A Self-Supervised Foundation Model for Robust and Generalizable Representation Learning in STED Microscopy

Bilodeau, Anthony[1, 2, †], Beaupré, Frédéric[1, 2, †], Chabbert, Julia[2], Bellavance, Jean-Michel[2], Lessard, Koraly[1,2], Deschênes, Andréanne[1,2], Bernatchez, Renaud[1,2], De Koninck, Paul[2, 3], Gagné, Christian[1, 4, 5], and Lavoie-Cardinal, Flavie[1, 2, 6, *]

[1]Institute Intelligence and Data, Université Laval
[2]CERVO Brain Research Centre, Université Laval
[3]Department of Biochemistry, Microbiology, and Bio-informatics, Université Laval
[4]Department of Electrical and Computer Engineering, Université Laval
[5]Canada-CIFAR Artificial Intelligence Chair
[6]Department of psychiatry and neurosciences, Université Laval
[*]Contact author: flavie.lavoie-cardinal@cervo.ulaval.ca

Abstract

Foundation Models (FMs) have dramatically increased the potential and power of deep learning algorithms in the fields of natural language processing and computer vision. However, their application in specialized fields like biomedical imaging and fluorescence microscopy remains difficult due to distribution shifts and the scarcity of high-quality annotated datasets. The high cost of data acquisition and the requirement for in-domain expertise further exacerbate this challenge in super-resolution microscopy. To address this we introduce STED-FM, a foundation model specifically designed for super-resolution STimulated Emission Depletion (STED) microscopy. STED-FM leverages a Vision Transformer architecture trained at scale with Masked Autoencoding on a new dataset of nearly one million STED images. STED-FM learns expressive latent representations without extensive annotations, leading to robust performance across diverse downstream microscopy image analysis tasks. Unsupervised experiments highlight the discriminative nature of its learned latent space. Our model significantly reduces the need for annotated data required to achieve strong performance in classification and segmentation tasks, both in- and out-of-distribution. Beyond its core capabilities, STED-FM enhances the quality of images generated by diffusion models, enabling latent attribute manipulation for the data-driven discovery of novel and subtle nanostructures and phenotypes. Moreover, its powerful structure retrieval capabilities are integrated into automated STED microscopy acquisition pipelines, paving the way for smart microscopy. In sum, we demonstrate that STED-FM lays a robust foundation for state-of-the-art algorithms across a wide array of tasks, establishing it as a highly valuable and scalable resource for researchers in super-resolution microscopy.

1 Introduction

Deep learning has become a powerful tool for predictive and quantitative analysis, driven by its ability to learn expressive representations of complex, high-dimensional data. The scaling hypothesis suggests that performance improves with larger models, datasets, and computational resources [1], which gave rise to the development of foundation models [2]: large-scale pre-trained models capable of excelling at various tasks. Foundation models (FMs) have been extensively studied for natural language processing [3, 4] and computer vision [5]. FMs have benefited from large curated datasets for training [6–9] and show impressive generalization capability on downstream tasks [10]. However, in specialized fields such as biomedical imaging, their effectiveness remains uncertain due to significant distribution shifts between downstream and upstream domains [11]. In biomedical imaging, limited data availability poses a challenge, as both fine-tuning of models

pre-trained on natural images and training a new model from scratch often result in suboptimal outcomes [12]. Consequently, there is a growing interest in developing FMs tailored specifically to biomedical and bioimaging applications [13–16].

One of the most significant challenges associated with the development of FMs for bioimaging tasks is the scarcity of high-quality annotated datasets [17]. This can be explained by the cost of data acquisition as well as the in-domain expertise and time required to annotate it [18–21]. Moreover, variability in the imaging modalities and experimental settings, including batch effects, will inevitably increase the complexity of dataset aggregation and curation. It will also lead to heterogeneous data distributions that need to be learned by the FMs. This can be challenging in a context where available datasets are of limited size [22]. These challenges can hinder the training of expressive FMs for biomedical imaging.

Yet, wide-ranging successes in other domains have motivated the design and training of FMs adapted to the bioimaging research field. Efforts have been put into the collection and curation of large bioimaging datasets [13, 23–26] and of datasets in more specialized fields, such as chemistry [27] and microscopy [15, 28–32]. Leveraging such datasets, researchers have proposed to use the FM paradigm to either train deep learning models in a fully-supervised fashion [13, 26, 31], using a combination of fully-supervised and self-supervised learning (SSL) [12], or with self-supervised learning only [18, 19]. SSL refers to the training procedures that aim to learn general representations from a dataset without any annotations [33–37]. On natural images, it has been shown that SSL can achieve similar or superior performance to their fully-supervised counterparts, in particular when few annotations are available [18]. One family of SSL techniques relies on generative approaches to learn meaningful representations of images by encoding them into a latent space. A decoder network then attempts to generate the image (or parts of an image) from its encoded vector [36]. More recent approaches have been introduced, using the strategy of corrupting inputs and training networks to reconstruct the corruption-free versions. For example, masked autoencoders [38] (MAEs) learn to reconstruct masked patches from a transformer encoder using only the visible encoded patches and a lightweight decoder. The assumption is that the latent representations should be expressive enough for the decoder to reconstruct the image, even from a corrupted version of the original signal. The encoder can afterwards be used as a general-purpose model. Such methods have been shown to scale well and to generalize efficiently to unseen datasets and tasks [38], including in biomedical imaging [32, 39, 40].

Due to their ability to learn meaningful representations from unannotated image data, SSL methods have already been successfully applied in microscopy for tasks such as image denoising and resolution enhancement [15, 41–46], object detection [47], and segmentation [48]. Recent work pointed out that in-domain pre-training can significantly impact the performance on downstream tasks, especially in the small data regime [16, 26]. In this work, we develop an FM tailored for super-resolution STimulated Emission Depletion (STED) microscopy, hence called STED-FM. Super-resolution microscopy, or optical nanoscopy, has transformed our ability to investigate biology at its molecular scale [49]. However, its complexity and extreme dependence on optimal parameters of acquisition and sample preparation [50] further limit data availability. Meanwhile, the richness of hidden information that could be extracted remains largely under-exploited [49]. Our model was specifically designed to perform a wide array of downstream tasks on small STED imaging datasets. To this end, we created a dataset of nearly one million STED images for the training of STED-FM [51], a general-purpose vision transformer (ViT) [52] (Fig. 1a). Our approach leverages the task-agnostic, auto-regressive MAE framework, originally introduced for natural images to emulate self-supervised training strategies used in natural language processing [38] (Fig. 1b). MAE is a well established method for downstream tasks such as classification and segmentation in biomedical [40, 53] and specialized fields, such as microscopy [32, 39].

We conduct a comprehensive evaluation of STED-FM's learned latent representations and demonstrate their utility across various tasks. The expressiveness of the learned manifold is validated through clustering (Fig. 1c: *Recursive Clustering*) and instance retrieval, where similar structures are grouped coherently (Fig. 1c: *Retrieval*). On standard downstream tasks such as classification and segmentation, STED-FM reduces the amount of annotated data required to achieve state-of-the-art performance (Fig. 1c: *Fine-tuning*). Furthermore, conditioning diffusion-based generative models on latent codes generated by our model improves the quality and visual fidelity of generated images (Fig. 1c: *Image Generation*), both enabling attribute manipulation and revealing structures that may be overlooked when screening images under different

experimental conditions. Finally, integrating STED-FM into the acquisition pipeline enables the automated detection of specific structures, thereby facilitating fully autonomous image acquisition on a STED microscope.

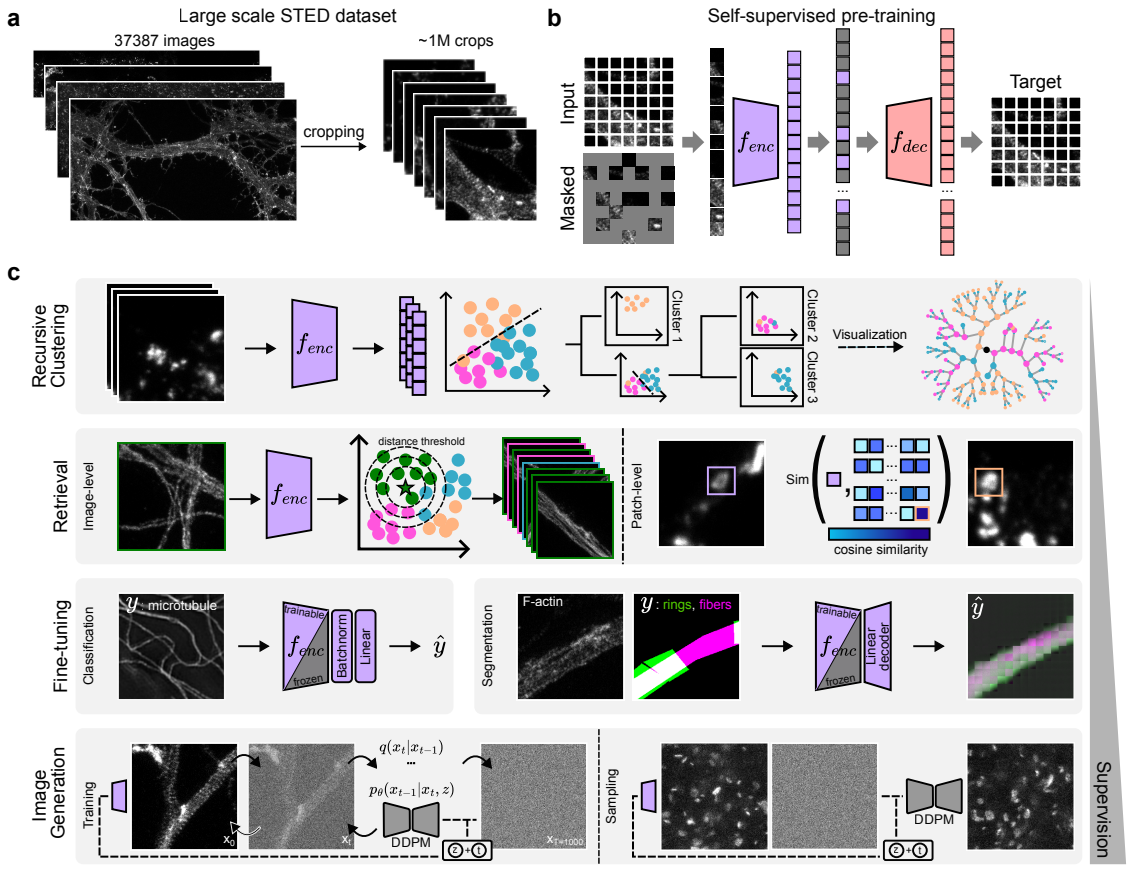


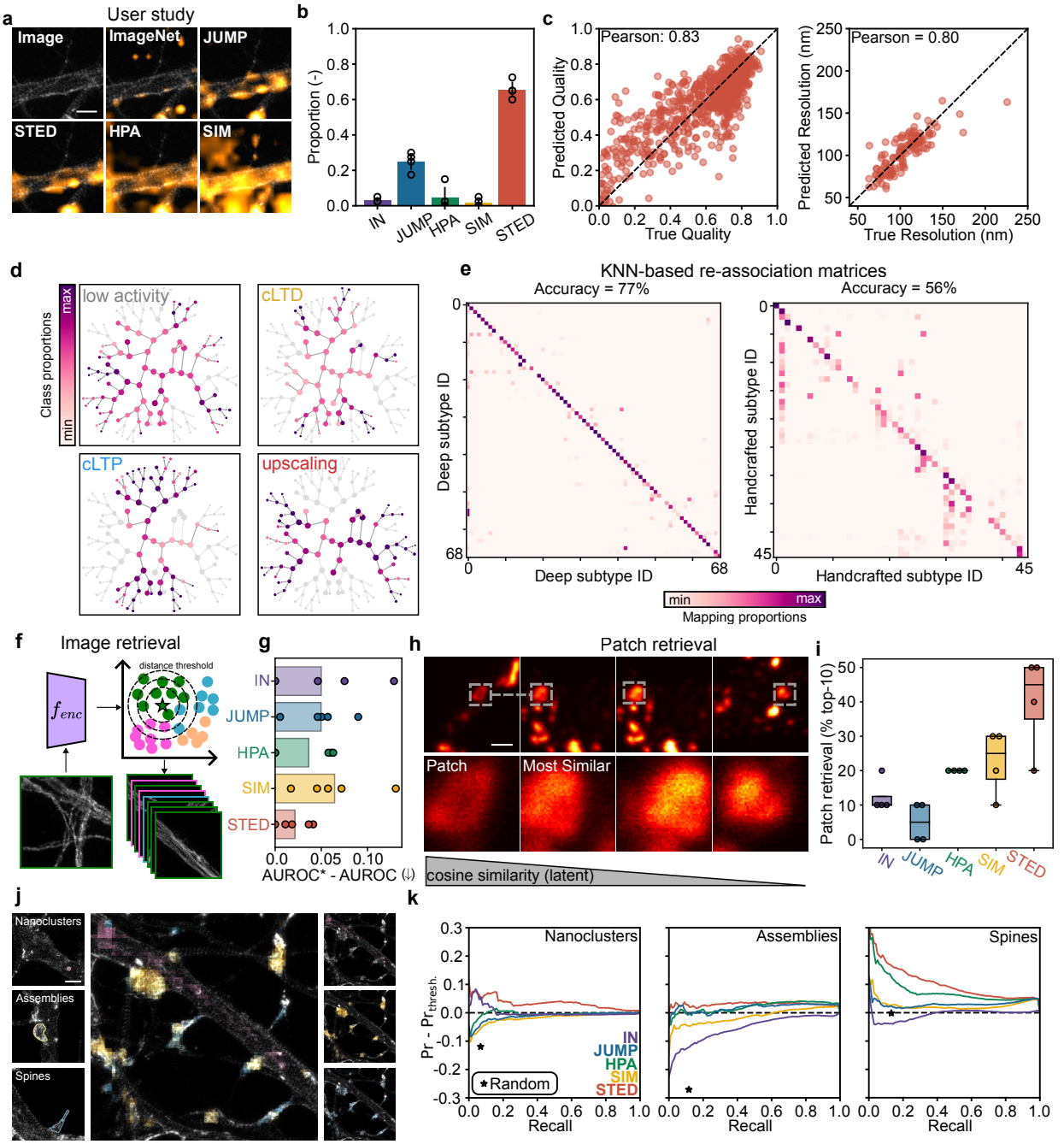
Figure 1: Overview of the deep learning framework. **a)** Large dataset of STED images cropped into smaller regions of 224×224 pixels resulting in a dataset of nearly one million crops used to train STED-FM. **b)** MAE training procedure for pre-training ViTs. **c)** Schematized overview of the different tasks performed by the pre-trained STED-FM model, listed in increasing order of supervision required.

