

METHODS

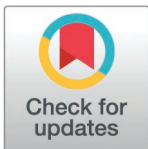
SynAPSeg: A novel dataset and image analysis framework for deep learning-based synapse detection and quantification

Pascal Schamber¹, Sahana Darbhamulla¹, Molly Boyer¹, Madison Pelletier¹, Helene Hartman¹, Olivia Friedman¹, Shiyu Zhang¹, Allison Blais¹, Seyun Oh¹, Haining Zhong², Alexei M. Bygrave^{1*}

¹ Department of Neuroscience, Tufts University, Boston, Massachusetts, United States of America,

² Vollum Institute, Oregon Health and Science University, Portland, Oregon, United States of America

* alexei.bygrave@tufts.edu



OPEN ACCESS

Citation: Schamber P, Darbhamulla S, Boyer M, Pelletier M, Hartman H, Friedman O, et al. (2026) SynAPSeg: A novel dataset and image analysis framework for deep learning-based synapse detection and quantification. PLoS Comput Biol 22(7): e1014571. <https://doi.org/10.1371/journal.pcbi.1014571>

Editor: William Redman, Johns Hopkins University, UNITED STATES OF AMERICA

Received: March 17, 2026

Accepted: July 13, 2026

Published: July 29, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pcbi.1014571>

Copyright: © 2026 Schamber et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution,

Abstract

Synapses are the fundamental units of neural computation, yet quantifying their organization across circuit-level scales remains a critical bottleneck in neuroscience. While advances in fluorescent labeling and imaging can generate vast datasets, analysis is often the limiting factor. Several deep learning-based tools have been proposed to ameliorate these issues. However, existing applications primarily focus on dendritic spines and lack robust solutions for segmenting synaptic puncta in dense tissue preparations. To address this, we introduce SynAPSeg, which encompasses an open-source framework for deep learning-based analysis and, to the best of our knowledge, the first large-scale, publicly available instance segmentation dataset specifically curated for synaptic puncta. We use this dataset to train deep learning models that reach the performance of human experts across a unique benchmark dataset. SynAPSeg integrates these models into an interactive interface, with support for multi-dimensional data, enabling fully automated segmentation and quantification pipelines alongside an annotation module for refinement and validation. We demonstrate the framework's scalability by performing the first comprehensive mapping of nearly 4 million excitatory postsynaptic PSD95 puncta within inhibitory interneurons across the dorsal hippocampus, revealing regional differences in synapse properties. Finally, we show SynAPSeg's utility for 3D quantification by applying these models to study aging-associated synaptic changes in CA1 parvalbumin (PV)-positive inhibitory neurons. Through this approach, we uncover a reduction in PSD95 density along PV dendrites in the aged CA1, indicating reduced glutamatergic recruitment of PV neurons which could contribute to age-related cognitive decline. Collectively, these results demonstrate that SynAPSeg provides a scalable solution for comprehensively studying synaptic architecture in health and disease.

and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The training datasets, benchmark datasets, and the custom trained models have deposited at the following DOI: <https://doi.org/10.5281/zenodo.18988899>. All code is open-source and available at our GitHub repository <https://github.com/pascalschamber/SynAPSeg>.

Funding: This work was supported by funding from the National Institutes of Health (R00MH124920 to AMB, RF1MH120119 and RF1MH130784 to HZ). <https://www.nih.gov>. The funders played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

The connections between brain cells, known as synapses, are the fundamental units of learning and memory. Synapses are often densely packed along dendrites making them difficult to accurately segment with automated image analysis methods. The current gold standard is manual synapse identification, but this is painstakingly slow and highly subjective. While artificial intelligence has shown promise in analyzing biological images, there has been a lack of specialized training data to teach these programs to identify synapses. To address this, we created SynAPSeg, an accessible software platform that includes a large-scale, publicly available collection of expert-annotated synapse images. We used this dataset to train deep learning models to accurately identify synapses in a fraction of the time of human experts—across a range of imaging conditions and sample preparations. In proof-of-principle experiments, we use SynAPSeg to map millions of synapses across the hippocampus to better understand how excitatory input onto inhibitory interneurons changes with aging. Our results indicated a significant decrease in the number of excitatory synapses within parvalbumin-positive inhibitory interneurons, which could be relevant to age-related cognitive decline. In summary, SynAPSeg is an open-source tool that provides a fast, reliable method to automatically segment synapses in fluorescent imaging datasets.

