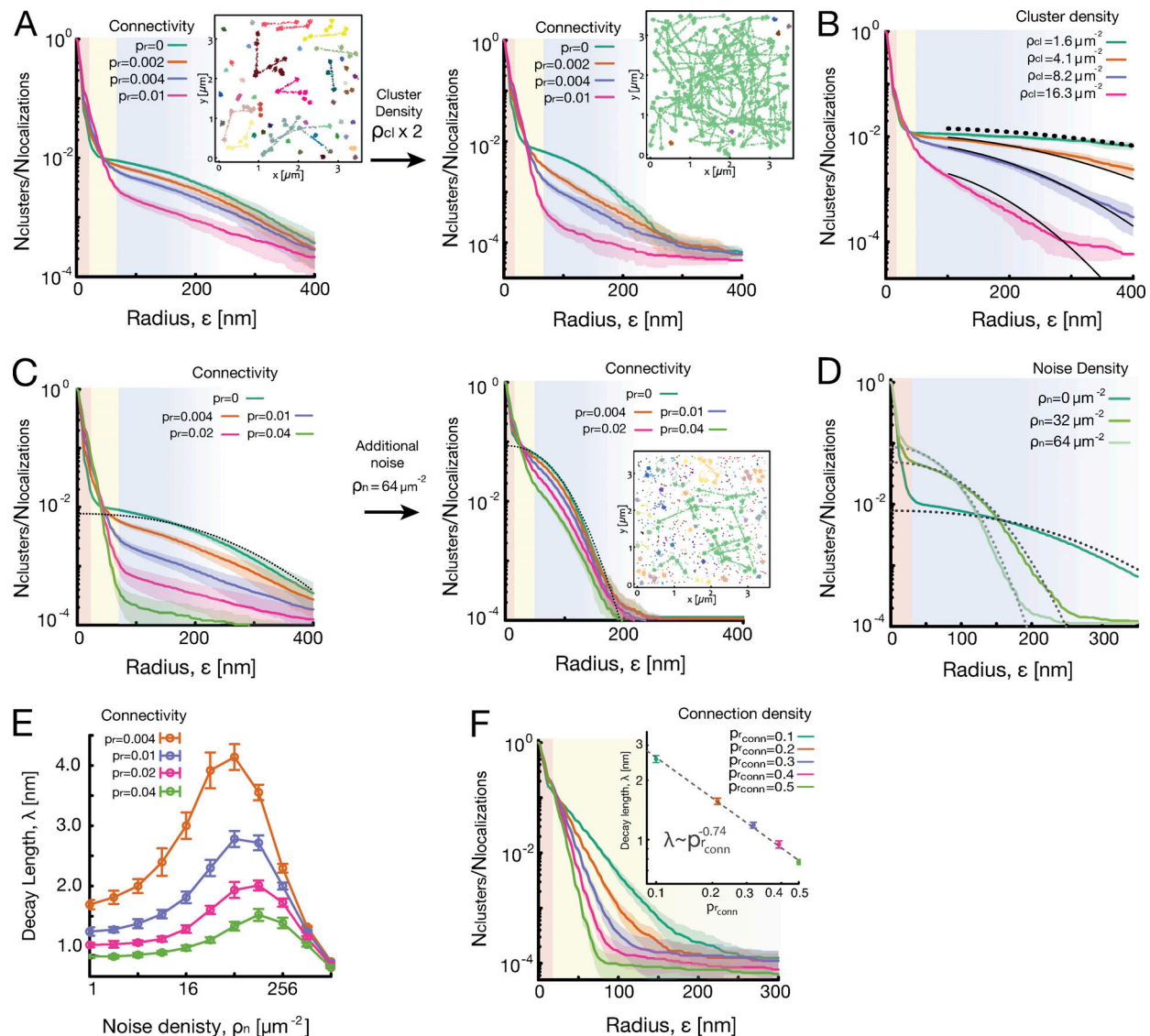
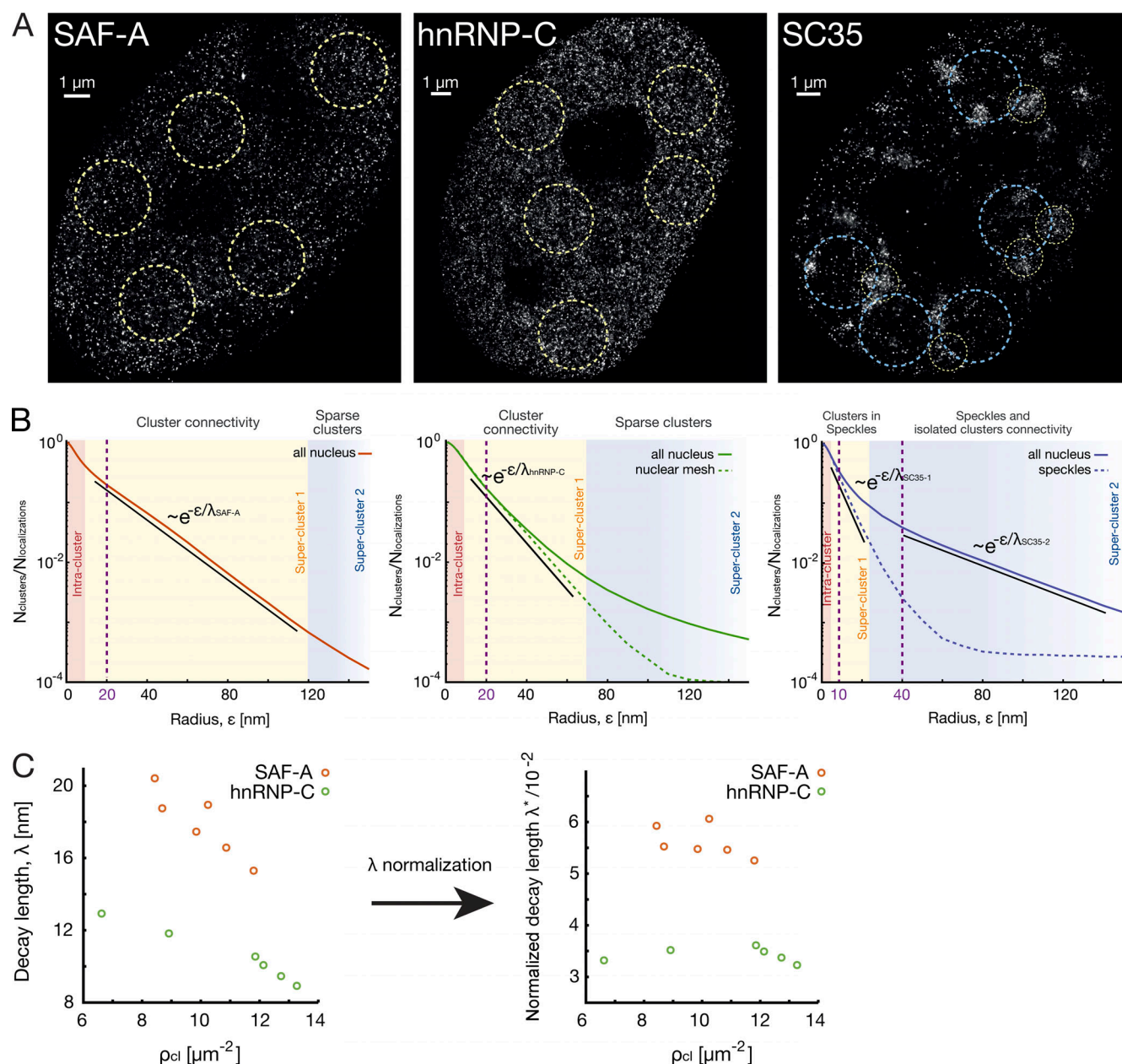


Figure S1. **The Poissonian functional form in the intra-cluster regime.** (A) To test the Poissonian functional form (Eq. 1) of the intra-cluster regime of SuperStructure curves, we simulated localizations inside clusters as a uniform distribution of  $N_{em}$  points distributed within a circle of radius  $R_{cl}$ . The resulting average density is  $\rho_{em}$ . The number of points included in any circular subregion of radius  $\epsilon$  is, on average,  $n(\epsilon) = \pi \rho_{em} \epsilon^2$ , and is in fact itself Poisson distributed. (B) To check the theoretical prediction of Eq. 1, we have created simulated datasets for various  $\rho_{em}$  and  $N_{em}$ . The theoretical predictions (dotted lines) with  $m = 2$  are in good agreement with the SuperStructure curves, indicating that indeed Eq. 1 correctly captures the behavior of uniformly distributed points forming one idealized cluster. However, note that for  $m = 2$ , there is already an overcounting of clusters at large values of  $\epsilon$  due to the fact that DBSCAN merges indirectly related emitters in a single big cluster. This suggests not to extend the summation to higher values of  $m$ . From Eq. 1, the end of the intra-cluster regime can be approximated by the width of the Poisson function, i.e.,  $\epsilon^* \simeq 3\kappa_0$  (at 99% confidence level), where  $\kappa_0 = 1/\sqrt{\pi\rho_{em}}$  is the decay length identified by Eq. 1. This is confirmed by observing that predicted  $\epsilon^*$  for the curves are  $\epsilon^*(\rho_{em} = 2,000 \mu m^{-2}) \simeq 38 \text{ nm}$ ,  $\epsilon^*(\rho_{em} = 10,000 \mu m^{-2}) \simeq 18 \text{ nm}$ , and  $\epsilon^*(\rho_{em} = 100,000 \mu m^{-2}) \simeq 5.3 \text{ nm}$ , which correspond to  $N_{cl}/N_{em} \simeq 10^{-3}$  (when most of the points have been merged in a single cluster).