2 Results

We first created a large curated dataset of approximately 1 million STED images (224×224 pixels) to enable the training of STED-FM (*STED-FM dataset*, Methods) [51]. The 224×224 pixel images were extracted from 37,387 large images of variable sizes (Fig. 1a). The dataset consisted of images of more than 20 different fluorescently-labeled proteins in different cell types (hippocampal and cortical neurons, HEK, NG108, Caco2), imaging conditions (live and fixed cells), and sample type (cell culture, tissue slices) acquired on an Abberior STED Expert Line microscope (Methods). We considered 4 different architecture sizes for STED-FM: ViT-tiny, small, base, and large. In early experiments, ViT-small (ViT-S) and ViT-base showed similar performance on downstream classification tasks (Supplementary Fig. 1). To minimize compute resources, we chose the ViT-S architecture. Thus, STED-FM refers to the ViT-S weights after pre-training on the *STED-FM dataset*.

2.1 Unsupervised Analysis

For baseline comparison, we used the publicly available pre-trained weights from *ImageNet* [6] and also pre-trained the ViT-S architecture on three broadly used microscopy datasets: the Joint Undertaking in Morphological Profiling Cell Painting dataset (*JUMP*) [28], the Human Protein Atlas dataset (*HPA*) [54], and



Caption is on the next page.

a Structured Illumination Microscopy (*SIM*) dataset [15]. The pre-training datasets cover a wide range of domains: natural images, microscopy, and super-resolution microscopy. More detailed information is provided for each dataset in this Supplementary Table 1.

As an initial evaluation, we aim to assess the quality and robustness of the representations learned by STED-FM during pre-training, and to explore the breadth of capabilities enabled by the latent representation space prior to any supervised post-training. We validated the learned representations of STED-FM against that of the ViT-S pre-trained on *ImageNet*, *JUMP*, *HPA*, and *SIM*.

Figure 2: Unsupervised analysis of the vector space learned during pre-training. **a)** Attention maps generated from the learned representation of STED-FM and ViT-S pre-trained on *ImageNet*, *JUMP*, *HPA* and *SIM* datasets used in the user study. **b)** Alignment of the different models' attention maps with expert preference. **c)** Correlation between ground truth and inferred image attribute values obtained from a ridge regression model trained on STED-FM's image embeddings. **d)** Graph visualizations of the recursive consensus clustering (RCC) algorithm performed on the *NAS Dataset* for neurons treated with different activity-modulating solutions (Methods). The different conditions show distinct patterns in the RCC graph, supporting the emergence of activity-dependent synaptic subtypes. Node sizes are given by the number of embeddings in the node. Nodes are color-coded by the proportion of the given class. **e)** Matrices obtained from re-assigning every image embedding to the majority class of its 10 nearest neighbors, where class refers to leaf node identifier. RCC performed on the latent manifold is more coherent than when performed on the hand-crafted manifold. **f)** Schematic of the image retrieval task. **g)** Image-based retrieval of distinct biological structures in 5 downstream datasets (*SO, NAS, Px, PR, DL-SIM*). For each dataset and pre-trained model, we report the difference to the best performing model (AUROC*) for structure-specific retrieval. On average, STED-FM shows near-optimal performance, with AUROC*-AUROC close to 0. **h)** Schematic and retrieved examples from the patch retrieval task using STED-FM templates for the identification of perforated synapses in the *Synaptic Protein Zoo Dataset*. **i)** Patch retrieval performance of the different pre-trained models for the identification of perforated synapses averaged over 4 experts' preferences. **j)** Retrieval of different F-actin conformations as templates: nanoclusters, compact assemblies, and spines using STED-FM on the *F-Actin Conformation Dataset*. **k)** Segmentation precision over the range of recall for each pre-trained model compared to using an intensity threshold-based approach on the pixel intensity distribution. Manual annotations were obtained for each conformation on 79 images. Random corresponds to re-assigning the manual annotations randomly on the foreground. Scale bars: 1 µm. *Random* refers to the random initialization of the weights of models trained from scratch.

The attention mechanism embedded within the backbone grants us a direct probe into how the model assigns importance and relevance to different regions of an image. Intuitively, the model's attention should align with an expert's notion of biological relevance. To quantitatively analyze this alignment, we sampled 40 images of the protein F-Actin and extracted the attention maps (Methods) from the models pre-trained on different datasets (Fig. 2a). We then presented these attention maps to 4 experts and asked them to select the attention map that most aligned with their preferences (Fig. 2a,b). The attention maps obtained from STED-FM were identified by the experts as overlapping better with the biological structures of interest in more than 60% of the time. This preference highlights that STED-FM's attention is better aligned with human preference and, in turn, that its predictions on ensuing downstream tasks are likely more interpretable, a key advantage for bioimage analysis.

Next, we used ridge regression to inspect whether physical attributes, such as image quality and spatial resolution, were encoded in the latent space of the models during pre-training (Methods, Fig. 2c). For both attributes, a Pearson correlation coefficient of at least 0.80 is obtained by STED-FM, indicating that it can encode relevant image features in its latent space. Interestingly, using the same procedure on the latent space learned from other pre-training datasets leads to decreased correlation, suggesting that these attributes are not as efficiently learned from pre-training on out-of-domain datasets (Supplementary Fig. 2).

Next, we investigated the global structure of the latent manifold using recursive consensus clustering (RCC) applied to the deep features of a given dataset. For this experiment, we used the *Neural Activity States Dataset (NAS)* [55], which consists of images of the post-synaptic protein PSD95 in fixed hippocampal neurons. Prior to fixation, the neurons were treated with different activity-modulating solutions (Methods) [55]. Our RCC procedure, adapted from Sonpatki & Shah [56], consists in embedding images into the latent space of STED-FM and then recursively applying the consensus clustering algorithm from Monti *et al.* [57] to these embeddings until all nodes in the resulting clustering tree are leaf nodes (Methods). The same RCC procedure is done from features that are manually extracted from the images (Supplementary Fig. 3). Once the clustering tree has been computed, we can visualize it as a graph (Fig. 2d) to evaluate if STED-FM is able to extract discriminative structural features associated with different neuronal activity states. RCC uncovered synaptic subtypes at leaf nodes which are rich in information (Supplementary Fig. 4a & 5). We performed K-Nearest Neighbors (KNN) classification ($K = 9$) on image embeddings (embedded with either STED-FM or hand-crafted feature extraction), using leaf node assignments from RCC as the ground truth labels. We then re-assigned every sample to the majority class/leaf node of its K nearest neighbors in the feature space,

obtaining an accuracy of 77% with STED-FM embeddings (left-hand side of Fig. 2e), compared to the 56% accuracy obtained with the hand-crafted graph (right-hand side of Fig. 2e). The latent manifold learned through pre-training STED-FM is thus better in terms of coherence, representational power, reduction of redundancy, and discriminative capability than a manifold built from manual extraction of hand-crafted features (Supplementary Fig. 3 and 4).

In microscopy image analysis, it is common for experts to go through a time-consuming visual inspection of the acquired images to identify specific structural phenotypes requiring further investigation. To automate the retrieval of similar substructures in large STED images rather than relying on manual inspection, we considered leveraging the latent representations learned by STED-FM for image and patch retrieval [58–60]. We evaluated the performance of STED-FM and of ViT-S with different pre-trainings for image and patch retrieval on different publicly available super-resolution microscopy datasets. We chose datasets acquired on the same microscope as the one used to generate the large scale dataset (in domain (ID); *STED Optimization Dataset - SO* [61], *NAS* [55]) and on different microscopes (out of domain (OOD); *Peroxisome Dataset - Px* [62], *Polymer Rings Dataset - PR* [63], and *Deep Learning Assisted Structured Illumination Microscopy Dataset - DL-SIM* [64]). For image-retrieval, we embedded an image (the *template*) and measured its similarity to other image embeddings within the dataset using cosine distance (Fig. 2f & Supplementary Figure 6, Methods). We evaluated the image retrieval task on all the images of *SO, NAS, Px, PR,* and *DL-SIM* datasets and found that STED-FM consistently outperforms other ViT-S pre-trainings, achieving either best or near-optimal performance on every downstream dataset (Fig. 2g, Methods).

A similar experiment was conducted using patch embeddings as templates. The experiment consisted of using STED-FM to automate the daunting task of retrieving rare structures in large microscopy images. We considered the retrieval of perforated synaptic structures from the annotated *Synaptic Protein Zoo Dataset - SPZ* [65] (Fig. 2h). We conducted a user study to compare the relevance of patches retrieved by STED-FM and ViT-S with different pre-trainings for the identification of perforated synapses in the images (Methods). STED-FM demonstrated a superior ability to retrieve patches similar to the perforated synapse template, suggesting that it is able to encode subtle nano-organization patterns enabling the identification of a rarely observable structure in the *SPZ* (Fig. 2i).

Finally, we leverage the patch-retrieval capability of the pre-trained models to automatically highlight regions of interest based on a single template on the *F-Actin Conformation Dataset* [66, 67] (Fig. 2j). Traditionally, quantification of such structures requires time-consuming manual annotations of identified structures of interest [66, 68]. Using one template for each class, we retrieved F-actin patches corresponding to: 1) F-actin nanoclusters, 2) F-actin compact assemblies, and 3) dendritic spines (Fig. 2j). We compared the performance of STED-FM to an intensity threshold-based approach and to ViT-S with different pre-trainings (Fig. 2k). We show that STED-FM outperforms the other approaches for the segmentation of all structures according to manual ground truth annotations on test images (Fig. 2k).

2.2 Supervised post-training

We evaluated the quality of the representations learned during pre-training by using them as initialization for supervised post-training. We evaluated the performance on two downstream tasks : 1) classification and 2) segmentation. For both tasks, we added a linear layer to the STED-FM backbone to predict either image classes or segmentation maps from the extracted features (Methods). We considered two training strategies: linear probing, where the backbone is frozen and only the linear layer is trained, and end-to-end fine-tuning, which allows for fine-tuning of the entire network. To simulate the scarcity of data often encountered in bioimaging, we also repeated both classification and segmentation tasks using limited fine-tuning data.

For the downstream classification tasks, we use two ID (*SO, NAS*) and three OOD datasets (*Px, PR, DL-SIM*) (Fig. 3a, Supplementary Table 2). We evaluated the performance of STED-FM and ViT-S with different pre-trainings on each dataset and observed that STED-FM achieves best performance across nearly all tasks, with the largest gaps in performance occurring on the ID datasets for the linear probing case (Fig. 3b, left). The only exception was the DL-SIM dataset, which consists of images from a different microscopy modality (SIM) for which the best performing model is ViT-S pre-trained in-domain on SIM data. We also benchmark a fully-supervised model trained from scratch (random initialization, end-to-end fine-tuning). Importantly, all

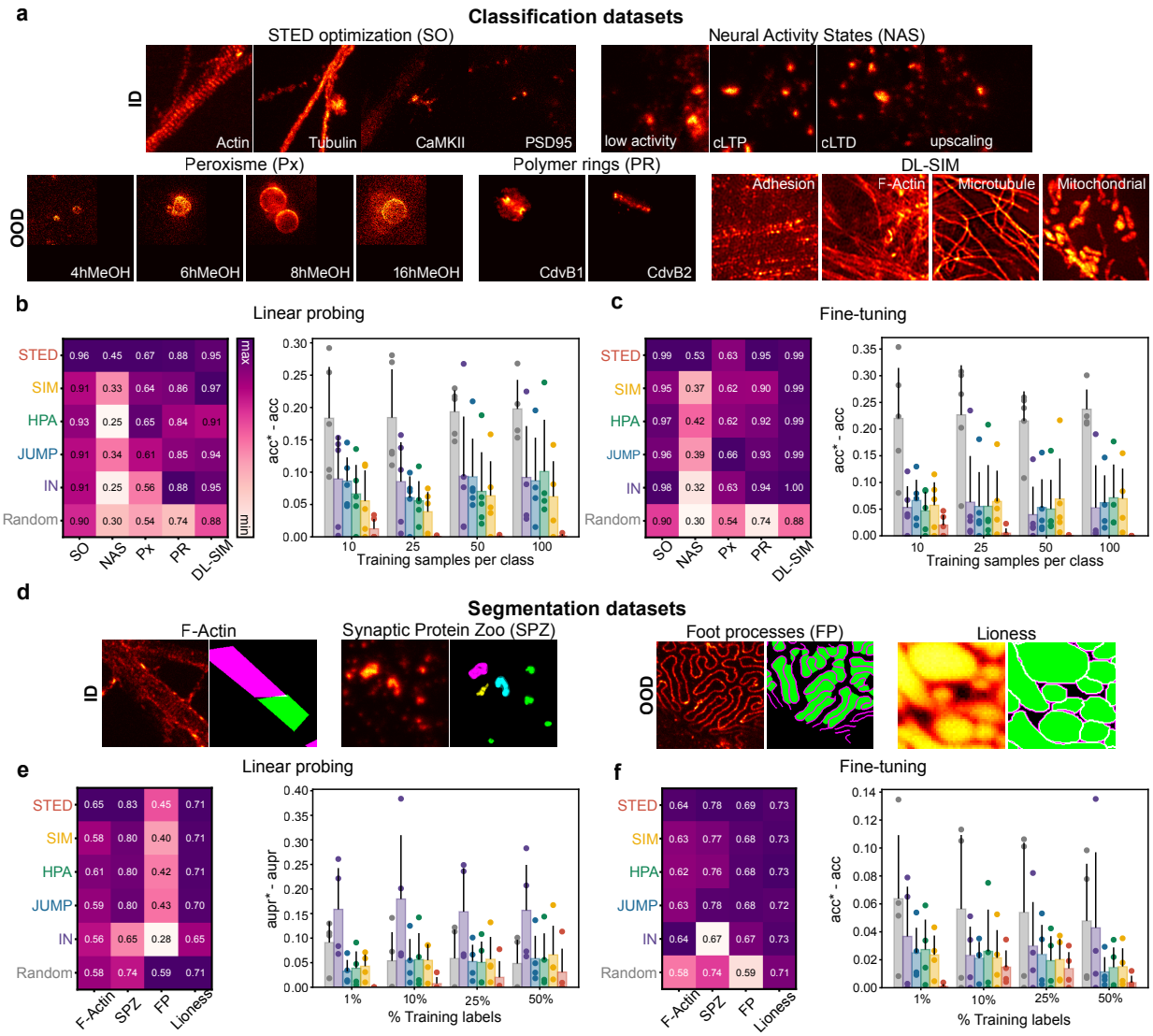


Figure 3: Supervised downstream task analysis. **a)** The classification performance was evaluated on 2 in-domain (ID) and 3 out-of-domain (ODD) super-resolution microscopy datasets. **b,c)** Classification accuracy in linear probing (b) and end-to-end fine-tuning (c) for the full datasets (left) and small subsets (small data regime, right). In the small data regime, performance is reported as the difference from optimal accuracy ($\text{acc}^* - \text{acc}$), where optimal accuracy represents the best performance achieved by any model for a given configuration. Intuitively, a model which performs best in all possible combinations of downstream dataset and number of samples per class would have $\text{acc}^* - \text{acc} = 0$. **d)** The segmentation performance was evaluated on 2 ID and 2 OOD super-resolution microscopy datasets. **e, f)** Segmentation accuracy in linear probing (e) and end-to-end fine-tuning (f) for the full datasets (left) and small subsets (small data regime, right). **b,c,e,f** The color-scale is obtained from a column-wise min-max adjustment.

networks with STED-FM as the backbone exceed the performance of the models trained from scratch. This demonstrates the challenges of training deep learning models from scratch on the limited amount of data typically available in super-resolution microscopy, and it shows the substantial advantage of using a network pre-trained in-domain and at scale, such as STED-FM. Next, we evaluated the performance of the different models with a limited number of samples to train the linear layer (Fig. 3b, right). Comparing all pre-trained models with each other, we observed that STED-FM is near optimal on all datasets (Figure 3b, Supplementary Fig. 7, Methods). Similar trends were observed with end-to-end fine-tuning (Fig. 3c). STED-FM remains the best model, with smaller performance margins compared to other models than in the linear probing case. This is likely due to the larger number of trainable weights in the end-to-end fine-tuning scenario, allowing

models with less effective initializations (pre-trainings) to inch closer to the performance of STED-FM. The latter, when fine-tuned end-to-end in the small data regime, outperforms the other approaches for all datasets (Fig. 3c, right).

