Computational and Structural Biotechnology Journal 21 (2023) 879–888

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Short review

Recent development of computational cluster analysis methods for single-molecule localization microscopy images

Yoonsuk Hyun [a], Doory Kim [b,c,d,e,*]

[a] Department of Mathematics, Inha University, Republic of Korea
[b] Department of Chemistry, Hanyang University, Republic of Korea
[c] Research Institute for Convergence of Basic Science, Hanyang University, Republic of Korea
[d] Institute of Nano Science and Technology, Hanyang University, Republic of Korea
[e] Research Institute for Natural Sciences, Hanyang University, Republic of Korea

ARTICLE INFO

Article history:
Received 15 November 2022
Received in revised form 7 January 2023
Accepted 7 January 2023
Available online 9 January 2023

Keywords:
Super-resolution fluorescence microscopy
Single-molecule localization microscopy
Cluster analysis
Machine learning

ABSTRACT

With the development of super-resolution imaging techniques, it is crucial to understand protein structure at the nanoscale in terms of clustering and organization in a cell. However, cluster analysis from single-molecule localization microscopy (SMLM) images remains challenging because the classical computational cluster analysis methods developed for conventional microscopy images do not apply to pointillism SMLM data, necessitating the development of distinct methods for cluster analysis from SMLM images. In this review, we discuss the development of computational cluster analysis methods for SMLM images by categorizing them into classical and machine-learning-based methods. Finally, we address possible future directions for machine learning-based cluster analysis methods for SMLM data.

1. Introduction

Understanding protein structure in terms of clustering and organization in a cell is imperative because it is known to be closely related to its function in the cell [1]. Identifying protein clusters on a molecular scale is important to precisely determine the protein-to-protein interactions, but it has been limited to conventional diffraction-limited optical microscopy [2]. However, it has become possible to map high-resolution protein structures in a cell on a molecular scale using recently developed super-resolution fluorescence microscopy (SRM) [3].

The past decade has witnessed the development of SRM techniques that overcome the diffraction-based far-field resolution limits of conventional light microscopy [4]. Various SRM techniques have been developed by several research groups, and they can be roughly categorized into two groups: one group employs the engineering of illumination patterns, and the other group uses single-molecule localization. The representative methods for the first group are stimulated emission depletion (STED) microscopy [5] and structured illumination microscopy (SIM) [6]. Briefly, STED sharpens the point-spread function (PSF) by employing an additional donut-shaped depletion beam, thereby improving the resolution of the microscope. In SIM, a sample is subjected to a structured illumination pattern with a known high spatial frequency in order to shift the sub-resolution structure of an unknown sample to a detectable lower frequency and thus restore the nanostructure. Representative single-molecule localization microscopy (SMLM) methods include stochastic optical reconstruction microscopy (STORM) [7] and (fluorescence) photoactivation localization microscopy ((F)PALM) [8]. These methods can achieve single-molecule localization with high precision by temporally separating the activation of individual fluorescent emitters to avoid their overlapping based on their stochastic "on–off" fluorescence photoswitching [9]. This high-resolution image reveals a pointillistic nature and requires a new method for cluster analysis because an SMLM image is reconstructed from individually localized points obtained by the detection and localization of single fluorescent molecules [10].

The analysis method for SMLM data is expected to differ from conventional cluster analysis methods for intensity grid-valued pixel-based images obtained from conventional microscopy because it is known that the computational cluster analysis methods developed for conventional microscopy images do not apply to pointillism SMLM data [1]. Therefore, in this review, we focus on the development of computational cluster analysis methods for SMLM images by

* Corresponding author at: Department of Chemistry, Hanyang University, Republic of Korea.
E-mail address: doorykim@hanyang.ac.kr (D. Kim).