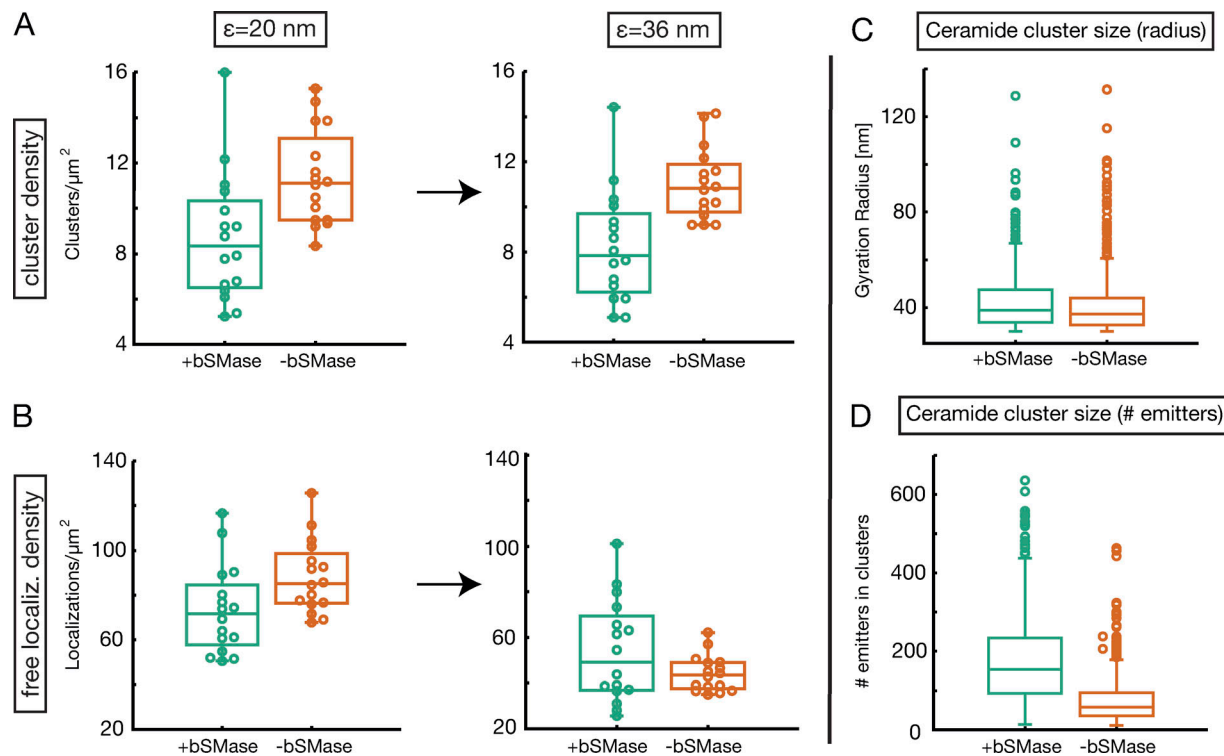


**Figure S2. Average SuperStructure curves for simulated datasets in different conditions.** SuperStructure analysis was run on 20 independent datasets (each in the same condition), and the resulting curves were then averaged. Shaded regions represent the standard deviation from the average. Parameters are set to their standard values if not otherwise specified (see Materials and methods). Palettes in the inset configurations represent cluster analysis at  $\epsilon = 80 \text{ nm}$ . **(A)** Locally connected clusters with different grades of connectivity and doubling the cluster density (from left to right):  $\rho_{cl} = 8.2 \mu\text{m}^{-2}$  (left) and  $\rho_{cl} = 16.3 \mu\text{m}^{-2}$  (right), connection density  $p_{conn} = 0.5$ , and no noise and different values of connectivity  $p_r$ . The higher cluster density makes SuperStructure curves more markedly distinct as a function of  $p_r$  compared with the same curves for a lower density. **(B)** Locally connected clusters with low connectivity and increasing cluster density: connectivity  $p_r = 0.002$ , connection density  $p_{conn} = 0.5$ , and no noise and different cluster densities  $\rho_{cl}$ . The first super-cluster regime maintains the single exponential decay, but the decay length  $\lambda$  decreases with the cluster density. In the main text, we showed that this dependence goes as  $\lambda \propto \rho_{cl}^{-1/2}$ . Also, the exponential decay  $\lambda_2$  of the second super-cluster regime decreases with the density of clusters, and this regime evolves from a Poisson-like (low  $\rho_{cl}$ ) to an exponential decay (high  $\rho_{cl}$ ). This behavior seems to be a pure effect of the cluster density, as all other parameters remain unchanged. Black curves are Poisson decays attempts  $\sim e^{-\pi \rho_{cl} \epsilon^2}$  to fit the second super-cluster regime. **(C)** Locally connected clusters with different grades of connectivity and sparse noise addition: cluster density  $\rho_{cl} = 8.2 \mu\text{m}^{-2}$ , connection density  $p_{conn} = 0.5$ , noise density  $\rho_n = 0 \mu\text{m}^{-2}$  (left)/ $\rho_n = 64 \mu\text{m}^{-2}$  (right), and different values of connectivity  $p_r$ . With high noise (eight times the cluster density), the second super-cluster regime becomes Poissonian; the first super-cluster regime maintains its typical exponential decay, but the decay length is altered. Dotted lines represent fit with Eq. 3 for  $\epsilon \in [70 : 300] \text{ nm}$ . **(D)** Unconnected clusters of points with increasing density of noise (other parameters are the same as C). Eq. 3 well describes the decay of the curves in the intercluster regime, with the density parameter  $\rho_{cl}$  and  $\rho_{cl} + \rho_n$ , respectively, in absence and presence of noise. **(E)** Average decay length of the first super-cluster regime for the connected systems represented in C as function of noise density  $\rho_n$ . The fit to calculate the decay length  $\lambda$  has been made for  $\epsilon \in [20, 60] \text{ nm}$  for 20 independent datasets. Values of  $\lambda$  are then averaged. Bars represent the standard deviation from the average. Decay lengths for systems with different connectivities  $p_r$  are distinguishable as long as the noise density is below the connection density ( $\sim 500 \mu\text{m}^{-2}$ ). However, low noise density also alters the estimation of the decay length. The alteration is less severe for highly connected clusters. **(F)** Fully connected meshes of clusters with increasing density of the mesh: cluster density  $\rho_{cl} = 8.2 \mu\text{m}^{-2}$ , connectivity  $p = 0.025$ , and no noise and different values of connection density  $p_{conn}$ . The super-cluster regime is unique, the decay is exponential, and the decay length  $\lambda$  decreases with the density of the mesh. Fit of  $\lambda$  was performed for  $\epsilon \in [20 : 60] \text{ nm}$ . The inset shows the dependence of  $\lambda$  on  $p_{conn}$  in a fully connected mesh, which is  $\lambda \sim p_{conn}^{-0.74}$ .