For the downstream segmentation tasks, we used two ID: *F-Actin* [66], *SPZ* [65] and two OOD datasets: *Foot processes - FP* [69], *Lioness* [30], all consisting of annotated STED microscopy images (Fig. 3d and Supplementary Table 3 & 8). The linear probing results in Fig. 3e show that STED-FM achieved the best overall performance for the segmentation of nanostructures across all downstream datasets, both in the full and small data regimes. *ImageNet*-pre-training performance drops below random initialization, highlighting the importance of in-domain pre-training for dense prediction tasks. The model trained from scratch also surpasses *JUMP, HPA* and *SIM* pre-training in certain small data regime configurations (Supplementary Fig. 9). Similar conclusions can be drawn from the end-to-end fine-tuning results (Fig. 3f). STED pre-training outperforms the other methods, but, as in the classification downstream tasks, the larger number of trainable weights decreases the gap between STED-FM and the other models.

Given the improved latent representation of STED-FM in both unsupervised and supervised downstream analysis in comparison to the other pre-trained models, we only consider STED-FM for further analysis in the next sections.

2.3 Image Generation

2.3.1 Diffusion models conditioned on pre-trained latent codes

Next, we explored how the learned representations from STED-FM can serve as a conditioning signal for a denoising diffusion probabilistic model (DDPM [70]). This *latent guidance* framework allows us to decode the latent representations back to pixel space, which serves as a powerful tool to analyze and interpret the learned manifold of STED-FM. Another approach that is often proposed for conditioning DDPMs is the classifier-free guidance method [71], which aims to generate images based on a learnable class token (Methods). While classifier-free guidance allows generation of class-specific images (e.g. F-actin, PSD95, Tubulin, Bassoon), we argue that using latent guidance to train a DDPM improves the conservation of key biological structures in the generated super-resolution microscopy images, while providing an interpretable tool to investigate the learned latent representations of STED-FM.

Our *latent guidance* approach (Fig 4a) closely follows the work of Preechakul *et al.* [72]. From the seminal DDPM, we added a conditioning signal $\mathbf{z}$, representing STED-FM's embedding of the input image (Methods). Once the latent-guided DDPM is trained, a noise image $\mathbf{x_T}$ can be given to the DDPM along with the latent code $\mathbf{z}$ of a test image $\mathbf{x_0}$, and it will generate a new image that preserves the features present in image $\mathbf{x_0}$. The intuition for using this approach is that a richer latent code should yield better quality and visual fidelity in the generated image. It should better preserve the test images' features and semantics when denoising $\mathbf{x_T}$ (conditioned on $\mathbf{z}$) back to $\mathbf{x_0}$ (Fig. 4a).

To assess the performance of our latent-guided DDPM, we conducted a user study in which we sampled 45 images from the *SO* dataset and asked experts to identify the primary structure present in the generated image (Actin, Tubulin, PSD95, or Other for unidentifiable patterns; Fig. 4b, left). We compared the image classification accuracy of synthetic images generated with latent and classifier-free guidance to the one obtained by the experts on real images (Fig. 4b, right). The results from the user study indicate that classification accuracy is similar for real and synthetic images generated with latent guidance and improved in comparison with classifier-free guidance (Supplementary Figure 10). This supports the ability of latent guidance to generate high-quality, high-fidelity images in which the features produced are coherent with those observed in real images.

2.3.2 Latent code manipulation reveals encoded attributes

A functionality that is inherent to the latent guidance procedure is the ability to manipulate the conditioning vector in the latent space of STED-FM and use the DDPM to observe how the generated images change upon latent code manipulation. These manipulations in STED-FM's latent space help us further assess the encoding quality and representation capacity of STED-FM. We adapted the methodology outlined by

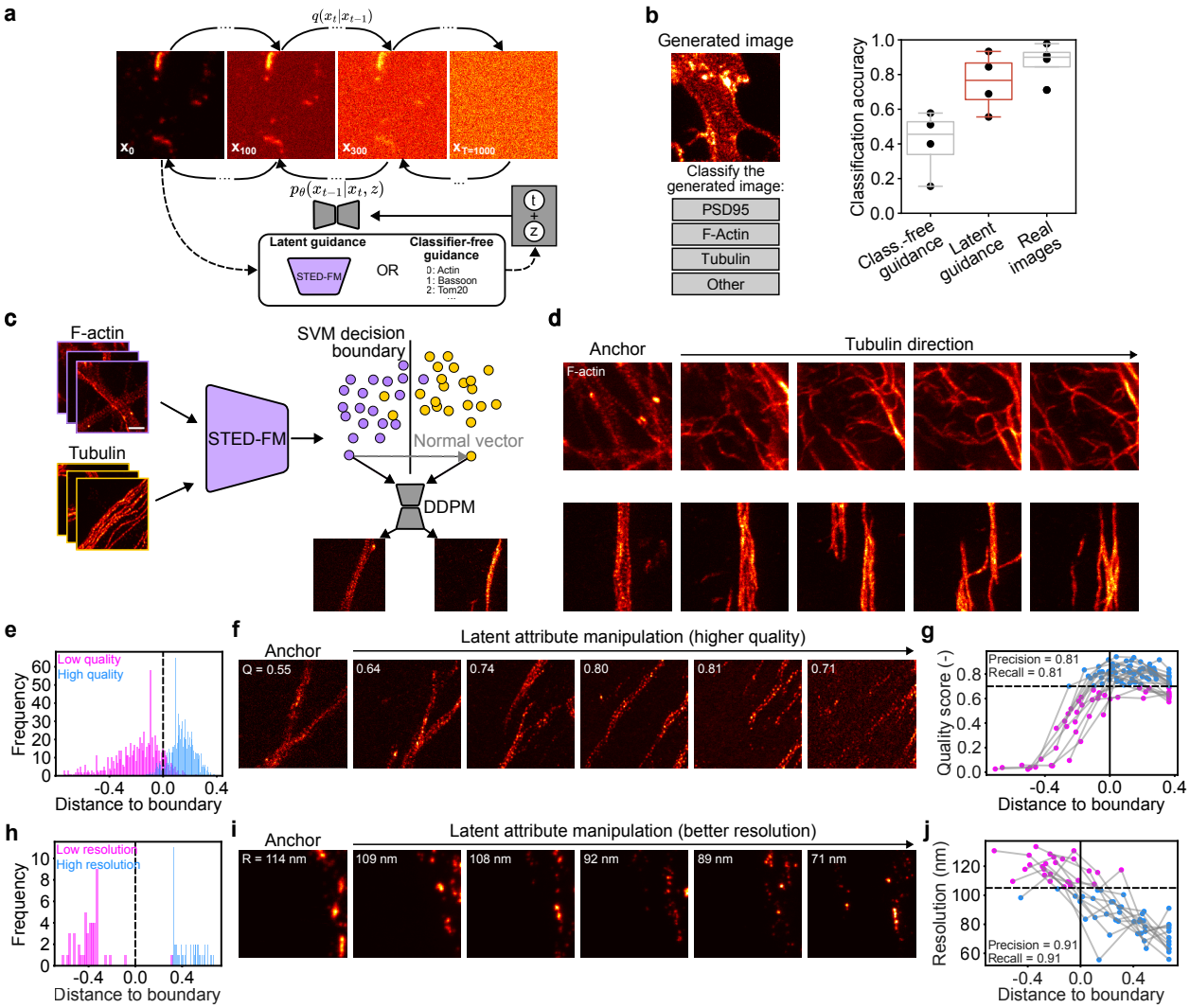


Figure 4: Image generation and latent attribute manipulation. **a)** Overview of the latent guidance and classifier-free guidance frameworks for training DDPMs. (q: diffusion kernel, p: learned denoising kernel, **t**: timestep embedding, **z**: conditioning vector) **b)** A user study was designed to evaluate experts' classification accuracy on real and generated images (N=4 experts). **c)** Schematic of the latent attribute manipulation pipeline. A SVM is trained on the latent representations from STED-FM to discriminate between two groups. The latent code of an image is manipulated by translation in the direction perpendicular to the decision boundary (normal vector). **d)** The latent code manipulation pipeline allows translation of structures in images of F-actin into features associated with tubulin. **e-j)** Latent code manipulation of spatial features encoded in super-resolution microscopy images, such as image quality (**e-g**) and spatial resolution (**h-j**) **e, h)** Histogram of embeddings' distance to the SVM's decision boundary. **f, i)** Image generation trajectories using latent code manipulation. **g, j)** The quality score and spatial resolution correlate with the distance to the boundary, enabling precision and recall measurements (Methods). Every data point is color-coded according to a threshold (high quality: > 0.7; high resolution: $< 105nm$; dashed horizontal line). By splitting the graph into four quadrants using this quality (g) or resolution (j) threshold and the SVM's decision boundary (solid vertical line), we define false positive, true positive, false negative and true negative regions (quality: read clockwise from the top-left region; spatial resolution: read clockwise from the bottom-left region). An ideal system, perfectly classifying points to the right of the boundary as high quality (high resolution) and those to the left as low quality (low resolution), would achieve a recall and precision scores of 1.0. All images are 4.48 µm × 4.48 µm.

Shen *et al.* [73], originally used to understand the learned latent space of generative adversarial networks (GANs) through latent attribute manipulation. Our latent guidance mechanism allows us to apply a similar procedure and investigate whether STED-FM can encode nanoscopy features and semantics. An overview of

the process is schematized in Fig. 4c. First, images are split into two classes which represent the endpoints of some attribute of interest. Afterwards, a support vector machine (SVM) is trained to classify the image embeddings, and the *attribute direction* is computed as the normal vector to the SVM's decision boundary (Methods). The main idea is to then manipulate latent codes obtained from unseen images along this normal vector and use the resulting manipulated code as the conditioning input to the latent-guided DDPM. The DDPM should then be able to generate a version of the input image which, after manipulation, exhibits features associated with the other class.

We first demonstrate STED-FM's ability to encode specific biological structures in an experiment that aimed to translate images of one protein (F-actin) into another (tubulin) (Fig. 4d). The latent *F-actin to tubulin* direction is determined through SVM classification of F-actin and tubulin image embeddings. Images generated by the DDPM conditioned on latent codes and manipulated towards the tubulin direction progressively lose the typical F-actin nanostructures, including periodic structures and longitudinal fibers [66]. Instead, these structures transform into continuous filaments, typical of STED tubulin images. This result highlights the ability of STED-FM to encode and discriminate between these two image attributes.

Next, we aimed to probe and manipulate spatial features in super-resolution microscopy images : image quality and spatial resolution. For the experiments on image quality, an SVM was trained to classify image embeddings of low and high quality classes [61]. The distribution of training image embeddings' distances relative to the SVM's decision boundary is shown in Fig. 4e, and an example trajectory of a low-quality image being manipulated into a higher quality image is shown in Fig. 4f (Supplementary Figure 11). To quantify the model's ability to improve image quality through this latent attribute manipulation, we plot all the trajectories (Fig. 4g) showing the distance to the SVM's decision boundary and the inferred quality score at each step. The analysis of generated trajectories reveals that the distance to boundary correlates with the quality scores, achieving recall and precision scores of 0.81 (Fig. 4g). The majority of false positives and negatives lie near the decision boundary. A similar behavior is observed when applying latent attribute manipulation of the modulation of spatial resolution (Fig. 4h,i & (Supplementary Figure 12)). The high precision and recall values (> 0.9) highlight the latent-guided DDPM's capacity to decode the resolution attribute from STED-FM's latent representation (Fig. 4k). Taken together, these results show that STED-FM effectively encodes relevant features, which can in turn be manipulated to enhance interpretability.

2.4 Pattern discovery

We now use the representational encoding of STED-FM with the decoding ability of the latent-guided DDPM to *discover* how biological phenotypes evolve as we navigate in the latent space. We first explored the age-dependent remodeling of synaptic structures in a dataset consisting of STED images of the synaptic protein PSD95 in cortical neuronal cultures that were fixed at 12 and 25 days in vitro (DIV12 and DIV25; *Synaptic development dataset* [74]). After training an SVM to classify embeddings of neurons at DIV12 and DIV25, we sampled independent images from DIV12 neurons ($N = 81$) as anchors, and manipulated their latent representations towards the one of DIV25 neurons (Methods). An example of such a trajectory is shown in Fig. 5a. In Fig. 5b, we plot all trajectories, with the x-axis representing the distance to the SVM's decision boundary and the y-axis indicating the distance of each generated image's embedding to the images of DIV25 neurons (Methods). Trajectories from the anchor point (DIV12) converge towards synthetic images whose embeddings are closer to those of the real STED images of PSD95 at DIV25. We next extracted six structural features used to describe synaptic nanoclusters in super-resolution microscopy images [55] (Methods), and investigated how they evolved along the manipulation trajectories (Fig. 5c). We observe a constant increase in area, perimeter, mean nanocluster intensity, nanocluster density and number of nanodomains for PSD95 protein clusters when moving from the representations associated with neurons at DIV12 to the ones of neurons at DIV25. Image generation with STED-FM and the DDPM helps better capture features which are discriminative of the two neuronal developmental stages. It allows measurement of continuous trajectories and reduces measurement noise associated with biological variability, thus capturing how synaptic nanostructures are reorganized in an age-dependent manner (Fig. 5c & Supplementary Fig. 13). Next, we performed consensus clustering on the 384-dimensional representations of the anchors (DIV12 and DIV25) and generated images using STED-FM to extract synaptic subtypes associated with the PSD95 nanoclusters (Fig. 5d & Supplementary Fig. 14; Methods). With this approach we can track the prevalence

of each subtype across the manipulation trajectory steps. In Fig. 5e, we see that the proportion of two of these four subtypes exhibit changes along the manipulation trajectory. Analysis of the subtype trajectories confirm observations from the literature obtained on real images [75], showing that PSD95 cluster size, and nanodomain number increases for neurons fixed at DIV25, in comparison to DIV12. Specifically, subtype s_0, characterized by relatively high area, perimeter, intensity, and number of nanodomains became more frequent for the steps closer to the DIV25 anchor. Subtype s_1, characterized by smaller area, perimeter, and number of nanodomains, decreases in proportion for synthetic images of older neurons. We also detect that one subtype (s_2) associated with the smallest PSD95 clusters is the most prevalent and its proportion remains constant throughout.