https://doi.org/10.1016/j.csbj.2023.01.006

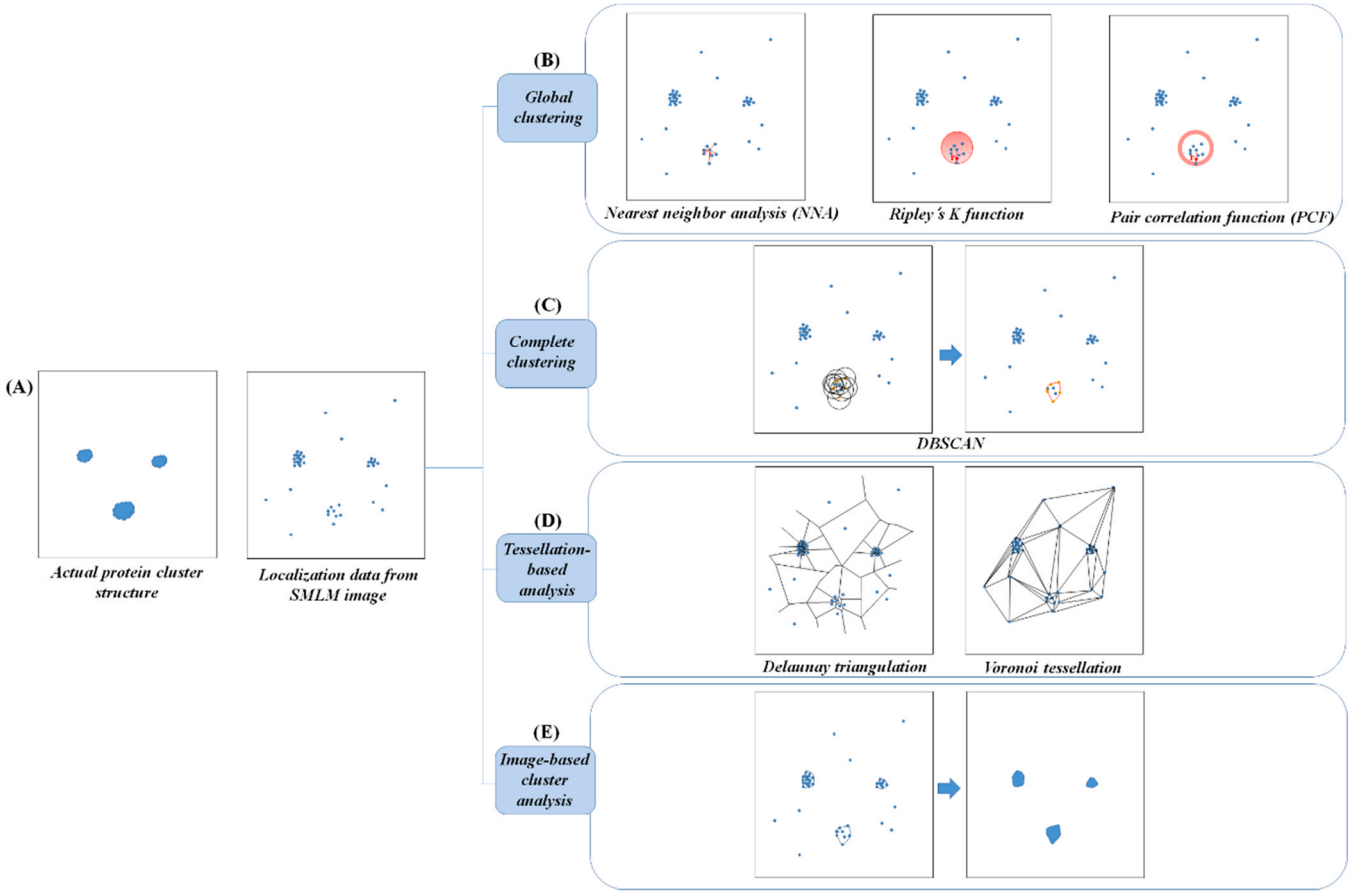

Fig. 1. Classical cluster analysis methods for single-molecule localization microscopy (SMLM) images. (A) Pointillism localization data generated from SMLM. (B) Global clustering. (C) Complete clustering. (D) Tessellation-based methods. (E) Image-based cluster analysis.

categorizing them into classical methods and machine learning-based methods.

2. Classical cluster analysis methods for SMLM images

Classical cluster analysis methods for SMLM images can be categorized into four groups: 1) global clustering, 2) complete clustering, 3) tessellation-based methods, and 4) image-based cluster analysis (Fig. 1, Table 1).

2.1. Global clustering approach

The global clustering analysis method provides a global description of protein clustering or organization by providing spatial statistics. This ensemble method includes nearest neighbor analysis (NNA), Ripley's K function, and pair correlation function (PCF).

NNA is the simplest metric for determining the level of clustering within data by calculating the distances between each molecule's position and its neighbor. The level of clustering within the data can be determined by comparing these distances with the calculated distances from a random distribution of molecules. For example, the mean nearest-neighbor distance for a clustered distribution is significantly lower than that for a random distribution. Bar-On et al. used NNA to analyze the distribution of syntaxin molecules in the cell membrane [11]. From the analysis, they could observe the non-homogeneously distributed single syntaxin molecules in the membrane and the concentrated molecules in the area adjacent to clusters.

A similar analysis is Ripley's K function, which calculates the average number of neighboring molecules around a single molecule within a given radius [12]. This calculation is repeated by increasing the radius, and the calculation result is compared with the expected values for a random distribution to determine the average cluster radius. This approach was used to quantify the clustering of membrane receptors, including tyrosine kinase Lck and EGFR, from SMLM images [13,14].

However, Ripley's K function-based methods have several issues, including the requirement of calibration data that can strongly affect the output, no consideration of the individual localization precisions, and the limited judgment of their performance [15]. These shortcomings can be overcome by a model-based Bayesian approach recently developed by Rubin-Delanchy et al. [15]. By completely taking the individual localization precisions for each emitter into account, this Bayesian approach enables accurate quantification of clustering behavior in SMLM images.

Another ensemble method is PCF, which is calculated to obtain the probability of finding localizations at a given distance from another localization. This method can provide not only great sensitivity to the changes in molecular clustering but also a correction to the overcounting, which results from multiple blinking of the same probe, by comparing the experimental data with the completely randomly distributed data [12,16,17]. Hartley et al. used PCF to determine the density of hybrid nanoconjugates crosslinked CD20 on the surface of malignant B cells [18]. The cluster size and the average number of molecules in the clusters were analyzed in their study to correlate with apoptosis induction after the treatment with the nanoconjugates. Such a global approach allows us to understand the level of organization or clustering; however, it is an ensemble method that provides limited statistical information on the data.

Table 1
Comparison of reported studies on the cluster analysis methods for single-molecule localization microscopy images.

Type	Method	Algorithm	Data	Target	Study	Ref.
Classical cluster analysis	Global clustering	NNA	dSTORM	Syntaxin	Syntaxin membranal clusters	[11]
		Ripley's K function	dSTORM, PALM	Tyrosine kinase Lck molecules	Clustering analysis from Lck distributions	[13]
			dSTORM	EGFR membrane clustering	Spatial distribution of inactive EGFR proteins on different membrane surfaces	[14]
		Bayesian approach	PALM	CD3ζ	Clustering behavior of CD3ζ, a subunit of the CD3 T cell receptor complex, in resting and activated primary human T cells	[15]
		PCF	PALM	Purified PAGFP, TfR-PAGFP, Lat-mEGFP, VSVG-mEGFP	Plasma-membrane proteins with different membrane anchoring and lipid partitioning characteristics	[16,17]
			dSTORM	CD20	Organizational changes in the plasma membrane after treatment with hybrid nanoconjugates	[18]
	Complete clustering	DBSCAN	PALM, dSTORM	T-cell receptor–CD3 complexes	Detection of TCR–CD3 complexes for quantification of their organization	[19]
			PALM, STORM	Clathrin, LFA, GPI, Lck	Cluster analysis of different proteins (clathrin, LFA, GPI, Lck) in the cellular plasma membrane	[20]
			dSTORM, PALM	Dopamine transporter (DAT)	The molecular organization of the dopamine transporter in cholesterol and neuronal activity-dependent nanodomains	[21]
		SuperStructure	dSTORM	Nuclear proteins (hnRNPs, SAF-A, SC35)	Development of SuperStructure as a parameter-free algorithm	[22]
		FOCAL	dSTORM, PALM	Clusters of eukaryotic RNAP II and the bacterial protein H-NS	Development of a grid-based clustering algorithm FOCAL	[23]
		ToMATo	dSTORM	Receptor clustering in platelets, nuclear pore components, endocytic proteins and microtubule networks	Development of ToMATo as a persistence-based clustering method	[24]
	Tessellation-based analysis	Delaunay triangulation	dSTORM, PALM	Mitochondrion	3D image processing of nucleoid clusters	[27]
		Voronoi tessellation	PALM	GluA1, integrin-β3, tubulin, GlyR, GPI	Development of SR-Tesseler software as a segmentation framework based on Voronoï tessellation	[28]
			STORM	Simulated SMLM images of microtubule and clusters, nucleoporin protein TPR	Development of ClusterViSu, a method for image reconstruction, visualization and quantification of labeled protein clusters, based on Voronoi tessellation	[29]
			STORM	Tubulin, nucleopore protein TPR	Development of SharpViSu as an interactive open-source software for performing processing steps for localization data in an integrated manner	[30]
			dSTORM	DHPR, RyR2, PLN, and SERCA2A	Analysis of changes in cluster properties of target sarcoplasmic reticulum proteins	[31]
	Image-based cluster analysis	Otsu's thresholding algorithm 8-point connectivity	STORM	Purinosome, mitochondria	Analysis of purinosome-mitochondria colocalization upon mTOR inhibition	[32]
			STORM	Purinosome, mitochondria, microtubule	Association of purinosomes with mitochondria and microtubules	[33]
			STORM	Xist RNA-PRC2 complex	Identification of Xist cloud and the determination of the nearest Xi marker localizations	[34]
			STORM	Actin, microtubules, DTS, mitochondria, autophagosome, granules	Distribution changes of various organelles in platelet during activation process	[35]
			STORM	Actin, microtubules, DTS, mitochondria, autophagosome, granules	Distribution changes of various organelles in platelet during platelet progeny production	[36]
			SR-STORM	Polymer blend film	Identification of boundary of phase separated patterns	[37]
		KDE	dSTORM	RAD51 and DMC1 foci on the synaptonemal complex axes	Investigation of the spatial organization of recombinase RAD51 and its meiosis-specific paralog DMC1 during meiotic DNA double-strand break repair	[38]