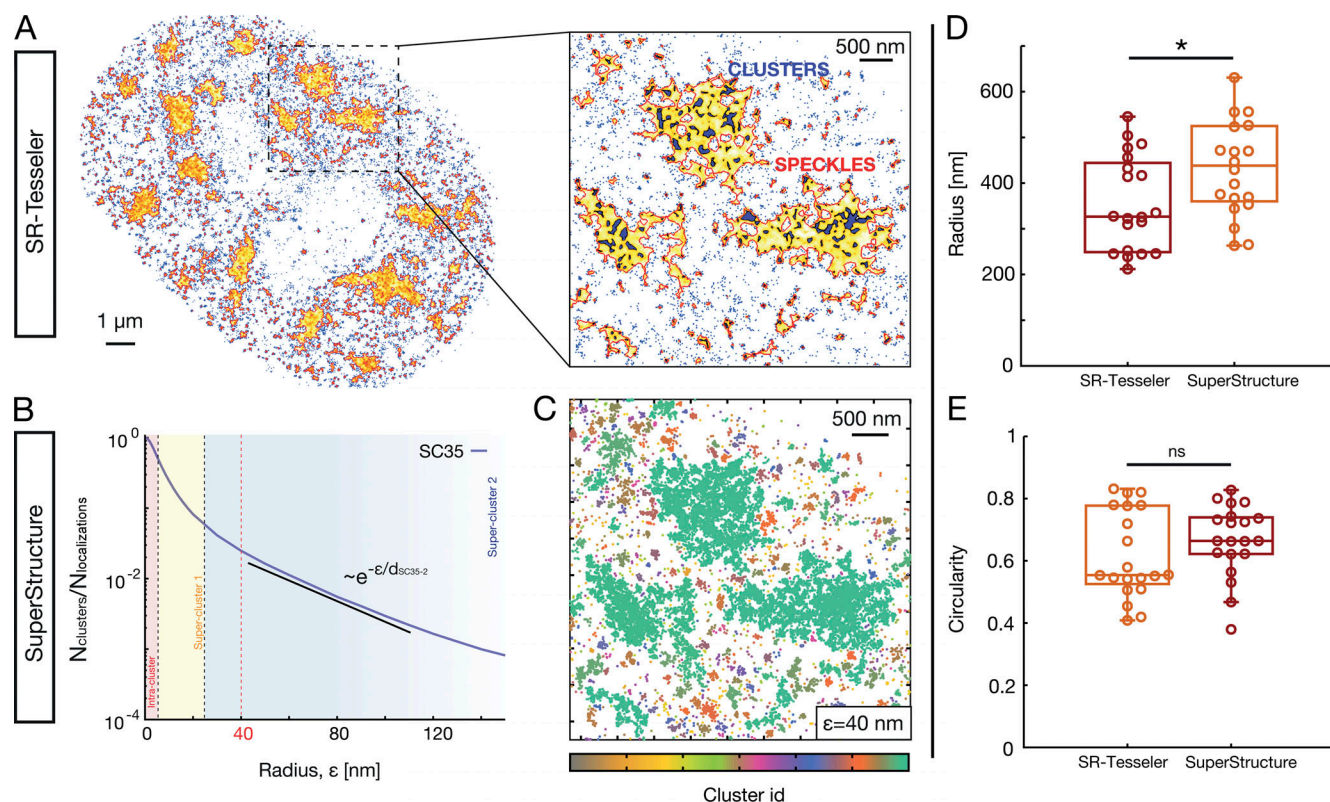




**Figure S3. Details on  $\lambda$  normalisation and proof that connections are not technical artifacts in nuclear protein data.** (A) dSTORM reconstructed images of SAF-A, hnRNP-C, and SC35 in a single cell where local circular regions for cluster density estimation purpose are highlighted. In the case of SC35, two different region types are used, one inside speckles for the first exponential regime and one outside speckles for the second exponential regime. In the case of hnRNP-C and SC35, local circular regions were also used to compute SuperStructure local curves and the decay length  $\lambda$  in the first super-cluster regime, as explained in Materials and methods. (B) Average SuperStructure curves for SAF-A, hnRNP-C, and SC35 are shown and explained in the main text. Solid lines are the result of all-nucleus analysis, while dashed lines are the result of a local analysis (in local circular regions). Exponential regimes of interest are highlighted, as well as the values of  $\epsilon$  at which the cluster analysis is made for cluster density estimation purposes (purple dashed vertical line). (C) Check that connections are not the result of technical artifacts due to bad blinking quality both in SAF-A and hnRNP-C data by monitoring  $\lambda$  (left) and  $\lambda^*$  (right) for different cluster densities  $\rho_{\text{cl}}$ . The bad blinking quality of fluorophores would lead to localization inaccuracy of emitters at the borders of protein clusters, and this in turn could lead to pseudo-connections between clusters. However, these pseudo-connections would be proportional to the cluster density; a higher cluster density would result in stronger pseudo-connections, which would reflect to a decrease of  $\lambda^*$  with the cluster density.  $\lambda$ ,  $\rho_{\text{cl}}$ , and  $\lambda^*$  were calculated for the six independent nuclei, as explained in Materials and methods, and are shown in Table S1. Every nucleus can be considered as a system where the blinking conditions are the same, but cluster densities may vary due to statistical fluctuations. While  $\lambda$  (left) decreases with  $\rho_{\text{cl}}$ , as expected,  $\lambda^*$  (right) is constant for different densities, ruling out the hypothesis that connections are artifacts due to bad blinking quality.