Next, we assessed the model's ability to differentiate phenotypes during the activity-dependent remodeling of neuronal F-actin in dendrites [66] (Fig. 5f). We used the segmentation of F-actin rings and fibers as a proxy to validate the generation of the images [66]. As expected, latent attribute manipulation towards the high-activity condition increases the prevalence of F-actin fibers while decreasing the presence of F-actin rings. When moving in the latent space from high- to low-activity, the opposite effect was observed (Fig. 5f, g). Interestingly, the transition from high- to low-activity revealed regions with higher F-actin density (Fig. 5f, arrow heads). These regions, previously reported as compact assemblies in Lavoie-Cardinal *et al.* [66] but not extensively analyzed, prompted us to quantitatively investigate their abundance in an activity-dependent manner. We screened the dataset from Lavoie-Cardinal *et al.* [66] for compact assemblies using STED-FM's patch-retrieval capability (Methods, Fig. 5f). Increasing neuronal activity gradually decreases the prevalence of compact assemblies (Fig. 5h,i & Supplementary Fig. 15). Combining STED-FM with a powerful decoder like DDPM enables discovery of phenotypes in a data-driven manner, which may be overlooked by manual screening of the images. Moreover, this is achieved without the need to obtain expert manual annotations required for supervised deep learning-based analysis [66].

2.5 Microscopy automation

In recent years, smart microscopy has become more popular, offering the possibility to adapt microscope acquisition parameters (regions of interest, imaging parameters, reactive microscopy, etc.) to address specific research questions [61, 76–79]. The application of data-driven microscopy generally requires the development of pipelines that can automatically and reliably detect events/structures of interest. We propose to leverage the retrieval capabilities of STED-FM to automate the region selection process and enable the detection of small substructures within the field of view. We draw inspiration from recent work using foundation model features for structure segmentation in microscopy images with minimal annotations [80] and demonstrate the integration of STED-FM into an automated acquisition pipeline. Briefly, the microscopist annotates a few regions of interest (ROIs) on a confocal overview image. In this application, the F-actin signal was used to identify spines on dendrites. Patch-level features corresponding to these annotations are then extracted from the latent space of STED-FM and used to train a Random Forest classification model for robust ROI detection. Afterwards, the predicted ROIs are used for automated region selection in machine learning-assisted parameter optimization pipelines (Fig. 6a).

We present two imaging pipelines: i) optimization of excitation laser power, STED laser power, and pixel dwell time in a 3D-DyMIN [81] acquisition of the presynaptic protein Bassoon, and ii) a multi-agent reinforcement learning (RL) pipeline for the simultaneous acquisition of 3-color STED images of F-actin, Bassoon, and PSD95. For the 3D-DyMIN acquisition, we used the LinTSDiag procedure proposed in Bilodeau *et al.* [76] to simultaneously optimize three imaging optimization objectives: photobleaching, signal ratio, and resolution (Methods). As expected, the z-resolution of the acquired images improved, as demonstrated by the full width at half maximum (FWHM) measurements obtained from manually extracted z-profiles, reaching values below 200 nm in the final images (Fig. 6b). During the optimization process, the z-profile was approximated by a center-line across the summed volume, which accounts for the difference between the automatic resolution computation and the manually determined FWHM. This improvement in resolution was achieved at the cost of a reduced signal ratio, while photobleaching was effectively stabilized (Supplementary Fig. 16).

We then adapted the RL pipeline originally proposed in Bilodeau *et al.* [76] to simultaneously run an independent RL model for each fluorophore. The ROIs for this pipeline were obtained using the random forest trained on the latent representations of STED-FM. At each ROI, the RL models first selected the set

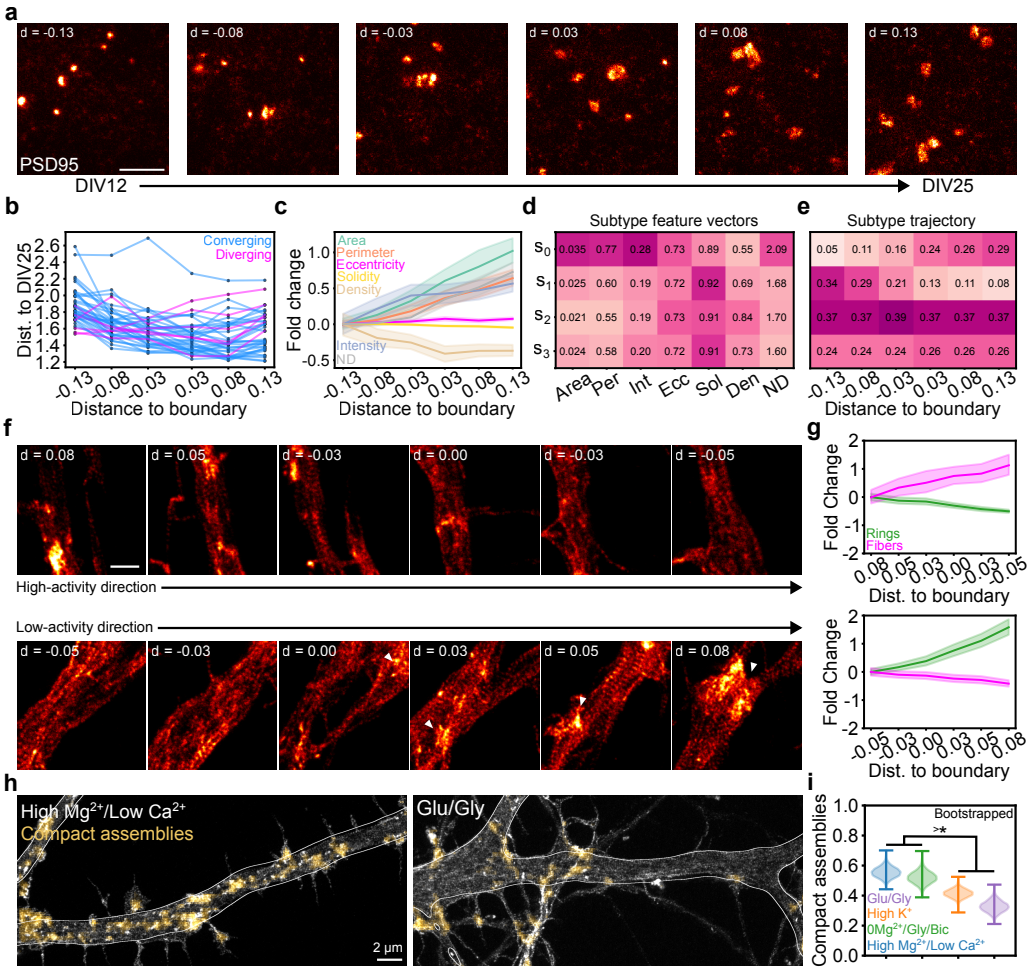


Figure 5: Discovery of structures. **a)** Latent attribute manipulation trajectory of PSD95 images from DIV12 neurons, manipulated towards images of DIV25 neurons. **b)** All *DIV12 to DIV25* trajectories. The y-axis shows the distance of each synthetic image to its 9 real nearest neighbors images of DIV25 neurons. **c)** Feature fold change along the trajectories, averaged over all trajectories. **d)** Average hand-crafted feature vectors of the different PSD95 subtypes. **e)** Proportion of PSD95 subtypes in the different trajectory time steps. **f)** Latent decoding from low- to high-activity and high- to low-activity in F-actin images. Example of a trajectory in the high-activity (top) and low-activity (bottom) directions from an anchor image. **g)** Quantification of the presence of F-actin rings (green) and fibers (magenta) depending on the direction (top: to high-activity; bottom: to low-activity). **h)** Quantification of compact assemblies from the images using a similar procedure as in Fig. 2j. **i)** The proportion of compact assemblies (Methods) is reduced with increasing activity. Scale bars: 1 µm unless otherwise noted.

of imaging parameters for their respective fluorophore. The imaging optimization objectives were evaluated, and the next ROI was processed. For all channels, the models effectively improved the signal ratio while reducing photobleaching and causing a small decrease in resolution (Fig. 6c, Supplementary Fig. 17 & 18). We also report the FWHM values obtained from manual annotations of the acquired images (Fig. 6c).

In sum, the integration of STED-FM's retrieval capabilities with algorithms designed to assist microscopists presents a novel and efficient approach to enhance acquisition throughput, enabling automated sub-region selection with minimal human interaction.

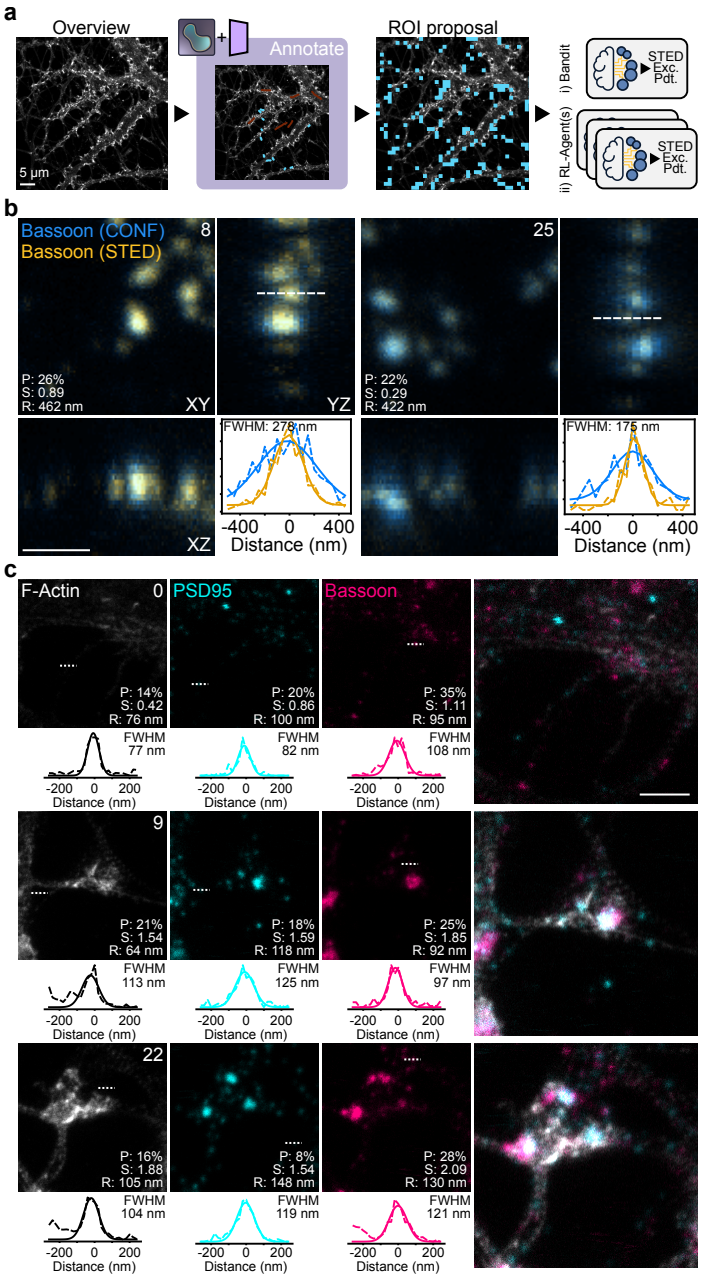


Figure 6: Automation of acquisition pipelines with STED-FM. **a)** Automatic region of interest (ROI) proposal is done by interactively training a Random Forest on the embedded patches from STED-FM on a large confocal overview (cyan: positive annotations; red: negative annotations; see Methods). In this example, spines are detected. The ROIs containing spines are used as part of the optimization pipeline of the imaging parameters. **b)** The imaging parameters of 3D DyMIN acquisition (Excitation and STED laser power, and pixel dwell time) of synaptic protein Bassoon are optimized using the Bandit framework. The imaging optimization objectives improve during the imaging session (P: Photobleaching, S: Signal Ratio, R: z-Resolution). The confocal (Blue) and DyMIN (Yellow) images of Bassoon are overlayed. The XY, XZ, and YZ plane are obtained from maximal projections. FWHM is calculated from manually traced line profiles (dashed). **c)** Imaging parameters (Excitation and STED laser powers, and pixel dwell time) are simultaneously optimized using a multi-agent RL pipeline for a 3-color STED acquisition (F-actin, PSD95, and Bassoon). The imaging optimization objectives improve during the imaging session (P: Photobleaching, S: Signal Ratio, R: z-Resolution). FWHM is obtained from manually traced line profiles. Scale bars: 1 µm unless otherwise noted.

3 Discussion

Here, we introduced STED-FM, a generalist FM for STED microscopy which was trained on a newly compiled large-scale dataset of STED images. STED-FM addresses a lack in the availability of robust FMs for super-resolution fluorescence microscopy. While various FMs exist for broader biomedical and conventional fluorescent microscopy, their efficacy on super-resolution data, as demonstrated across multiple tasks in this study, remains limited. We trained STED-FM in a self-supervised manner, using a ViT backbone with the MAE training procedure. The model learned a high-quality latent manifold with strong expressive and discriminative capabilities. This translates into strong downstream performances on a wide variety of tasks, with little to no supervision.