(*continued on next page*)

Table 1 (*continued*)

Type	Method	Algorithm	Data	Target	Study	Ref.
Machine learning based cluster analysis	Methods for constructing clustered data	3D scene reconstruction using convolutional neural network	STORM	CEP 152 Complex	Reconstructing 3D clusters from 2D SMLM image	[44]
		K-means clustering	dSTORM	HER2 molecules	Quantifying HER2 molecules in clusters	[45]
		Regression using fully connected, convolution and LSTM layers.	dSTORM	C-terminal Src kinase (Csk)	Clustering algorithm of Csk points	[46]
		Support vector machine classifier	STORM	Proteins within centrioles and procentrioles in KE 37 cells	Point classification and cluster partitioning	[48]
	Method for cluster identification and classification	K-means clustering	GSD SMLM	Cav1, Cavin-1 in PC3 cells	Clustering analysis of Cav1.	[49]
		Random forest classifier, K-means clustering	GSD SMLM	Cav1, Cavin-1 in PC3 cells	Blob identification on SMLM images of Cav1.	[50]
		Random forest, convolutional neural net based, and PointNet[43] based classifier.	GSD SMLM	Cav1, Cavin-1 in PC3 cells	Detection of caveolae and scaffolds.	[51]

NNA: nearest neighbor analysis; PCF: pair correlation function; DBSCAN: density-based spatial clustering of applications with noise; FOCAL: fast optimized cluster algorithm for localizations; ToMATo: topological mode analysis tool; KDE: kernel density estimation; GSD SMLM: ground state depletion single-molecule localization microscopy.

2.2. Complete clustering approach

A complete clustering approach can overcome the limitation of a global approach by extracting rich information from the data at the single-cluster level, such as the number and shape of individual clusters. It includes density-based spatial clustering of applications with noise (DBSCAN) and SuperStructure.

First, DBSCAN is the most commonly used complete clustering method that directly groups localizations into discrete clusters. It can not only determine a cluster but also identify noise by calculating the number of points within a given neighborhood radius. Given a minimum number of localizations as a threshold and a radius of search as an input, if there are more localizations than the threshold within a given radius of search, it is classified as a cluster, whereas it is classified as noise if there are a smaller number of localizations within a given radius of the search. The identified clusters were further categorized into several groups based on their size and the number of localizations. This method was applied to differentiate the phosphorylated from the nonphosphorylated T-cell receptor (TCR)-CD3 clusters in intact T cells, demonstrating that the molecular density within a TCR-CD3 cluster can determine signal initiation [19]. However, a cluster generated from multiple blinks could result in cluster artifacts by overcounting the multiple blinks of a fluorophore, which was not easily differentiable from a real cluster consisting of protein assemblies in this method. Baumgart et al. overcame this limitation by varying the labeling density, such as in the titration of fluorescent antibodies, to distinguish random from clustered distributions of molecules, which is insensitive to the blinking statistics of the used fluorophores [20]. Nanodomain patterns for the dopamine transporter (DAT) and nanocluster formation in resting and activated immune cells were analyzed based on this method by ruling out multiple observations of single fluorophores [20,21].

Another complete clustering approach is SuperStructure, which is an extended DBSCAN by detecting the number of localizations within a neighborhood radius [22]. However, in contrast to DBSCAN, SuperStructure is a parameter-free detection and quantification method that extracts connectivity information from the change rate ($N_c(\varepsilon)$ curve) of the number of localizations with the neighborhood radius, which is overlooked in DBSCAN. This method was demonstrated to investigate protein organization, such as nuclear proteins involved in RNA processing and ceramide membrane lipids involved in cellular trafficking [22].

Another complete clustering analysis method to overcome the limitations of DBSCAN is fast optimized cluster algorithm for localizations (FOCAL) [23]. Because DBSCAN scales with the number of localizations such as O(n log (n)), it can be slow for analyzing dense structures. Additionally, the choice of parameters strongly affects performance. These limitations can be overcome using the grid-based clustering algorithm, FOCAL. In contrast to DBSCAN, FOCAL scales such as O(n) have only one set parameter, allowing a fast and efficient analysis. Additionally, FOCAL is effective for filtering out focus clusters, which could increase the local background and deteriorate localization precision [23].

Recently, a persistence-based clustering method has also been developed to overcome the limitations of density-based clustering analysis with a single density threshold. For complex biological structures with varying densities, a single-density threshold is insufficient, and persistence can be utilized to identify clusters [24]. By employing both persistence-based clustering [25] and persistent homology [26], Pike et al. developed the topological mode analysis tool (ToMATo) to quantify complex biological nanostructure [24]. By demonstrating this method to various biological structures, including receptor clustering in platelets, nuclear pore components, endocytic proteins, and microtubule networks, the authors reported

that their method outperforms existing approaches, including DBSCAN [24].

2.3. Tessellation-based method

The next category of cluster analysis for SMLM images is a tessellation-based method that includes Delaunay triangulation and Voronoi tessellation. These examples are generally used in mathematics and computational geometry. This method creates a tessellated surface in which tiles are generated from localizations to determine the presence of clusters.

First, the Delaunay triangulation method creates triangular tiles in which the localizations comprise the corners or vertices [12]. In this approach, the process of organizing localizations in a triangular mesh is repeated until all triangles in the mesh satisfy the Delaunay criterion, which specifies a rule to determine a favorable representation of the spatial relationships between points. It can not only determine the highly clustered points in the tile area but also separate noise by defining the lowest density of points forming the cluster. This method was successfully used to segment individual DNA nucleoid clusters from the 3D STORM data of the mitochondrial cristae [27].