**Figure S4. Absence of local connectivity and confirmation of original paper results in ceramide data. (A and B)** The absence of local connectivity was confirmed by analyzing cluster density (A) and sparse localization density (B) in the crossover range. We monitored the density of ceramides clusters and that of free emitters at  $\epsilon_1 = 20$  nm and  $\epsilon_2 = 36$  nm. To calculate cluster density, DBSCAN was run at  $N_{min} = 0$  and at the given value of  $\epsilon$ , and we kept only clusters with at least 10 particles. The remaining particles were considered as free localizations. Clusters and free localizations were detected at  $N_{min} = 0$  for 16 independent circular regions. The number of clusters remains constant in the considered  $\epsilon$  regime, while the free localizations density significantly decreases (more severely for -bSMase cells). As a consequence, we can state that there is no significant merging of ceramide clusters, only embedding of nearby free localizations in already-formed clusters. **(C and D)** Confirmation of the original paper's results by calculating the ceramide cluster size both as gyration radius (C) and number of emitters (D). Protein clusters were detected at  $N_{min} = 0$  at  $\epsilon^+ = 20$  nm and  $\epsilon^- = 24$  nm. In accordance with the analysis in the paper, we looked at the size of clusters with a radius  $>30$  nm. Note that +bSMase ceramide clusters consist of (on average) 180 emitters in a circle of radius 42 nm. The resulting density is  $32,500 \mu\text{m}^{-2}$ . This result is approximately in line with our prediction obtained with the Poisson intra-cluster fit by considering that the standard deviation of both cluster radius and emitters is high. Similarly, -bSMase clusters have on average 78 emitters in an average cluster radius of 40 nm. The resulting density is  $15,500 \mu\text{m}^{-2}$ .



**Figure S5. Size and shape estimation of local super-structures emerging in SC35 dSTORM data (i.e., nuclear speckles) by using both SuperStructure and SR-Tesseler.** Analysis was performed on a single cell as proof of concept. **(A)** Super-structure detection using SR-Tesseler software, a segmentation framework based on Voronoï tessellation (constructed from the localization coordinates). Adjustments of the density factor allows the detection of structures at different density levels, such as clusters (violet) or speckles (yellow). Blue dots represent no-segmented localizations. The software was downloaded from <https://github.com/fleivet/SR-Tesseler/releases/tag/v1.0> and run on a Windows operating system. **(B)** SuperStructure curve of the same data. Analysis of decay regimes allows the identification of  $\epsilon = 40$  nm as a suitable value for super-structure identifications. **(C)** Identified clusters at  $\epsilon = 40$  nm with SuperStructure. Speckle detections are visually compatible with those of SR-Tesseler. **(D and E)** Radius and circularity of super-structures using both SR-Tesseler and SuperStructure. Both radius and circularity are very similar, showing the power of SuperStructure in computing shape and size properties. In the analysis, we considered the 20 largest identified structures (i.e., speckles). For SuperStructure, the 2D symmetric gyration tensor  $\vec{R}^2$  was computed and diagonalized for identified super-structures. The gyration tensor components  $R_{xy}^2$  are defined as  $R_{xy}^2 = \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N (x_i - x_j)(y_i - y_j)$ , where  $N$  is the total number of localizations in a superstructure, and  $x_i$  and  $y_i$  are the  $x$  and  $y$  positions of the localization  $i$ . The diagonalization is necessary to obtain the square of the major and minor semi-axis of the speckles, namely  $y_1$  and  $y_2$ . We then calculated the speckle radius  $R_g = \sqrt{y_1 + y_2}$  and their circularity  $c = \sqrt{\frac{|y_1 - y_2|}{y_1 + y_2}}$ . For SR-Tesseler, radius and circularity parameters were obtained as output after Voronoï tessellation. P values were calculated using a Student's  $t$  test: ns,  $P > 0.05$ ; \*,  $P < 0.05$ .

**Table S1 is provided online and shows the decay length, detected cluster density, and normalized decay length for SAF-A, hnRNP-C, and SC-35 in both super-cluster regimes (SC35-1 and SC35-2).**