STED-FM's ability to encode inherent physical attributes of microscopy images, such as image quality and spatial resolution, was demonstrated. On the other hand, out-of-domain pre-training was less effective at capturing these attributes. Furthermore, our RCC experiments showcased the quality of STED-FM's latent representations. By leveraging the learned latent manifold, RCC enhances our ability to differentiate between experimental conditions. STED-FM can also be used for effective patch- and image-level retrieval, which can be readily applied to automatic extraction of rare structures in STED images or integrated into automated acquisition pipelines. We demonstrated that STED-FM's deep representations can be fine-tuned for both classification and segmentation tasks using a simple linear layer, alleviating the annotation process and reducing the size of the required training dataset in comparison to a fully-supervised training pipeline. The high performance achieved by training only this linear layer underscores the quality of the learned representation, which yields a linearly-separable feature space. We validated our approach under realistic conditions, specifically in data-scarce regimes over a wide range of tasks. Our evaluations showed that models fine-tuned from the STED-FM backbone generally outperformed ViT-S pre-trained on other large microscopy datasets. Notably, ImageNet initialization led to performance which varied considerably depending on the downstream dataset and on the number of samples used for fine-tuning. This is consistent with previous reports suggesting that ImageNet pre-training does not always improve the performance in medical imaging tasks [11]. To fully benefit from transfer learning, pre-training should ideally occur on a large dataset whose distribution closely matches that of the target domain.

Beyond direct task fine-tuning, we showed that STED-FM's deep representations can be used as building blocks for more complex analysis pipelines. We showed how the latent vector of an image can serve as a conditioning input for the training of a DDPM and how it can be manipulated to alter specific features of interest. We confirmed that the latent vector of an image extracted from STED-FM contains sufficient semantic information for high quality and faithful reconstruction by the DDPM, highlighting the expressive nature of the latent space. This latent representation can be used to manipulate physical attributes like image quality and spatial resolution, but more importantly, it facilitates discovery-driven experiments. Our results show that this pipeline can effectively interpolate between experimental conditions, generating deep features representative of real experimental outcomes. While acknowledging the need for careful validation of discovered phenotypes on a real microscope when feasible, this pipeline allowed us to identify structures missed in initial screening of F-actin images. Often obstructed and overlooked in favor of more obvious dissimilarities which are often caused by biological variability, the generated images provide researchers with a powerful tool to visualize changes associated with different experimental conditions, thereby guiding further in-depth analysis. Furthermore, STED-FM can be integrated into microscopy acquisition pipelines for automated regions of interest detection with minimal user interaction. When combined with optimization methods like Bandit or RL, this integration creates fully automated, high-throughput acquisition pipelines for specific structures of interest, assisting microscopists during their imaging session.

By making the dataset publicly available [51], we anticipate its broad adoption by the scientific community and its integration into diverse training pipelines. This collaborative effort has the potential to foster the creation of multi-modal training datasets. Such datasets, incorporating images from various super-resolution microscopy techniques, could enable the training of even more general and powerful FMs for the bioimaging field. While our evaluation of STED-FM has primarily focused on STED datasets, its competitive performance on a downstream classification task using SIM images demonstrates promising generalization capabilities to other super-resolution modalities. This suggests that STED-FM holds great potential for incorporation into a

wide range of quantitative image analysis pipelines beyond those specifically designed for STED microscopy.

In the future, we envision a continuous improvement of STED-FM through the ongoing incorporation of newly acquired STED images into its training dataset. This iterative scaling of the training data holds the potential to unlock even greater performance gains, aligning with the scaling hypotheses observed in natural language processing and natural image computer vision [1]. The currently compiled dataset predominantly features STED images from neuronal cell cultures with a notable imbalance towards specific proteins like F-actin and PSD95. Most images were acquired on the same Abberior Expert-Line STED microscope over the past decade. Increasing the diversity of the experimental conditions, sample types, and imaging facilities could improve the robustness of the learned representations to broader experimental contexts. Despite this, we are proactively addressing the inherent limitations in diversity compared to internet-scale computer vision datasets [59]. To effectively broaden the scope and variability of our training data, we have established a public repository, inviting the super-resolution microscopy community to contribute their own image datasets. By fostering this community-driven expansion, we anticipate a significant boost in STED-FM's robustness, generalizability, and overall performance, creating a more powerful tool for the entire super-resolution microscopy field.

Acknowledgments

We thank the Neuronal Cultures Platform (Plateforme de Culture Neuronale, PCN) of the CERVO Brain Research Center for the preparation of the dissociated hippocampal cultures and the Digital Research Alliance of Canada for the compute resources that were used to train and validate the deep learning models. We thank Renée Hložek for fruitful discussion on the development of foundation models and William Witteman for careful proofreading of the manuscript.

Funding was provided by grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-06171, P.D.K., RGPIN-2019-06704, F.L.C., and RGPIN-06704-2019 to F.L.C.), Fonds de Recherche Nature et Technologie (FRQNT) Team Grant (2021-PR-284335 to F.L.C, C.G., and P.D.K.), Canadian Institutes of Health Research (CIHR) (202109PJT-471107-NSB-CFBA-12805 to F.L.C. and P.D.K.), Neuronex Initiative (National Science Foundation 2014862, Fond de recherche du Québec - Santé 295824 to F.L.C. and P.D.K.), CERVO Brain Research Center Foundation (F.L.C.), the Canadian Foundation for Innovation (32786 to P.D.K. and 39088, F.L.C.), a New Frontiers in Research Fund Exploration Grant (NFRFE-2020-00933 to F.L.C., and C.G.) F.L.C. is a Canada Research Chair Tier II (CRC-2019-00126, F.L.C.) and C.G. is a CIFAR AI-Chair.

A.B. and K.L. were supported by scholarships from NSERC. K. L. was supported by a scholarship from the Institute for Intelligence and Data. F.B. was supported by scholarships from FRQNT, NeuroQuébec, and the Québec Bio-Imaging Network. A.B., R.B., and F.B. were awarded excellence scholarships from the FRQNT strategic cluster UNIQUE.

Author contributions

A.B. and F.B. developed the pipeline for the foundation model design and training. A.B. and F.B. performed the unsupervised, supervised post-training, image generation, latent attribute manipulation, and automated microscopy experiments. A.B., F.B., and F.L.C designed the user studies. J.C. acquired and annotated the F-actin conformations dataset. J.M.B and F.L.C. designed the experimental protocol and acquired the synaptic development dataset. R.B. managed the Synaptic Protein Zoo platform and aggregated the data and annotations for the corresponding dataset. A.D. acquired the image resolution dataset. K.L. performed preliminary self-supervised learning experiments. A.B., F.B., and F.L.C wrote the manuscript. P.D.K, C.G and F.L.C co-supervised the project.

Competing interests

The authors declare no competing interests.

Data availability

The in-house datasets (*Synaptic protein zoo* [65], *Synaptic development dataset* [74], *F-actin conformations dataset* [67], and the *STED-FM dataset* [51]) are publicly available from https://github.com/FLClab/STED-FM. Other datasets used in this study are available from the original publications.

Code availability

All the code used in this manuscript is open source and available at: https://github.com/FLClab/STED-FM.

4 Methods

4.1 Datasets

4.1.1 Self-supervised pre-training datasets

The *STED-FM dataset* [51] consisted of 37 387 images of varying size which were split into 224×224 crops. The resulting size of the dataset was 976 022 crops, all of which were used for pre-training STED-FM. Within the *STED-FM dataset*, the protein of interest is labeled in 238 683 crops. In total, 24 different proteins were identified in this subset, and the total number of crops per protein is reported in Supplementary Table 4.

We benchmarked STED-FM against pre-training of ViT-S on four other datasets which cover the natural image, microscopy and super-resolution microscopy domains. In order of similarity to the STED image domain: Structured Illumination Microscopy images (*SIM*; [15]), the Human Protein Atlas (*HPA*; [54]), a subset of the JUMP-CP dataset (*JUMP*; [28]), and *ImageNet-1k* [6]. For *ImageNet*, we did not re-train our own ViT-S model on the ImageNet-1k dataset, but used the pre-trained weights available from the Pytorch Image Models library [82].

For *SIM*, we used all the reconstructed images from the datasets used in [15]. The dataset is aggregated from the 3D-RCAN dataset [83, 84], the BioSR dataset [85, 86], DeepBacs dataset [87–89], and the EMTB dataset [90, 91].

In the case of *HPA*, we treated the 4 channels of an image as 4 independent images. We performed random cropping to sample 224×224 crops from the 512×512 original images during training.

For *JUMP*, we used the JUMP Cell Painting datasets [28], available from the Cell Painting Gallery [92] on the Registry of Open Data on AWS. We sampled 3M images from the different cpg-0016 sub-folders. Since images were of variable sizes, we performed random cropping of the images during training to obtain crops of size 224×224.

Each image in the dataset was normalized to the [0, 1] range using min/max values (*JUMP*, *HPA*), or the 0.001 and 0.999 quantile (*STED-FM dataset*, *SIM*). The number of independent images and the number of crops for each dataset can be found in Supplementary Table 1.

4.1.2 Supervised downstream datasets

For the different downstream experiments performed, we used a total of 15 super-resolution microscopy datasets.

Classification We used 5 downstream classification datasets: 2 in-distribution (ID) and 3 out-of distribution (OOD) (Supplementary Table 2).

- *STED optimization* (*SO*; ID): Dataset of STED microscopy images from Durand *et al.* [61] containing 4 different structures: the cytoskeleton proteins F-Actin and Tubulin, the calcium-dependent protein kinase II (CaMKII), and the post-synaptic protein PSD95. Only the images with an image quality above 0.7 were used. The quality rating was obtained by a manual expert annotation in Durand *et al.* [61]. The train and test splits from the original publication were used. A validation dataset was created

by sampling 10% of the training images from each class. The downstream task consisted in classifying the four aforementioned structures (Fig. 3).

- *Neural activity states* (*NAS*; ID): STED microscopy dataset of PSD95 imaged under 4 different experimental conditions from Wiesner *et al.* [55]. It highlights activity-dependent remodeling of PSD95 following different stimulations: high Mg^{2+}/low Ca^{2+} (low activity), $0Mg^{2+}/Gly/Bic$ (cLTP), Glutamate/Glycine (cLTD), and Tetrodotoxin (upscaling). The dataset was split into train (70%), validation (15%) and test (15%) sets. The downstream task consisted in classifying the four experimental conditions: low activity, cLTP, cLTD, and upscaling (Fig. 3).

- *Peroxisome* (*Px*; OOD): STED microscopy dataset of images containing peroxisome organelles from de Lange *et al.* [62, 93]. It contains images of yeast cells grown for four different time periods in a medium containing 0.5% methanol (MeOH): 4hMeOH, 6hMeOH, 8hMeOH, and 16hMeOH. The dataset was acquired in triplicate. We combined two replicates for training and validation, with a random split of 80%/20% respectively. The third replicate was used for testing. The downstream task consisted in classifying the four time periods (Fig. 3). Images were padded with zeros to match the required image size of 224×224 pixels.

- *Polymer rings* (*PR*; OOD): STED microscopy images from Hurtig *et al.* [63, 94] showing two proteins which self-assemble into filamentous polymers, and which are part of the cell division (Cdv) system: CdvB1 and CdvB2. We randomly split the dataset into training, validation, and testing (respectively 70%, 15%, 15%). The downstream task consisted in classifying CdvB1 vs. CdvB2 (Fig. 3). Images were padded with zeros to match the required image size of 224×224 pixels.

- *DL-SIM* (OOD): SIM images of four different sub-cellular structures obtained from Jin *et al.* [64]: Microtubules, Adhesions, Mitochondria, and F-actin. We used the high-quality reconstructed SIM images and the training and testing dataset splits from the original publication. We subsampled 30% of the training dataset to obtain a validation dataset. The downstream task consisted in classifying the aforementioned structures (Fig. 3).

Segmentation We used 2 ID and 2 OOD downstream segmentation datasets (Supplementary Table 3). Example images and ground truth segmentations are shown in Supplementary Fig. 8.

- *F-actin* (ID): STED microscopy dataset of images containing two segmentation classes: F-actin rings and F-actin fibers from Lavoie-Cardinal *et al.* [66]. The training, validation, and test splits from the original publication are used. The downstream task consisted in segmenting regions containing F-actin rings and/or fibers (Fig. 3).

- *Synaptic Protein Zoo* (*SPZ*, ID): STED microscopy dataset of PSD95 and Bassoon images from Wiesner *et al.* [55] were annotated using the citizen science platform *Zooniverse* [65]. The ground truth consists of semantic segmentation masks obtained from the majority class defined from 10 annotations per synaptic nanocluster for each task (segmentation and classification). The segmentation masks correspond to the aggregation (pixel-wise) of all voted masks, defining a probability at every pixel. Not all synaptic proteins in the images have been annotated, resulting in weak labels. The classes correspond to nanocluster morphology: round, elongated, perforated and multi-domain proteins. The downstream task consisted in the semantic segmentation of the four nanocluster morphologies (Fig. 3).

- *Foot processes* (*FP*; OOD): STED microscopy dataset of podocyte foot processes from Unnersjö-Jess *et al.* [69]. The ground truth masks are two-channel segmentations of foot process and slit diaphragm regions. We used the training and testing splits provided by the authors of the original publication. We subsampled 20% of the training set to obtain the validation dataset. The downstream task consisted in the semantic segmentation of foot process and slit diaphragm regions (Fig. 3).

- *Lioness* (OOD): STED microscopy dataset of structures and their boundaries acquired using the LIONESS procedure from Velicky *et al.* [30, 95]. The four imaging volumes from the original publication were split into two volumes for training, one for validation, and one for testing. The downstream task consisted of semantic segmentation of structures and their boundaries (Fig. 3).

Latent attribute manipulation For the latent attribute manipulation experiments (including those from Section 2.4), we used variants of datasets introduced beforehand, as well as new ones. Each dataset was split into training, validation, and test sets. The train and validation sets were used for training the SVM and obtaining the normal vector to the decision boundary. Afterwards, we sampled images from the test set to serve as anchors to the manipulation trajectories.

- *F-actin to tubulin dataset*: To train the SVM, we sampled all F-actin and tubulin images from the *SO* dataset with image quality scores above 0.7 [61]. This resulted in 1 101 images for training, 111 images for validation, and 60 F-actin test images as anchors (Fig. 4d).
- *Image quality dataset*: The ground truth quality scores (ranging from 0 to 1) for this dataset were obtained from the annotations of the original dataset in Durand *et al.* [61]. A small convolutional neural network was trained to assign a quality score to images of F-Actin. We used this network to measure the quality of the generated images. Images with quality scores lower than 0.60 were assigned to the low quality class, while those with scores higher than 0.70 belonged to the high quality class. Images with scores between these thresholds were discarded to reduce the ambiguity near the SVM's decision boundary. This resulted in 1,623 images for training, 147 for validation, and 20 low quality F-actin images as anchors (Fig. 4e-g).
- *Image resolution dataset*: The image resolution dataset consisted of images of the pre-synaptic protein Bassoon (STAR635P) imaged using two different STED laser powers [96]. The high-resolution images were acquired with a STED depletion power of 67.7 mW and the low-resolution images with a depletion power of 22.4 mW. The resulting images had average resolutions of 86 nm (high resolution) and 117 nm (low resolution; Supplementary Fig. 19). The dataset consists of 97 training images, 23 validation images, and 12 low resolution image anchors (Fig. 4h-j). The image resolution was measured using the decorrelation analysis [97].
- *Synaptic development dataset*: The synaptic development dataset consisted of STED images of PSD95 from neuronal cortical cultures fixed at : 12, 16, 18, 21, and 25 days in vitro (DIV12 to DIV 25) [74]. We used the DIV12 and DIV25 classes for the latent attribute manipulation. We filtered out images with low signal (foreground intensity < 0.15 after min-max normalization). The final dataset resulted in 311 images for training, 144 images for validation, and 38 images of the DIV12 class as anchors (Fig. 5a-e).
- *Activity-dependent F-actin dataset*: We used the F-actin activity-dependent dataset provided in Bilodeau *et al.* [20] for training the SVM. For the latent attribute manipulation experiment (Fig. 5f,g) we used the images of DIV13 neurons treated with a High Mg^{2+}/Low Ca^{2+} solution (low activity) and a glutamate/glycine solution (high activity). For the discovery experiments, we used all images from DIV13 neurons provided in Lavoie-Cardinal *et al.* [66] (Fig. 5f-i).