In contrast to Delaunay triangulation, Voronoi creates polygonal regions called Voronoi cells, in which molecules are localized at the center of the tiles according to the Euclidean distance between points [1]. There is no intersection between Voronoi cells when the equidistant Voronoi edges are generated from the two nearest molecules. Such segmentation of molecular clusters in an SMLM image can be used to describe the density and organization of molecules [1]. This method has been recently implemented in open-source software, such as SR-Tesseler [28], ClusterVisu [29], and SharpViSu [30]. For example, SR-Tesseler was recently used to quantify clustering characteristics, such as cluster area, localization number per cluster, and cluster density, for Ca^{2+} handling regulators, including dihydropyridine receptor (DHPR), ryanodine receptor 2 (RyR2), phospholamban (PLN), and sarco/endoplasmic reticulum Ca^{2+}-ATPase 2A (SERCA2A) in TAC hearts [31]. It should be noted that localizations on the border of a cluster can be excluded, and the multiple localizations from a blinking fluorophore cannot be differentiable from the true molecular cluster since such a tessellation-based method can determine the presence of highly clustered points by the tile area [1].

2.4. Image-based cluster analysis

The aforementioned approaches, including global clustering, complete clustering, and tessellation-based methods, are based on the coordinates of localizations; hence, noise localizations can be unavoidably included for cluster analysis, interrupting the identification of actual molecular clusters. Although they can be filtered out based on their density, they can be more easily filtered out by rendering localization points based on various photoswitching properties, such as photon number and photoswitching rates, because the localizations from noise exhibit distinct photoswitching properties. In addition, the empty area within a cluster can be filled by rendering the localization points, allowing straightforward cluster analysis. Therefore, reconstructed and rendered SMLM images are preferable for cluster analysis rather than localization coordinate data.

For example, French et al. used high-resolution, rendered STORM images for cluster analysis of mitochondria and purinosomes [32,33]. Each rendered STORM image of mitochondria and purinosomes was median filtered and intensity thresholded using Otsu's thresholding algorithm to remove the background noise localizations. The boundaries of the filtered clusters were then identified using 8-point connectivity and dilated for erosion. The identified clusters were further analyzed for the colocalization of mitochondria and purinosomes based on their boundary-to-boundary distances. Such a rendered image-based cluster analysis was also successfully demonstrated for the analysis of the nearest distance between the Xist and EZH2 clusters to investigate X-chromosome inactivation. Each image was separately processed by medial filtering, and the identified clusters of EZH2 were cross-correlated with EZH2 localization coordinates to calculate the nearest neighbor's distance [34]. This method has also been shown to quantify the ultrastructural changes of various organelles, such as mitochondria, dense tubular system (DTS), autophagosomes, α-granules, and dense granules, in platelet during the platelet activation and division process [35,36]. It was also successfully employed to identify and quantify the phase separation in polymer blend films on the nanoscale [37]. The identified boundaries of each phase were further processed for the analysis of specific interfacial lengths and areas.

The kernel density estimation (KDE) is another image-based approach. Based on the pixel information, KDE determines the localization density at a certain position as the kernel density. The points are clustered using kernel size and density as user-defined inputs. Slotman et al. employed the KDE approach to investigate the spatial organization of recombinase RAD51 and its meiosis-specific paralog DMC1 during meiotic DNA double-strand break repair (DSB) [38]. In their study, the single-molecule localization data were fitted to a 2D KDE function to define DNA repair foci at different cell stages. This KDE-based cluster analysis not only demonstrated variability in foci composition but also defined functional consensus configurations during the DSB process.

Collectively, such a rendered SMLM image-based cluster analysis is quite similar to the conventional cluster analysis for pixel-based images obtained from conventional microscopy, even though a rendered SMLM image-based cluster analysis uses much higher resolution images with high pixel numbers. This method was found to be particularly useful for removing small objects, possibly background signals, in a simple way, even though each localization coordinate information can be lost in this method.

3. Machine learning based cluster analysis methods for SMLM images

Machine learning is a subfield of artificial intelligence that enables systems to learn from experience and improve without being explicitly programmed. It focuses on developing an algorithm that can access data and use it to learn for itself. The type of algorithm can be determined by its input data and desired output data. Various machine learning algorithms have been demonstrated to be useful and effective for cluster analysis of SMLM images [39]. To understand these works, we first describe machine learning algorithms that are useful for cluster analysis and then discuss each reported work (Fig. 2, Table 1).

3.1. Backgrounds for machine learning algorithms

3.1.1. Classical machine learning-based algorithm

Over the past decade, deep neural network-based algorithms, such as recurrent neural networks and convolutional neural networks, have been dominant in solving AI-related problems. The main reason for this is that deep learning demonstrates superior performance in various tasks. Although deep learning shows such high performance, there are still some advantages to using classical machine learning algorithms. Classical machine learning algorithms work better on small datasets. In addition, they are cost-effective and easy to interpret. Some notable classical machine-learning algorithms are useful for the cluster analysis of SMLM data.

First, a decision tree is a machine learning algorithm that produces an output from yes-or-no questions. It uses a tree-like

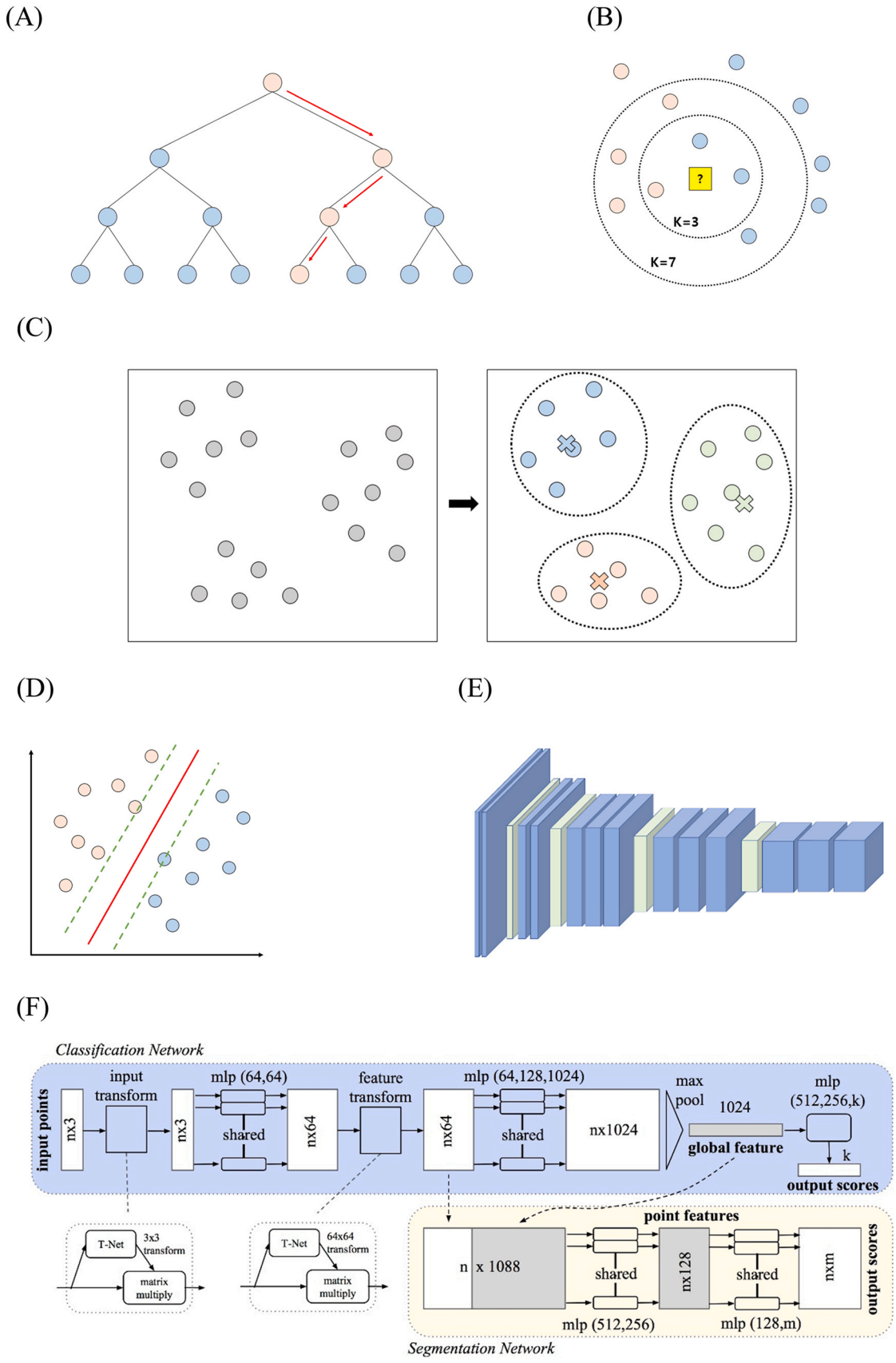

Fig. 2. Machine learning algorithms used in cluster analysis for single-molecule localization microscopy (SMLM) images. (A) Decision tree. (B) K-nearest neighbors. (C) K-means Clustering. (D) Support vector machine. (E) Typical architecture of convolutional neural network. (F) The architecture of PointNet [43].

structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each terminal node contains a class label. Decision trees classify instances by sorting them from the root to a leaf node, which provides the classification of the instance. It is prone to overfitting and instability despite being simple to understand and necessitating minimal data preparation. This is because small variations in the data can result in a completely different tree being generated. Random forest is an ensemble learning method that constructs several levels of decision trees during training.

K-nearest neighbors algorithm (KNN) is a non-parametric supervised learning method that uses proximity to make classifications or predictions for grouping individual data points. The KNN tries to predict the correct class for the test data by calculating the instance between the test data and all training points. Then, the K number of points is chosen to be closest to the test data, followed by the computation of the probability for each class based on these K training points. The class with the highest probability was selected. KNN can be used for a regression problem where the value is determined by the mean of K-selected training points.

K-means clustering is an unsupervised learning algorithm used to identify clusters of data objects in a dataset. This algorithm first selects k centroids, where k denotes the chosen number of clusters. Centroids are data points that represent the center of a cluster. Each data point was assigned to the closest centroid, and the position of the centroids was updated based on the newly assigned points. The algorithm continues the reassignment process repeatedly until the positions of the centroids are stable. The k-means clustering algorithm is easy to construct and compute, and it can be applied to a large dataset. However, it is too sensitive to the initial choice of the value “k” and centroids, and it often fails on complex datasets because its concept is based on separable spherical clusters, with the mean convergent towards the cluster center.

Support vector machine (SVM) is one of the most popular supervised machine algorithms that can be utilized for both classification and regression but is commonly used in classification. SVM was designed for binary classification problems. It identifies a hyperplane that separates data points into different classes. SVM can be easily extended to complex instances that are not linearly separable by mapping training examples to a higher-dimensional space, where they become linearly separable using a kernel trick. SVM has been successful in various applications, such as medical imaging and natural language.

3.1.2. Image classification algorithm based on deep learning

One of the fundamental problems in computer vision is determining whether image data contains specific objects, features, or activities. The image-classification algorithm assigns an input image to one label from a fixed set of categories. More complicated computer vision tasks, such as object detection and segmentation, can utilize architectures developed for image classification.

Although there are classical machine algorithms for image classification, the most common architecture in recent years is neural network-based models. It is typically composed of feature networks and a small number (typically two or three) of additional fully connected layers (or convolution layers). The output of the entire model architecture is a one-dimensional vector whose number is the same as the number of classes. Each element in the final vector represents the probability of each class, and the class with the highest probability was selected as the final prediction. Although a series of fully connected layers can be sufficient for a feature network for a small dataset, CNN-based architectures such as ResNet [40], EfficientNet [41], or Vision Transformer (ViT)-based architectures [42] generally show much better accuracy for complex datasets.

3.1.3. Deep learning algorithm on point cloud

A point cloud is a set of data points in three-dimensional space. Each point may represent a 3D scene environment, the 3D shape of a target object, or the distribution of particles. Such data have been widely used for robot navigation, scene classification, and cluster analyses. Detailed algorithms should be considered with the point cloud input for clustering points and object classification. There are several common approaches based on the deep learning of a point cloud. One approach is to use a 3D-convolution layer to capture the features of point clouds. The points are generally sparse, so the point cloud can be divided into several voxel grids, and the number of points in each grid is considered valid data. In another approach, PointNet [43] uses a symmetric function to impose permutation invariants of points during the input phase and computes the features of points using a multi-layer perceptron architecture. Another approach is to consider the point cloud as a graph structure, where each point serves as a node. Various graph neural networks can be used to analyze reformulated input data.