Image retrieval For the quantification of the image retrieval experiments, we used a dataset of F-actin images (Fig. 2j).

- *F-actin conformations dataset*: The F-actin conformations dataset consists of F-actin STED images from several rat hippocampal neuronal cultures fixed between DIV12 and DIV15 (79 images) [67]. Nanoclusters are defined as very dense small round F-actin aggregation (around 100 nm in diameter) along the dendritic shaft. Compact assemblies are large dense F-actin aggregations that are on the dendritic shaft. Spines are identified as protrusions that extend from the dendritic shaft.

4.2 Deep neural network pre-training

4.2.1 Vision Transformers - Masked autoencoders

The vision transformers (ViTs) were pre-trained on microscopy datasets using the masked autoencoding procedure introduced in [38]. In Supplementary Fig. 1, we show the training of ViTs of different sizes: tiny, small, base, and large, with a patch size of 16 in all cases. We used the architectures provided in the PyTorch Image Models library [82]. According to the results obtained, we selected the ViT-S architecture for STED-FM as it offered the best overall performance. Given an input size of 224×224, this resulted in a

sequence length of 196 for all models, excluding the class token. All models were trained for 1000 epochs with a batch size of 1024. We used the same training hyper-parameters as in the seminal implementation of the masked autoencoder (MAE) [38]. We used the AdamW optimizer [98] with a learning rate of 1.5×10^{-3}, weight decay of 0.05, and decay rates β_1 and β_2 of 0.9 and 0.95, respectively. We used a cosine annealing scheduler with a period of 20 epochs. For the configuration of the MAE, we used a masking ratio of 0.75 and a fixed decoder embedding dimension of 512 for all models. After pre-training, the lightweight decoder of the MAE procedure is removed and the ViT encoder serves as the foundation model.

4.3 Unsupervised analysis

4.3.1 ViT attention maps

To extract an attention map from the pre-trained ViT-S, the self-attention of the class token from the 6 attention heads in each of the 12 transformer blocks is averaged, resulting in 12 attention maps. The global attention map of the model is obtained by summing these 12 maps. A 0.5 quantile threshold is applied to the sum, and the thresholded attention map is then rescaled to input size. The global attention maps are overlayed on the input intensity image and shown to domain experts in a user study (Fig. 2a, b).

4.3.2 Ridge regression

We trained a ridge regression model using the `scikit-learn` library [99] to predict the spatial resolution, or image quality, given the latent feature vector of an image (Fig. 2c). A ridge regression model was trained independently for each pre-training dataset. A value of $\alpha = 1$ was used for the regularization term (l2-norm).

4.3.3 Recursive consensus clustering

We performed clustering on the deep feature space of STED-FM for the *NAS* dataset. We followed the work of Sonpatki & Shah [56], which is a recursive version of the consensus clustering algorithm introduced in Monti *et al.* [57].

Consensus Consensus clustering is a method which performs n trials of a clustering algorithm. In each trial we sample a subset representing ratio P of the dataset. These n resampling trials are performed for C different clustering algorithms or variants of the same clustering algorithm (such as K-Means with different values of K), such that the final number of trials is given by $N = n \times C$. At the end of the trials, for every possible pair of elements, the number of times the pair clusters together is divided by the number of trials in which the two elements were both part of the sampled subset. This produces a consensus matrix M. For our experiment, we use $N = 30$, $P = 0.8$, and use the K-Means clustering algorithm with values of K ranging from 2 to 10 ($C = 8$).

Recursion We determine the best number of clusters for recursion by computing the consensus matrices for the different values of K. We calculate how close each matrix is to a perfect consensus matrix (one which would have all elements either 0 or 1). The selected K^* which produces a matrix closest to a perfect one is chosen as the optimal number of clusters. It is then on those K^* clusters that recursion is performed.

Stopping The recursion is performed until all recursion paths are completed. As proposed in Sonpatki & Shah [56], this happens when clustering returns an intra-cluster stability < 0.80, or an inter-cluster overlap > 0.20. Letting S be the set of pairs that cluster together and W be the set of pairs that do not cluster together, the intra-cluster stability and inter-cluster overlap are defined as:

$$\text{Intra cluster stability} = \frac{1}{|S|} \sum_{(i,j) \in S} M_{i,j} \ ;$$

$$\text{Inter cluster overlap} = \frac{1}{|W|} \sum_{(i,j) \in W} M_{i,j} \ .$$

4.3.4 Retrieval

Image retrieval

For the image retrieval task, latent embeddings of all images within a classification dataset (*SO*, *NAS*, *Px*, *PR*, *DL-SIM*) were generated using the STED-FM model. Subsequently, an iterative retrieval process was performed. In each iteration, a single image embedding served as the *template*. The cosine similarity was computed between this template embedding and all other image embeddings in the dataset. These similarity scores are then used as prediction scores for performance evaluation.

More specifically, to evaluate the retrieval performance, the Area Under the Receiver Operating Characteristic curve (AUROC [100]) was employed. The ROC curve, in this binary classification context, was constructed by calculating the true and false positive rates across a range of similarity score thresholds (Fig. 2f). For each template, a one-vs-all approach was adopted, where the positive class consisted of images belonging to the same class as the template, and all other images constituted the negative class. The AUROC was calculated for each template and averaged across all images for all classification datasets to provide a single metric performance. On all downstream datasets, the performance of a model is reported as the difference between its performance and the best model's performance.

Patch retrieval

The patch retrieval task follows a similar procedure, but instead of considering image embeddings as templates and candidates, we use patch embeddings, which can be easily retrieved from the ViT architecture. We start by choosing a patch in an image (or a collection of patches) containing a structure of interest, and extract its embedding from ViT-S as template. In the case of a structure of interest spanning more than a single patch (a region of more than 16 pixels by 16 pixels), we average the patches in which the structure accounts for more than 50% of the patch's pixels. For all images of a given dataset, we gather patch embeddings and rank them in order of similarity to the template patch using cosine distance.

We performed the patch retrieval experiment with all pre-trained models, and for a given template, only keep the top 10 candidates from each model (50 images in total from 5 pre-trained models). For quantification we performed a user study in which experts were asked to rank the 50 images (pairwise ranking, corresponding to the 50 most similar embeddings set) according to their similarity to the template image. This produced an expert-defined similarity ranking from which we extracted the top 10 most similar images. We counted the proportion of these 10 candidates which were produced by each of the different pre-trained models (Fig. 2h,i), which is used to quantify patch retrieval performance.

4.3.5 Automatic patch retrieval experiment

The automatic patch retrieval experiment was performed on the *F-actin conformation dataset* [67]. The template was obtained by manually annotating a structure of interest in an image. In the case of a template spanning more than a single patch, we averaged the patches in which the template accounts for more than 50% of the patch's pixels.

This template embedding was compared to the embeddings of all patches on large field of views. The large field of view was cropped by extracting overlapping crops of 224×224 pixels (25% overlap). To avoid potential orientation bias of the selected structure of interest, the initial template was augmented through rotation. Four distinct template orientations were generated by applying successive 90° rotations. Each of these four rotated template embeddings was then compared against the embeddings of the image crops from the field of view. The maximum similarity score (cosine distance) achieved between any of the four templates and an image patch was used for that patch. For overlapping crops the similarity score is averaged. The resulting maximum projection heatmap is upscaled to input size and overlayed on the large of field of view in Fig. 2j.

The capacity of the pre-trained models to accurately identify relevant regions was evaluated against a manually annotated dataset. The dataset was composed of 79 manually annotated images for three structures: nanoclusters, compact assemblies, and spines. The performance of each pre-trained model is compared to a baseline detection method employing class-specific intensity thresholds. The primary evaluation metric was the difference in precision between a pre-trained model (Pr) and the thresholding method ($\mathrm{Pr}_{\mathrm{thresh.}}$), analyzed over the range of recall. Additionally, the performance of a random baseline is reported. This random model simulated chance-level detection by randomly reassigning the expert annotations within the foreground pixels of the field of view of the image.

4.4 Fine-tuning

4.4.1 Classification

Downstream classification performance is evaluated by adding a linear layer with batch normalization on top of the pre-trained encoders. We then either keep the pre-trained features frozen and only fine-tune the linear layer (linear probing), or allow for end-to-end fine-tuning (fine-tuning). All models are trained with the cross-entropy loss. Early stopping is used to checkpoint the best validation loss during fine-tuning.

Linear probing We use stochastic gradient descent (SGD) with a linear warmup schedule that takes the learning rate from 0.0 to 0.001 in the first 30 epochs. Afterwards, a cosine annealing scheduler [101] with a period of 30 epochs is applied for the remaining 270 epochs, for a total of 300 epochs of fine-tuning. Cosine annealing is done with maximum and minimum learning rate values of 10^{-3} (the value reached after the linear warmup) and 10^{-5}, respectively.

Fine-tuning The models are fine-tuned for 300 epochs. The Adam optimizer is used. The scheduler uses a linear warmup schedule that takes the learning rate from 0.0 to 10^{-4} in the first 30 epochs. A cosine annealing scheduler is applied for the remaining training epochs, annealing the learning rate from 10^{-4} and 10^{-6}.

Random initialization - from scratch We also trained the models from scratch (random initialization) using a fully-supervised training scheme with the annotated data associated with the downstream task. For this configuration we used the Adam optimizer with a linear warmup scheduler for 30 epochs followed by a cosine decay scheduler over the remaining epochs.

4.4.2 Segmentation

We evaluate segmentation performance by appending a linear layer to the pre-trained encoder. Letting B be the batch size, N_p the number of patches, D_p the patch embedding dimension, P the patch size (in pixels, assuming square patches) and C the number of channels to predict (classes), the encoder outputs a tensor of dimensions (B, N_p, D_p). The linear layer is trained to map these patch embeddings to pixel values (in the range [0, 1]), transforming the encoder's output (excluding the class token embedding, hence $N_p - 1$ patches) to a tensor of dimension $(B, N_p - 1, P^2C)$. The output of this linear layer is then reconstructed into image space (a process called "unpatchifying") to match the original input dimensions, enabling the computation of a Mean Squared Error (MSE) loss against the ground truth segmentation mask. For classification, we consider linear probing and fine-tuning. For all training, early stopping is used to checkpoint the best validation loss during fine-tuning. The models are trained for 300 epochs.

Linear Probing When using linear probing, we use SGD with a linear warmup of 30 epochs, followed by cosine annealing for the remaining epochs using maximum and minimum learning rates of 10^{-3} and 10^{-5}, respectively, along with a period of 30 epochs.

Fine-tuning For end-to-end fine-tuning we use the same schedulers, this time removing the 30-epoch period and replacing the SGD optimizer with the AdamW optimizer, using a learning rate of 10^{-3}, weight decay of 0.1 and β parameters of 0.9 and 0.95.

Random initialization - from scratch When training from scratch, only the hyper-parameters of the AdamW optimizer are different from the end-to-end case, with learning rate of 10^{-3}, weight decay of 0.05, and β parameters of 0.9 and 0.999.

4.4.3 Small data regime

For both classification and segmentation, we also considered the small data regime, in which the number of annotated instances for fine-tuning is limited. For classification, we used {10, 25, 50, 100} annotated instances per class for the linear probing or fine-tuning steps. For datasets with insufficient instances to meet these counts, the maximum available number of instances per class was used (Supplementary Table 2). For segmentation tasks, we considered a percentage of training samples rather than a raw count, using a ratio of {0.01, 0.1, 0.25, 0.50} of available annotated samples for linear probing or fine-tuning steps.

To reduce variability in performance caused by different number of weight updates when using smaller datasets, we adopted a constant training budget. The budget was set to match the number of parameter updates used in the full dataset fine-tuning scenario, with an upper limit of 1000 training epochs. More specifically, we define the training budget in the full dataset scenario as $B_f = N_t \times N_e$, where N_t is the number of annotated training instances and N_e is the fixed number of training epochs. The small data regime training budget $B_{s,\text{classif}}$ in classification is computed as:

$$B_{s,\text{classif}} = \min\left(1000, \frac{B_f}{N_c \cdot C}\right)$$

where N_c is the number of training instances per class and C is the number of classes. In segmentation, the training budget $B_{s,\text{seg}}$ is computed as:

$$B_{s,\text{seg}} = \min\left(1000, \frac{B_f}{P \cdot N_t}\right)$$

where P is the percentage of training samples. The optimizers, loss functions, and hyper-parameters (other than those pertaining to training epochs) were identical to those described in the full-dataset experiments.

4.5 Image generation

4.5.1 Latent-guided denoising diffusion probabilistic models (DDPMs)

Following the latent guidance framework of Preechakul *et al.* [72], we conditioned the DDPM using a 384-dimensional latent vector $\mathbf{z}$ obtained by passing an image through STED-FM. To condition the DDPM, we map $\mathbf{z}$ to the dimensionality of the timestep embedding's $\mathbf{t}$ latent space (256 dimensions). The resulting embeddings are then added ($\mathbf{z} + \mathbf{t}$, Fig. 4a), and serve as the conditioning input. Once the latent-guided DDPM is trained, a noise image $\mathbf{x_T}$ can be given to the DDPM along with the latent code $\mathbf{z}$ of a test image $\mathbf{x_0}$. The DDPM generates a new image which preserves the features present in image $\mathbf{x_0}$.

For classifier-free guidance [71], a conditioning vector was obtained by creating an embedding layer for 24 protein classes that are defined in a subset of the *STED-FM dataset* (Supplementary Table 4), producing a 256-dimensional vector which matches the timestep embedding dimension. The latent vector is then added to the timestep embedding for conditioning ($\mathbf{z} + \mathbf{t}$, Fig. 4a).