3.2. Machine learning based cluster analysis methods for SMLM images

3.2.1. Methods for constructing clustered data

Several studies have been conducted on the construction of 3D structures from SMLM images. This reconstructed point cloud can be used for cluster analysis. Blundell et al. used a convolutional neural network (CNN) to retrieve 3D structures from SMLM images [44]. The input for CNN is batches of images obtained from SMLM, and the direct output of the network is six rendering parameters, which represent the position and orientation. The 3D structure of the object can be obtained using a differentiable renderer with these parameters. The authors demonstrated their method, HOLLy, for reconstructing the 3D structures of the CEP152 complex, which is a part of the centriole. The central torus for the CEP152 complex converged after training sets of ~ 2000 2D SMLM images, which was consistent with the previously reported structure of this protein complex, confirming the performance of this method.

Tobin et al. proposed a machine-learning-based clustering analysis method to detect clustered and unclustered molecules using a k-means clustering algorithm [45]. Tobin et al. used the k-means clustering algorithm to quantify the fraction of the targeted molecules residing in clusters. In their study, a k-means-like clustering algorithm was employed to quantify the fraction of human epidermal growth factor receptor 2 (HER2) molecules in clusters, whose level has recently been considered an important indicator of breast cancer. In this algorithm, the cluster radius and average localization precision were used as parameters from the pair-correlation analysis. They described the clustering features of membrane HER2 receptor molecules using this method.

Williamson et al. developed a supervised machine learning approach for cluster analysis that is fast and accurate [46]. The purpose of the suggested algorithm is to classify SMLM points as either clustered or non-clustered. The distances to the nearest neighborhood points were computed for each query point and then used as input data for the neural networks. The difference between these models is the number of neighborhood points and layer operations. The simple model uses only a fully connected layer, whereas the more complex model uses one-dimensional convolution and LSTM (long short-term memory) [47] layers for better accuracy. They demonstrated this method using simulated and experimental SMLM data of the cytosolic kinase Csk and the transmembrane adapter PAG in primary human T cell immunological synapses, since it has been suggested that Csk is regulated through its association with PAG. From this analysis, changes in Csk and PAG clustering were observed in naive and previously stimulated T cells.

One could suggest employing a computational and analytical framework similar to the single-particle reconstruction (SPR)

analysis of electron microscopy (EM) images to reconstruct the 3D image from a 2D SMLM image. Sieben et al. developed a computational and analytical framework that reconstructs and coaligns multiple proteins from 2D super-resolution fluorescence images, which is similar to SPR analysis for EM images used in structural biology [48]. One of the key steps is to align two protein volumes obtained from different images. They carried out orientational filtering using a support vector machine classifier to identify the top-view and side-view projections of the reference protein from a combination of 12 calculated shape descriptors. They remarked that this method is advantageous for direct application to other datasets using the same reference once the model is trained as a reference protein. They reconstructed the 3D four-color map of the human centriole, revealing their relative locations, dimensions, and orientations using the SMLM images of several proteins within the purified human centrosomes immunolabeled for Cep152, Cep164, Cep57, and Cep63.

3.2.2. Methods for cluster identification and classification

Khater et al. [49] developed a computational pipeline for analyzing the large image datasets generated by SMLM images. They applied it to analyzing 3D point clouds of SMLM localizations of the caveolar coat protein caveolin-1(Cav1). Random decision forest algorithms were used to identify features that distinguish the regions of interest of PC3 and PC3-PTRF. After filtering out low-degree blinks, unsupervised k-means clustering algorithms were applied to identify the different blobs. They demonstrated that this method could successfully define not only the molecular structure of plasma membrane-associated caveolae but also the coat protein Cav1 localization signatures for scaffolds. They demonstrated that this method could successfully identify plasma membrane-associated coat protein Cav1 scaffolds that combine to form caveolae and larger scaffolds by defining Cav1 localization signatures for scaffolds and caveolae.

In their subsequent work, they performed blob identification on SMLM images of Cav1 antibodies in prostate cancer cells [50]. Based on their previous work [49], the SMLM data are processed to be represented as a 3D point cloud, which is divided into several blobs (clusters). These are the input data for the machine learning-based classification algorithms. In this study, both unsupervised and supervised learning techniques were used. The CAVIN1/PTRF mask was used to label blobs as either PTRF+ or PTRF- classes used for supervised learning. This mask was also used to assign learned and matched groups as S2 scaffolds, caveolae, S1B scaffolds, and S1A scaffold blobs for unsupervised learning. The main purpose of these algorithms is to calculate the graphlet frequency distribution (GFD), which is a combinatorial object obtained from graph theory. The class of blobs is decided based on the GFD of each blob using a random forest classifier. They demonstrated this method to define the changes in the structural organization in caveolae and scaffolds independently of the association with CAVIN1/PTRF.

They also identified biological structures from the SMLM data using three different machine learning-based methods [51]. For the detection of caveolae and scaffolds, they developed and compared three binary classification methods to identify whether a given 3D cluster of Cav1 proteins is a caveola. The input SMLM dataset is a three-dimensional point cloud, and it is reformulated to be suitable for each classification method. The first method employs a random forest, which uses expert features obtained by hand-designed features from a point cloud. The second method employs a CNN-based architecture that uses multi-view 2D images as input data. The third method uses the PointNet [43] architecture, which takes a point cloud as input data. Although the latter methods are more modern and newly developed, the first method exhibits higher accuracy in their study, probably because of the relatively small size of the experimental dataset.

4. Future direction

Although classical cluster analysis methods demonstrate satisfactory results, there is still room for development, especially in multiple blinking artifacts correction in clustering analysis. Multiple blinking can cause repeated localizations with various numbers from a single molecule, resulting in artificial clustering, followed by misidentification of cluster localization and size measurement. This limitation can be overcome by applying the recently developed blinking-caused artifact correction approaches to classical clustering analysis methods. Although the artifact-free analysis of membrane-protein nanoclusters developed by Baumgart et al. was discussed above as a method to resolve the problem of blinking fluorophore overcounting, other blinking-caused artifact correction approaches can be employed for cluster analysis as well [20]. One example includes the quantitative PALM analysis method developed by Annibale et al. based on the detailed knowledge of the fluorophore photophysical behavior [52]. Based on the systematic investigation of the effect of blinking and fluorescence dark times on PALM imaging, this method enables reliable quantification of photoblinking molecules in a biological sample from PALM images. Another blinking-caused artifact correction method for PALM images is 'model-based correction' (MBC) developed by Jensen et al.[53], which utilizes calibration-free estimation of realistic photophysics fluorescent protein models to provide corrected localization data with enhanced localization precision. Although this method cannot be utilized to correct STORM or other SMLM methods, another recently developed method can also be applied to STORM data. Bohrer et al. developed a distance distribution correction (DDC) algorithm to eliminate multiple blinking-caused artifacts in general SMLM images [54]. Based on the true pairwise distance distribution of different fluorophores, this method could produce a set of localizations without blinking artifacts. Therefore, the application of such recently developed blinking-caused artifact correction approaches to classical cluster analysis may enable not only accurate reconstruction and quantification of single molecules but also accurate quantitative cluster analysis without overclustering.

Further development in machine-learning-based methods is also expected. There have been remarkable results on the cluster analysis of SMLM images using algorithms in machine learning, but the common methods usually come from classical machine learning algorithms. Deep learning algorithms in computer vision, including image classification, object detection, and segmentation, have led to drastic improvements in their performance and resulted in varied applications, which may contribute to cluster analysis. There is a notable effort for adopting deep-learning-based algorithms in this field, but extraordinary results have not yet been achieved. Modern deep neural networks have various options for the model architecture, training methods, and type of input data. Any newly developed architecture for image data or a graph neural network would be helpful for cluster analysis. Therefore, further attempts are expected to be made to exploit recent deep learning-based methods for analyzing clusters in SMLM images. Any newly developed architecture for image data or point cloud data would be helpful for various algorithms related to cluster analysis. For example, image segmentation is a general computer vision algorithm; however, it has not been utilized for cluster analysis in SMLM. Old and new high-performance deep-learning-based segmentation algorithms may be helpful for cluster classification and colocalization.

Additionally, cluster analysis in SMLM image data can be assisted by a graph neural network (GNN) because the clusters in the point cloud can also be considered graph structures. The GNN is a deep-learning-based method designed to be conducted on data described by graphs. GNN can be applied to diverse algorithms, including node classification, graph classification, graph visualization, and graph clustering. Because GNN has recently become a popular research

topic in deep learning, cluster analysis in SMLM could take advantage of these results.

The application of a recently developed 3D analysis method to cluster analysis in the SMLM may generate new perspectives and results. The observed data in SMLM images can have a three-dimensional structure explicitly or implicitly, which can be represented as a 3D point cloud. There are various deep learning methods for 3D point clouds, including 3D shape classification, point cloud segmentation, and object detection and tracking. The applications of these deep-learning methods would expand our understanding of the target 3D structure, which cannot be obtained using conventional analysis methods.

We expect that different cluster analysis methods would reveal distinct performance depending on the type of structures in SMLM images, necessitating the evaluation of their performances for comparison. For example, Nieves et al. reported a framework to evaluate cluster analysis performances of DBSCAN, ToMATo, and KDE by scoring the result of clustering algorithms based on the metrics, including the Adjusted Rand Index (ARI) and Intersection over Union (IoU) [55]. Such a framework to compare the success of clustering results in different cluster analyses would not only provide guidelines for choosing the method but also contribute towards the development of future methodologies.

Collectively, further development of cluster analysis for SMLM data in such a way is anticipated to provide fruitful structural information about targets, allowing extensive opportunities for its application.

5. Conclusion

The recent development of SMLM methods, such as STORM and PALM, has increased the demand for new methods for cluster analysis, due to their distinct pointillism data. Modern computational cluster analysis methods for SMLM images can be categorized into classical and machine learning-based methods.

Classical cluster analysis methods for SMLM images include the global clustering analysis method, the complete clustering analysis method, the tessellation-based method, and the reconstructed image-based method. The global clustering analysis method provides a global description of protein clustering or organization by providing spatial statistics. The complete clustering approach extracts rich information from the data at the single-cluster level, such as the number of clusters and the shape of individual clusters. The tessellation-based method creates a tessellated surface in which tiles are generated from the localizations to determine the presence of clusters. The image-based method uses the image itself for cluster analysis instead of the coordinates of localization.

Various machine learning-based methods have also been exploited for cluster analysis. Classical machine learning algorithms, such as decision trees, random forests, and KNN, are used for cluster identification, and the k-means clustering algorithm has been useful to group related points in the SMLM image as a cluster. Algorithms based on neural networks have also been used for cluster classification. A CNN-based architecture is effective for image-type data, and 3D convolution or PointNet is applied for point cloud processing.

Although these machine learning-based cluster analysis methods have shown satisfactory results, it is expected that there will be more attempts to exploit recent deep learning-based methods for analyzing clusters in SMLM images. Such a further improvement of cluster analysis for SMLM images with the modern deep learning-based methods is expected to clarify demanding questions in a wide range of biology by providing ultrastructural information about targets, finally playing a significant role in super-resolution image analysis.

CRediT authorship contribution statement

Yoonsuk Hyun: Conceptualization, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Doory Kim:** Conceptualization, Writing – original draft, Writing – review & editing, Visualization, Supervision.

Acknowledgements

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2022R1A4A5033271) and INHA UNIVERSITY Research Grant.