4.5.2 Latent attribute manipulation

For latent attribute manipulation, we followed the method from Shen *et al.* [73], originally developed for generative adversarial networks (GANs) and later adapted to DDPMs by Preechakul *et al.* [72]. The method manipulates image semantics in latent space using the normal vector of a SVM's decision boundary trained to differentiate between two classes (Fig. 4c). The SVM's decision boundary is a hyperplane in d-dimensional space, from which we can extract its normal vector $\mathbf{n} \in \mathbb{R}^d$. For an image embedding $\mathbf{z}$ in this latent space, we compute the signed distance from $\mathbf{z}$ to the hyperplane as

$$d(\mathbf{n}, \mathbf{z}) = \mathbf{n}^\top \mathbf{z} + b \,, \tag{1}$$

where b is the intercept of the decision boundary. To manipulate the image embedding $\mathbf{z}$ in the direction of the normal vector to a target distance d_t from the hyperplane, we update it along the direction of the normal vector $\mathbf{n}$

$$\mathbf{z}' = \mathbf{z} + (d_t - d(\mathbf{n}, \mathbf{z})) \cdot \mathbf{z} \,. \tag{2}$$

Manipulating image semantics requires partitioning of an image dataset into two classes that represent opposing ends of the desired semantic attributes. In our case, the images are embedded using STED-FM, and a SVM is trained to classify the image embeddings. The direction of the normal vector to the SVM's hyperplane corresponds to the semantic attribute direction. We can then investigate how the image has changed as a result of the manipulation by decoding the manipulated embedding $\mathbf{z}'$ back to pixel space using the conditioned DDPM described in the previous section.

4.6 Pattern discovery

4.6.1 Synaptic development of PSD95

To investigate age-dependent changes in PSD95 morphology, we used the *Synaptic development dataset* [74]. A SVM was trained to discriminate between the synaptic structures from neurons fixed at two different time points: DIV12 and DIV25. The normal vector to the decision boundary in the latent space was then used to manipulate the latent representation of an image of PSD95 at DIV12 neurons towards images corresponding to PSD95 in neurons at DIV25. For each DIV12 anchor image, we generated a series of five interpolated latent codes at constant intervals between the average distance of the DIV12 and DIV25 classes from the SVM boundary. These manipulated latent codes were fed to the DDPM to generate the corresponding images, which were used for further analysis.

To quantify the morphological changes induced by the latent space manipulation, we calculated the average cosine distance between the top-10 KNN of the generated images and the real images of the opposing class (DIV25). Protein clusters were identified with a wavelet segmentation method, similar to the approach described in Wiesner *et al.* [55]. Several morphological and intensity features were manually extracted from the images: area, perimeter, eccentricity, solidity, density, intensity, and number of nanodomains. Except for density, which is an image-level feature, protein cluster-level features were averaged across all identified protein clusters within each image. To measure the general tendency of each feature's evolution across the manipulated steps, the feature trajectories were bootstrapped 1,000 times. The average and standard deviation of the bootstrapped feature trajectories are presented (Fig. 5c).

To gain a deeper understanding of the features that best discriminate between PSD95 nanostructures at DIV12 and DIV25, we performed a consensus clustering analysis. K-means clustering was applied by varying the number of clusters (K) from 2 to 25. A pair-wise consensus matrix was constructed, indicating the frequency with which each pair of data points was clustered together out of the 23 independent K-Means runs. Hierarchical clustering was then applied on the consensus clustering, from which we were able to identify four subtypes of PSD95 synaptic clusters in the images. The proportion of each subtype was measured at each step of the latent space manipulation.

4.6.2 F-actin nanostructures

The activity-dependent remodeling of F-actin from rings to fibers was investigated using the training and validation datasets provided in Bilodeau *et al.* [20]. To ensure unambiguous training, image crops containing both F-actin rings and fibers were excluded from the dataset used to train the SVM boundary.

For image generation, the images of the testing dataset provided in Lavoie-Cardinal *et al.* [66], corresponding to High Mg^{2+}/Low Ca^{2+} (low-activity) and Glu/Gly (high-activity), conditions were used as anchors. For each crop in the testing dataset, its latent code was manipulated along the normal vector of the SVM boundary. We generated a series of five interpolated latent codes at constant intervals between the average distance of the low and high activity classes from the SVM boundary. Five images were generated for each anchor with uniform step sizes along the manipulation vector using the DDPM conditioned on the manipulated embeddings. To validate the morphological changes, we used the pre-trained U-Net model from Lavoie-Cardinal *et al.* [66] to quantify the evolution of segmented F-actin rings and fibers as a function of image manipulation.

F-actin compact assemblies Given the apparent prevalence of F-actin compact assemblies when manipulating the latent representation towards the low-activity condition, we conducted a more in-depth analysis of this pattern. The patch retrieval capability of the model was used. A template of an F-actin compact assembly was manually annotated from a real image. The same procedure described in Automatic patch retrieval experiment was then applied to generate similarity maps to this template across the activity-dependent F-actin dataset. To obtain segmentation masks of the compact assemblies from the similarity maps, a threshold of 0.65 was applied. This threshold was visually validated by an expert. The area proportion of F-actin compact assemblies was quantified within dendritic regions identified using the MAP2 signal. Any segmented region that touched the dendrite mask was included in the proportion calculation. The bootstrap distribution of the average proportion was calculated using 10,000 resamplings and is reported in Fig. 5i.

4.7 Automated Microscopy

The automated microscopy experiment proceeded as follows:

1. A confocal overview image of the sample is acquired on the microscope.
2. Regions of interests (ROIs) are identified and extracted from the overview using the developed napari-plugin which uses the STED-FM model to extract the latent vectors of image patches.
3. Each extracted ROI is subsequently imaged using STED microscopy. The imaging parameters for each ROI are selected by machine learning optimization routines, such as Bandit optimization or reinforcement learning (RL) models [76].

4.7.1 STED microscopy experiments

STED imaging. Super-resolution imaging was done on an Abberior Expert-Line STED system (Abberior Instruments GmbH, Germany). This system was equipped with a 100x oil immersion objective lens with a numerical aperture of 1.4 (Olympus, model UPLSAPO100XO), a motorized stage, and an auto-focus unit. Three excitation wavelengths were used (485 nm, 561 nm, and 640 nm) for the experiments. The system is equiped with a 775 nm depletion laser (40 MHz, 1.2 ns pulse duration; MPB Communications). All lasers were operated at 40 MHz. The emitted fluorescence photons were detected by avalanche photodiode detectors (APDs) through a pinhole set at approximately 1 Airy unit. Red dyes were detected with a FF02-615/20-25 filter (Semrock) and far-red dyes were detected with a ET685/70 filter (Chroma). The acquired images were visualized and processed using the FIJI software (ImageJ, version 1.54p) or Python.

Neuronal cell culture. Extraction and culture of hippocampal neuronal cells was performed following the protocol of Nault & De Koninck [102] and in accordance to guidelines of the animal care committee of Université Laval. Briefly, neuronal cultures from the hippocampus were obtained using neonatal Sprague-Dawley rats. P0-P1 rats were sacrificed by decapitation before dissection of the hippocampi. Cells were dissociated using enzymatic digestion and mechanical dissociation, and then seeded onto 12 mm coverslips coated with poly-D-lysine and laminin (40,000 cells/coverslip). Neurons were cultivated in a growth medium composed of Neurobasal and B27 (in a 50:1 ratio), enriched with penicillin/streptomycin (25 U/mL; 25 μg/mL) and 0.5 mM L-GlutaMAX (by Invitrogen). After 5 days, Ara-C (5 μM; from Sigma-Aldrich) was added into the medium to limit the proliferation of non-neuronal cells. Cultures were fed twice a week, by replacing $\sim$50% of the growth medium with serum- and Ara-C–free medium.

Sample preparation and staining procedures. Neuronal cultures were fixed for 20 minutes in 4% PFA solution (PFA 4%, Sucrose 4%, Phosphate Buffer 100mM, Na-EGTA 2mM). Permeabilization and blocking of aspecific binding sites were done simultaneously with a solution of Triton X-100 and 2% goat serum in PBS 20 mM, for 30 min. Primary and secondary antibodies were successively incubated for 2h and 1h respectively. Labeling of F-actin with fluorescently-tagged phalloidin was performed in a subsequent step, for 2 h. Incubation of antibodies was performed in the blocking solution at room temperature. Immunostained coverslips were mounted in PVA with DABCO for imaging. Supplementary Table 5 & 6 summarizes the antibodies used in this study with associated concentrations. F-actin was stained with phalloidin-STAR635 (Sigma Aldrich, cat. 30972-20µg, 1:50 dilution).

4.7.2 Interactive ROI selection

For interactive ROI selection, a random forest model was trained on patch embeddings of the overview confocal image generated by STED-FM. This approach is inspired by the work of Seifi *et al.* [80]. The interactive annotation was implemented in a napari-plugin, allowing users to annotate as many structures of interest as desired. In the experiments, the regions containing spines (or not) were annotated.

Since STED-FM requires images of 224×224 pixels, the overview image was tiled with a 50% overlap and fed to the model. A 16×16 pixel patch within these crops was considered annotated if at least 25% of its area had been annotated. The extracted patch embeddings and corresponding annotations were split into training and validation sets with an 80/20 ratio. The napari-plugin provides user-configurable parameters

for the random forest model such as the number and maximum depth of trees. During the experiments, a configuration of 100 trees with a maximum depth of 5 was used.

To extract individual ROIs from the binary masks predicted by the random forest model, the following procedure was applied: 1) the predicted mask was converted into a labeled image, where each contiguous region (island) of positive pixels was assigned a unique ID, 2) the centroid of the bounding box center of each island was obtained by using `regionprops` from the `scikit-image` Python library [103]. In the experiments, ROIs of a constant, pre-defined size were used and centered on the centroid of the bounding box.

4.7.3 Bandit optimization

3D-DyMIN acquisitions were performed on neuronal cultures stained for Bassoon and F-actin. For interactive ROI selection using STED-FM as previously described, spines were chosen from a confocal overview of the F-actin. At each selected location, a volume of $2.88\,\mu m \times 2.88\,\mu m \times 2\,\mu m$ was acquired. Only the STED and excitation powers, and pixel dwell time were optimized. Supplementary Table 8 provides the range of each parameter. DyMIN parameters were set at a threshold of 1 photon detected within 20% of the pixel dwell time for steps 1 and 2 [81]. The imaging optimization objectives were photobleaching, signal ratio and z-Resolution. Photobleaching and signal ratio were calculated following the procedure described in Bilodeau *et al.* [76]. Z-resolution was approximated by fitting a Gaussian function to the line profile obtained from the central slices (10 above and 10 below the center) of the maximal-projection of each frame. The contextual LinTSDiag model was used for optimization with the default parameters of Bilodeau *et al.* [76].

4.7.4 Reinforcement learning

3-color STED acquisitions were done on neurons stained for Bassoon, PSD95, and F-actin. For each channel, a dedicated RL model, the same proximal policy optimization (PPO) model trained and used in Bilodeau *et al.* [76], was employed for parameter optimization. The range of imaging parameters for each channel is provided in Supplementary Table 9. The imaging session followed the same methodology and default parameters as in Bilodeau *et al.* [76]. Briefly, ROIs were obtained using STED-FM as previously described from an initial confocal image, spines were chosen from the F-actin signal. A confocal acquisition of a ROI was then fed to the corresponding RL model. Based on the acquisition history, the model selected the next imaging parameters, and a STED image was acquired. A second confocal image was then acquired to quantify photobleaching. The imaging optimization objectives are measured from the acquired images and fed back to the RL model to inform future parameter selections. This iterative process was repeated for a total of 30 ROIs with the STED images acquired in this order: F-actin, PSD95, and Bassoon.

4.8 Resampling statistics

We employed resampling to verify statistical differences between two groups [104]. The detailed procedure is described in Bilodeau *et al.* [20]. When analyzing more than two groups, the approach involved resampling of the F-statistic for 10,000 iterations [104]. We tested the null hypothesis that all groups originate from the same underlying mean distribution, rejecting it if the calculated p-value is below 0.05. Upon rejection, we conducted pairwise comparisons between each group using a two-group randomization, also performed with 10,000 repetitions [104]. This test randomly reassigned the data points between the two groups being compared and calculated the difference in their means. We then assessed the likelihood of observing our actual mean difference under the assumption of no true difference (the null hypothesis), again using a significance level of 0.05 to determine statistical significance.

References

1. Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
2. Bommasani, R. *et al. On the Opportunities and Risks of Foundation Models* July 2022.
3. Brown, T. *et al. Language Models Are Few-Shot Learners* in *Advances in Neural Information Processing Systems* **33** (Curran Associates, Inc., 2020), 1877–1901.

4. DeepSeek-AI *et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* Jan. 2025.

5. Radford, A. *et al. Learning Transferable Visual Models From Natural Language Supervision* in *Proceedings of the 38th International Conference on Machine Learning* (PMLR, July 2021), 8748–8763.

6. Deng, J. *et al. Imagenet: A large-scale hierarchical image database* in *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.

7. Sun, C., Shrivastava, A., Singh, S. & Gupta, A. *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era* in *2017 IEEE International Conference on Computer Vision (ICCV)* (IEEE, Venice, Oct. 2017), 843–852. ISBN: 978-1-5386-1032-9.

8. Zhai, X., Kolesnikov, A., Houlsby, N. & Beyer, L. *Scaling Vision Transformers* in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New Orleans, LA, USA, June 2022), 1204–1213. ISBN: 978-1-6654-6946-3.

9. Kirillov, A. *et al. Segment Anything* Apr. 2023.

10. Abnar, S., Dehghani, M., Neyshabur, B. & Sedghi, H. *Exploring the Limits of Large Scale Pre-Training* in *International Conference on Learning Representation* (2022).

11. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. in *Advances in Neural Information Processing Systems 32* (eds Wallach, H. *et al.*) 3347–3357 (Curran Associates, Inc., 2019).

12. Azizi, S. *et al.* Robust and Data-Efficient Generalization of Self-Supervised Machine Learning for Diagnostic Imaging. *Nature Biomedical Engineering* **7,** 756–779 (June 2023).

13. Zhao, T. *et al.* A Foundation Model for Joint Segmentation, Detection and Recognition of Biomedical Objects across Nine Modalities. *Nature Methods* **22,** 166–176 (Jan. 2025).

14. Zhang, K. *et al.* A Generalist Vision–Language Foundation Model for Diverse Biomedical Tasks. *Nature Medicine* **30,** 3129–3141 (Nov. 2024).

15. Ma, C., Tan, W., He, R. & Yan, B. Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration. *Nature Methods,* 1–10 (2024).

16. Wang, E. Y. *et al.* Foundation Model of Neural Activity Predicts Response to New Stimulus Types. *Nature* **640,** 470–477 (Apr. 2025).