References

[1] Khater IM, Nabi IR, Hamarneh G. A review of super-resolution single-molecule localization microscopy cluster analysis and quantification methods. Patterns 2020;1(3):100038.
[2] Bates M, et al. Multicolor super-resolution imaging with photo-switchable fluorescent probes. Science 2007;317(5845):1749–53.
[3] Jeong D, Kim D. Super-resolution fluorescence microscopy-based single-molecule spectroscopy. Bull Korean Chem Soc 2022;43(3):316–27.
[4] Kim D, et al. Correlative stochastic optical reconstruction microscopy and electron microscopy. PLoS One 2015;10(4):e0124581.
[5] Hell SW, Wichmann J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. Opt Lett 1994;19(11):780–2.
[6] Gustafsson MG, Agard DA, Sedat JW. Sevenfold improvement of axial resolution in 3D wide-field microscopy using two objective-lenses. In: Three-dimensional microscopy: image acquisition and processing II. SPIE; 1995.
[7] Rust MJ, Bates M, Zhuang X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Methods 2006;3(10):793–6.
[8] Betzig E, et al. Imaging intracellular fluorescent proteins at nanometer resolution. Science 2006;313(5793):1642–5.
[9] Chung J, et al. Development of a new approach for low-laser-power super-resolution fluorescence imaging. Anal Chem 2021;94(2):618–27.
[10] Kim Gh, et al. Single-molecule sensing by grating-based spectrally resolved super-resolution microscopy. Bull Korean Chem Soc 2021;42(2):270–8.
[11] Bar-On D, et al. Super-resolution imaging reveals the internal architecture of nano-sized syntaxin clusters. J Biol Chem 2012;287(32):27158–67.
[12] Nieves DJ, Owen DM. Analysis methods for interrogating spatial organisation of single molecule localisation microscopy data. Int J Biochem Cell Biol 2020;123:105749.
[13] Rossy J, et al. Conformational states of the kinase Lck regulate clustering in early T cell signaling. Nat Immunol 2013;14(1):82–9.
[14] Gao J, et al. Mechanistic insights into EGFR membrane clustering revealed by super-resolution imaging. Nanoscale 2015;7(6):2511–9.
[15] Rubin-Delanchy P, et al. Bayesian cluster identification in single-molecule localization microscopy data. Nat Methods 2015;12(11):1072–6.
[16] Sengupta P, et al. Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. Nat Methods 2011;8(11):969–75.
[17] Sengupta P, Lippincott-Schwartz J. Quantitative analysis of photoactivated localization microscopy (PALM) datasets using pair-correlation analysis. Bioessays 2012;34(5):396–405.
[18] Hartley JM, et al. Super-resolution imaging and quantitative analysis of membrane protein/lipid raft clustering mediated by cell-surface self-assembly of hybrid nanoconjugates. ChemBioChem 2015;16(12):1725–9.
[19] Pageon SV, et al. Functional role of T-cell receptor nanoclusters in signal initiation and antigen discrimination. Proc Natl Acad Sci USA, Vol. 113(no. 37); 2016, p. E5454–63.
[20] Baumgart F, et al. Varying label density allows artifact-free analysis of membrane-protein nanoclusters. Nat Methods 2016;13(8):661–4.
[21] Rahbek-Clemmensen T, et al. Super-resolution microscopy reveals functional organization of dopamine transporters into cholesterol and neuronal activity-dependent nanodomains. Nat Commun 2017;8(1):1–14.
[22] Marenda M, et al. Parameter-free molecular super-structures quantification in single-molecule localization microscopy. J Cell Biol 2021;220(5).
[23] Mazouchi A, Milstein J. Fast Optimized Cluster Algorithm for Localizations (FOCAL): a spatial cluster analysis for super-resolved microscopy. Bioinformatics 2016;32(5):747–54.
[24] Pike JA, et al. Topological data analysis quantifies biological nano-structure from single molecule localization microscopy. Bioinformatics 2020;36(5):1614–21.
[25] Chazal F, et al. Persistence-based clustering in Riemannian manifolds. J ACM (JACM) 2013;60(6):1–38.
[26] Ghrist R. Barcodes: the persistent topology of data. Bull Am Math Soc 2008;45(1):61–75.
[27] Dlasková A, et al. 3D super-resolution microscopy reflects mitochondrial cristae alternations and mtDNA nucleoid size and distribution. Biochim Biophys Acta (BBA)-Bioenerget 2018;1859(9):829–44.
[28] Levet F, et al. SR-Tesseler: a method to segment and quantify localization-based super-resolution microscopy data. Nat Methods 2015;12(11):1065–71.

[29] Andronov L, et al. ClusterViSu, a method for clustering of protein complexes by Voronoi tessellation in super-resolution microscopy. Sci Rep 2016;6(1):1–9.
[30] Andronov L, et al. SharpViSu: integrated analysis and segmentation of super-resolution microscopy data. Bioinformatics 2016;32(14):2239–41.
[31] Hadipour-Lakmehsari S, et al. Nanoscale reorganization of sarcoplasmic reticulum in pressure-overload cardiac hypertrophy visualized by dSTORM. Sci Rep 2019;9(1):1–17.
[32] French JB, et al. Spatial colocalization and functional link of purinosomes with mitochondria. Science 2016;351(6274):733–7.
[33] Chan CY, et al. Microtubule-directed transport of purine metabolons drives their cytosolic transit to mitochondria. Proc Natl Acad Sci USA, Vol. 115(no. 51); 2018, p. 13009–14.
[34] Sunwoo H, Wu JY, Lee JT. The Xist RNA-PRC2 complex at 20-nm resolution reveals a low Xist stoichiometry and suggests a hit-and-run mechanism in mouse cells. Proc Natl Acad Sci USA, Vol. 112(no. 31); 2015, p. E4216–25.
[35] Chung J, et al. Super-resolution imaging of platelet-activation process and its quantitative analysis. Sci Rep 2021;11(1):1–18.
[36] Go S, et al. Super-resolution imaging reveals cytoskeleton-dependent organelle rearrangement within platelets at intermediate stages of maturation. Structure 2021;29(8):810–22. e3.
[37] Park Y, et al. Polarity nano-mapping of polymer film using spectrally resolved super-resolution imaging. ACS Appl Mater Interfaces 2022;14(40):46032–42.
[38] Slotman JA, et al. Super-resolution imaging of RAD51 and DMC1 in DNA repair foci reveals dynamic distribution patterns in meiotic prophase. PLoS Genet 2020;16(6):e1008595.
[39] Hyun Y, Kim D. Development of deep-learning-based single-molecule localization image analysis. Int J Mol Sci 2022;23(13):6896.
[40] He K, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
[41] Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. In: Proceedings of the international conference on machine learning. PMLR; 2019.
[42] Dosovitskiy A, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929; 2020.
[43] Qi CR, et al. Pointnet: deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
[44] Blundell B, et al. 3D structure from 2D microscopy images using deep learning. Front Bioinf 2021;1:740342.
[45] Tobin SJ, et al. Single molecule localization microscopy coupled with touch preparation for the quantification of trastuzumab-bound HER2. Sci Rep 2018;8(1):1–13.
[46] Williamson DJ, et al. Machine learning for cluster analysis of localization microscopy data. Nat Commun 2020;11(1):1–10.
[47] Hochreiter S, Schmidhuber J. Long short-term memory. Neur Comput 1997;9(8):1735–80.
[48] Sieben C, et al. Multicolor single-particle reconstruction of protein complexes. Nat Methods 2018;15(10):777–80.
[49] Khater IM, et al. Super resolution network analysis defines the molecular architecture of caveolae and caveolin-1 scaffolds. Sci Rep 2018;8(1):1–15.
[50] Khater IM, et al. Identification of caveolin-1 domain signatures via machine learning and graphlet analysis of single-molecule super-resolution data. Bioinformatics 2019;35(18):3468–75.
[51] Khater IM, et al. Caveolae and scaffold detection from single molecule localization microscopy data using deep learning. PLoS One 2019;14(8):e0211659.
[52] Annibale P, et al. Quantitative photo activated localization microscopy: unraveling the effects of photoblinking. PLoS One 2011;6(7):e22678.
[53] Jensen LG, et al. Correction of multiple-blinking artifacts in photoactivated localization microscopy. Nat Methods 2022;19(5):594–602.
[54] Bohrer CH, et al. A pairwise distance distribution correction (DDC) algorithm to eliminate blinking-caused artifacts in SMLM. Nat Methods 2021;18(6):669–77.
[55] Nieves DJ, et al. A framework for evaluating the performance of SMLM cluster analysis algorithms. bioRxiv; 2021.