17. Wang, J. *et al.* Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine* **31,** 609–617 (2025).

18. Huang, S.-C. *et al.* Self-Supervised Learning for Medical Image Classification: A Systematic Review and Implementation Guidelines. *npj Digital Medicine* **6,** 1–16 (Apr. 2023).

19. Singh, P. *et al.* Shifting to Machine Supervision: Annotation-Efficient Semi and Self-Supervised Learning for Automatic Medical Image Segmentation and Classification. *Scientific Reports* **14,** 10820 (May 2024).

20. Bilodeau, A. *et al.* Microscopy Analysis Neural Network to Solve Detection, Enumeration and Segmentation from Image-Level Annotations. *Nature Machine Intelligence* **4,** 455–466 (May 2022).

21. Beaupré, F. *et al.* Quantitative analysis of miniature synaptic calcium transients using positive unlabeled deep learning. *Digital Discovery* **4,** 105–119 (2025).

22. Arevalo, J. *et al.* Evaluating Batch Correction Methods for Image-Based Cell Profiling. *Nature Communications* **15,** 6516 (Aug. 2024).

23. Kuş, Z. & Aydin, M. MedSegBench: A Comprehensive Benchmark for Medical Image Segmentation in Diverse Data Modalities. *Scientific Data* **11,** 1283 (Nov. 2024).

24. Yang, J. *et al.* MedMNIST v2 - A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification. *Scientific Data* **10,** 41 (Jan. 2023).

25. Allen, E. J. *et al.* A Massive 7T fMRI Dataset to Bridge Cognitive Neuroscience and Artificial Intelligence. *Nature Neuroscience* **25,** 116–126 (Jan. 2022).

26. Schäfer, R. *et al.* Overcoming Data Scarcity in Biomedical Imaging with a Foundational Multi-Task Model. *Nature Computational Science,* 1–15 (July 2024).

27. Jumper, J. *et al.* Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **596,** 583–589 (Aug. 2021).

28. Chandrasekaran, S. N. *et al.* JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *BioRxiv,* 2023–03 (2023).

29. Pontén, F., Jirström, K. & Uhlen, M. The Human Protein Atlas—a tool for pathology. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* **216,** 387–393 (2008).

30. Velicky, P. *et al.* Dense 4D nanoscale reconstruction of living brain tissue. *Nature Methods* **20,** 1256–1265 (2023).

31. Archit, A. *et al.* Segment Anything for Microscopy. *Nature Methods* **22,** 579–591 (Mar. 2025).

32. Kraus, O. *et al. Masked Autoencoders for Microscopy Are Scalable Learners of Cellular Biology* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), 11757–11768.

33. Gidaris, S., Singh, P. & Komodakis, N. *Unsupervised Representation Learning by Predicting Image Rotations* in *International Conference on Learning Representations* (Feb. 2018).

34. Oord, A. v. d., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

35. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. *Momentum contrast for unsupervised visual representation learning* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), 9729–9738.

36. Balestriero, R. *et al. A Cookbook of Self-Supervised Learning* June 2023.

37. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. *A simple framework for contrastive learning of visual representations* in *International conference on machine learning* (2020), 1597–1607.

38. He, K. *et al. Masked autoencoders are scalable vision learners* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), 16000–16009.

39. Sauer, A., Tian, Y., Bewersdorf, J. & Rittscher, J. *Refining Biologically Inconsistent Segmentation Masks with Masked Autoencoders* in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (June 2024), 6904–6912.

40. Zhou, L. *et al. Self Pre-Training with Masked Autoencoders for Medical Image Classification and Segmentation* in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (Apr. 2023), 1–6.

41. He, Y. *et al.* Self-Supervised Deep-Learning Two-Photon Microscopy. *Photonics Research* **11,** 1–11 (Jan. 2023).

42. Wang, Y. *et al.* Image Denoising for Fluorescence Microscopy by Supervised to Self-Supervised Transfer Learning. *Optics Express* **29,** 41303–41312 (Dec. 2021).

43. Qu, L. *et al.* Self-Inspired Learning for Denoising Live-Cell Super-Resolution Microscopy. *Nature Methods* **21,** 1895–1908 (Oct. 2024).

44. Chen, R. *et al.* Single-Frame Deep-Learning Super-Resolution Microscopy for Intracellular Dynamics Imaging. *Nature Communications* **14,** 2854 (May 2023).

45. Zhang, G. *et al.* Bio-Friendly Long-Term Subcellular Dynamic Recording by Self-Supervised Image Enhancement Microscopy. *Nature Methods* **20,** 1957–1970 (Dec. 2023).

46. Kim, J. *et al.* Self-Supervised Machine Learning Framework for High-Throughput Electron Microscopy. *Science Advances* **11,** eads5552 (Apr. 2025).

47. Midtvedt, B. *et al.* Single-Shot Self-Supervised Object Detection in Microscopy. *Nature Communications* **13,** 7492 (Dec. 2022).

48. Robitaille, M. C., Byers, J. M., Christodoulides, J. A. & Raphael, M. P. Self-Supervised Machine Learning for Live Cell Imagery Segmentation. *Communications Biology* **5,** 1–8 (Nov. 2022).

49. Hell, S. W. *et al.* The 2015 Super-Resolution Microscopy Roadmap. *Journal of Physics D: Applied Physics* **48,** 443001. (2020) (Oct. 2015).

50. Mund, M. & Ries, J. How Good Are My Data? Reference Standards in Superresolution Microscopy. *Molecular Biology of the Cell* **31,** 2093–2096 (Sept. 2020).

51. Bilodeau, A., Beaupré, F. & Lavoie-Cardinal, F. *STED-FM dataset* (Zenodo, 2025). https://doi.org/10.5281/zenodo.15608143.

52. Dosovitskiy, A. *et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale* in *International Conference on Learning Representations* (Oct. 2020).

53. Wolf, D. *et al.* Self-Supervised Pre-Training with Contrastive and Masked Autoencoder Methods for Dealing with Small Datasets in Deep Learning for Medical Imaging. *Scientific Reports* **13,** 20260 (Nov. 2023).

54. Thul, P. J. *et al.* A Subcellular Map of the Human Proteome. *Science* **356,** eaal3321 (May 2017).

55. Wiesner, T. *et al.* Activity-dependent remodeling of synaptic protein organization revealed by high throughput analysis of STED nanoscopy images. *Frontiers in neural circuits* **14,** 57 (2020).

56. Sonpatki, P. & Shah, N. Recursive Consensus Clustering for novel subtype discovery from transcriptome data. *Scientific reports* **10,** 11005 (2020).

57. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* **52,** 91–118 (2003).

58. Caron, M. *et al. Emerging properties in self-supervised vision transformers* in *Proceedings of the IEEE/CVF international conference on computer vision* (2021), 9650–9660.

59. Oquab, M. *et al.* Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

60. Chen, C. *et al.* Fast and scalable search of whole-slide images via self-supervised deep learning. *Nature Biomedical Engineering* **6,** 1420–1434 (2022).

61. Durand, A. *et al.* A machine learning approach for online automated optimization of super-resolution optical microscopy. *Nature communications* **9,** 5247 (2018).

62. De Lange, E. M., Mol, F. N., van der Klei, I. J. & Vlijm, R. STED super-resolution microscopy unveils the dynamics of Atg30 on yeast Pex3-labeled peroxisomes. *Iscience* **27** (2024).

63. Hurtig, F. *et al.* The patterned assembly and stepwise Vps4-mediated disassembly of composite ESCRT-III polymers drives archaeal cell division. *Science advances* **9,** eade5224 (2023).

64. Jin, L. *et al.* Deep Learning Enables Structured Illumination Microscopy with Low Light Levels and Enhanced Speed. *Nature Communications* **11,** 1934 (Apr. 2020).

65. Wiesner, T. *et al. Synaptic protein segmentation - Zooniverse - STED-FM* (Zenodo, May 2025). https://doi.org/10.5281/zenodo.15480470.

66. Lavoie-Cardinal, F. *et al.* Neuronal activity remodels the F-actin based submembrane lattice in dendrites but not axons of hippocampal neurons. *Scientific reports* **10,** 11960 (2020).

67. Chabbert, J. *et al. F-actin conformations - STED-FM* (Zenodo, June 2025). https://doi.org/10.5281/zenodo.15608267.

68. van Bommel, B., Konietzny, A., Kobler, O., Bär, J. & Mikhaylova, M. F-actin Patches Associated with Glutamatergic Synapses Control Positioning of Dendritic Lysosomes. *The EMBO Journal* **38,** e101183 (Aug. 2019).

69. Unnersjö-Jess, D. *et al.* Deep learning–based segmentation and quantification of podocyte foot process morphology suggests differential patterns of foot process effacement across kidney pathologies. *Kidney International* **103,** 1120–1130 (2023).

70. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33,** 6840–6851 (2020).

71. Ho, J. & Salimans, T. *Classifier-Free Diffusion Guidance* in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications* (Dec. 2021).

72. Preechakul, K., Chatthee, N., Wizadwongsa, S. & Suwajanakorn, S. *Diffusion autoencoders: Toward a meaningful and decodable representation* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), 10619–10629.

73. Shen, Y., Yang, C., Tang, X. & Zhou, B. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence* **44,** 2004–2018 (2020).

74. Bellavance, J.-M., Beaupré, F., Bilodeau, A., Sephton, C. & Lavoie-Cardinal, F. *Synaptic development - STED-FM* (Zenodo, June 2025). https://doi.org/10.5281/zenodo.15490254.

75. Metzbower, S. R., Levy, A. D., Dharmasri, P. A., Anderson, M. C. & Blanpied, T. A. Distinct SAP102 and PSD-95 Nano-organization Defines Multiple Types of Synaptic Scaffold Protein Domains at Single Synapses. *The Journal of Neuroscience* **44,** e1715232024. (2025) (June 2024).

76. Bilodeau, A. *et al.* Development of AI-assisted Microscopy Frameworks through Realistic Simulation with pySTED. *Nature Machine Intelligence,* 1–19 (Sept. 2024).

77. Mahecic, D. *et al.* Event-Driven Acquisition for Content-Enriched Microscopy. *Nature Methods,* 1–6 (Sept. 2022).

78. Alvelid, J., Damenti, M., Sgattoni, C. & Testa, I. Event-Triggered STED Imaging. *Nature Methods,* 1–8 (Sept. 2022).

79. Morgado, L., Gómez-de-Mariscal, E., Heil, H. S. & Henriques, R. The Rise of Data-Driven Microscopy Powered by Machine Learning. *Journal of Microscopy* **295,** 85–92 (2024).

80. Seifi, M. *et al. FeatureForest: The Power of Foundation Models, the Usability of Random Forests* Dec. 2024.

81. Heine, J. *et al.* Adaptive-Illumination STED Nanoscopy. *Proceedings of the National Academy of Sciences* **114,** 9797–9802 (Sept. 2017).

82. Wightman, R. *PyTorch Image Models* https://github.com/rwightman/pytorch-image-models. 2019.

83. Chen, J. *et al.* Three-Dimensional Residual Channel Attention Networks Denoise and Sharpen Fluorescence Microscopy Image Volumes. *Nature Methods* **18,** 678–687 (June 2021).

84. Chen, J. *3D residual channel attention networks denoise and sharpen fluorescence microscopy image volumes* (June 2021). https://doi.org/10.5281/zenodo.4651921.

85. Qiao, C. *et al.* Evaluation and Development of Deep Neural Networks for Image Super-Resolution in Optical Microscopy. *Nature Methods* **18,** 194–202 (Feb. 2021).

86. Qiao, C. & Li, D. *BioSR: a biological image dataset for super-resolution microscopy* (Nov. 2020). https://figshare.com/articles/dataset/BioSR/13264793.

87. Spahn, C. *et al.* DeepBacs for Multi-Task Bacterial Image Analysis Using Open-Source Deep Learning Approaches. *Communications Biology* **5,** 1–18 (July 2022).

88. Pereira, P. M. & Pinho, M. *DeepBacs – S. aureus SIM prediction dataset and CARE model* (Zenodo, Oct. 2021). https://doi.org/10.5281/zenodo.5551141.

89. Pereira, P. M. & Pinho, M. *DeepBacs – E. coli SIM prediction dataset and CARE model* (Zenodo, Oct. 2021). https://doi.org/10.5281/zenodo.5551153.

90. Li, X. *et al.* Three-Dimensional Structured Illumination Microscopy with Enhanced Axial Resolution. *Nature Biotechnology* **41,** 1307–1319 (Sept. 2023).

91. Li, X. *Three-dimensional structured illumination microscopy with enhanced axial resolution* version Version 1. July 2022. https://doi.org/10.5281/zenodo.6727773.

92. Weisbart, E. *et al.* Cell Painting Gallery: An Open Resource for Image-Based Profiling. *Nature Methods* **21,** 1775–1777 (Oct. 2024).

93. De Lange, E. M., Mol, F. N., van der Klei, I. J. & Vlijm, R. *STED super-resolution microscopy unveils the dynamics of Atg30 on yeast Pex3-labeled peroxisomes* version V2 (DataverseNL, 2024). https://doi.org/10.34894/FNUHXH.

94. Burgers, T., Mol, F. & Vlijm, R. *Dataset of ESCRT-III polymer rings (CdvB, CdvB1 and CdvB2) in wildtype archaea (Sacidocaldarius-DSM639) imaged by (dyMIN) STED nanoscopy.* version V1 (DataverseNL, 2022). https://doi.org/10.34894/Q6BYSQ.

95. Danzl, J. G. *Research data for the publication "Dense 4D nanoscale reconstruction of living brain tissue"* (Institute of Science and Technology Austria, 2023). https://doi.org/10.15479/AT:ISTA:12817.

96. Deschênes, A., Santiague, J.-G. S. & Lavoie-Cardinal, F. *Confocal- and STED-FLIM images of neuronal proteins* (Zenodo, May 2025). https://doi.org/10.5281/zenodo.15438495.

97. Descloux, A., Grußmayer, K. S. & Radenovic, A. Parameter-Free Image Resolution Estimation Based on Decorrelation Analysis. *Nature Methods* **16,** 918–924. (2021) (Sept. 2019).

98. Loshchilov, I. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

99. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

100. Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters* **27,** 861–874 (2006).

101. Paszke, A. *et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library* in *Advances in Neural Information Processing Systems* (eds Wallach, H. *et al.*) **32** (Curran Associates, Inc., 2019). https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

102. Nault, F. & De Koninck, P. Dissociated Hippocampal Cultures. *Protocols for Neural Cell Culture,* 137–159. (2021) (2009).

103. Van der Walt, S. *et al.* Scikit-Image: Image Processing in Python. *PeerJ* **2,** e453 (June 2014).

104. Good, P. I. *Resampling Methods* 3rd ed. ISBN: 978-0-8176-4444-4 (Birkhäuser Basel, 2006).