Introduction

Synapses are fundamental units of computation in the brain; their strength and connectivity define the functional architecture of neural circuits and underlying plasticity. Quantifying synaptic changes at scale is essential for bridging molecular and circuit-level neuroscience yet it remains a significant technical challenge. While electron microscopy remains the gold standard for studying ultrastructure, laser scanning microscopy (LSM) offers scalability, and the capacity for multiplexed imaging of synaptic proteins across broad neural circuits. LSM also benefits from recent innovations in genetic mouse models and the development of intrabodies, which enable the fluorescent labeling of endogenous synaptic proteins with enhanced signal-to-noise ratios (SNR) and cell-type specificity compared to traditional immunohistochemistry (IHC) [1–4]. Collectively, this progress shifts the bottleneck in achieving scale from image acquisition to quantification.

While manual annotation is widely accepted as the gold standard for quantifying synapses, its prohibitive time requirements make it impractical for large datasets. In addition, recent studies have shown high inter-rater and intra-rater variability even among experts, challenging the idea that manual annotations reliably reflect an absolute ground truth [4–6]. This ambiguity can be partially attributed to the small size of synapses; as nanoscale structures they often approach the resolution limits of LSM making precise delineation of object boundaries challenging

[7,8]. Consequently, annotations from multiple experts are frequently pooled to generate a consensus. Though this represents a highly rigorous method for establishing ground truth when there is uncertainty among experts, it further exacerbates resource demands [6,8–10]. Automated thresholding-based approaches have been employed to mitigate subjectivity. However, despite the advances made by these methods, they often provide fragile solutions that require parameters to be manually tuned per-image to accommodate varying SNR and illumination levels [11–14], effectively reintroducing the bottleneck and reproducibility issues they were meant to combat. Conventional techniques also typically struggle to resolve object boundaries when signal is dense, especially in volumetric (3D) contexts, as they depend on heuristic-based methods like flood-filling or watershed segmentation to delineate individual objects that are prone to merging errors, leading to under-segmentation, and compromise the ability to analyze the features of individual synaptic structures [7,11,14,15].

Deep learning-based methods have been applied to address similar fluorescent imaging challenges in other areas of biology [16–20], and recently, to dendritic spine segmentation. DeepD3 for example, can accurately segment fluorescent cell fill signal to detect dendritic spines in challenging in vivo domains [6]. Others have shown incorporating image restoration techniques can further improve performance [8,21]. While these approaches work well for sparse dendritic spines, the underlying U-Net-like architecture performs semantic segmentation and does not resolve individual objects. Instead, they must rely on conventional heuristic-based methods to separate touching objects and thus face similar challenges on dense labels. In addition, the punctate signals produced by fluorescently labeled synaptic proteins are morphologically distinct from cell filled spines. This unique detection challenge potentially limits the analysis of various synaptic puncta including somatic synapses, most inhibitory synapses, as well as excitatory inputs onto inhibitory neurons which typically form shaft synapses along dendrites rather than dendritic spines [22].

The largest challenge in developing robust deep learning-based approaches for specialized domains remains a lack of publicly available training data [23,24]. It has been repeatedly shown that deep learning models struggle with data far outside the training distribution [6,8], therefore without dedicated training datasets synaptic puncta segmentation will remain limited. To the best of our knowledge, the DeepD3 dataset is the only significant collection of manually annotated training data in this domain. While we generally lack specialized end-to-end deep learning tools capable of segmenting individual puncta, generalist architectures such as StarDist [25–27], Cellpose [28–30], and MASK R-CNN [31] have been widely applied for instance segmentation across biological domains as well as in dendritic spine segmentation [32]. However, the coding expertise required to utilize and properly implement these tools can create a technical barrier, especially for 3D data, as popular image analysis tools such as ImageJ lack full support even for popular techniques like StarDist.

To facilitate deep-learning-based synaptic quantification, we introduce SynAPSeg, encompassing both open-source datasets and a framework for image analysis. The datasets span a diverse range of labeling modalities to promote generalizability and represent, to our knowledge, the first large-scale, publicly available instance segmentation dataset for synaptic puncta. We use this dataset to train models on established deep learning architectures and evaluate the performance of several recent algorithms for synapse segmentation—finding that custom-trained StarDist architectures show the highest performance on 2D and 3D instance segmentation tasks. We benchmark these models using a novel multi-rater dataset and demonstrate that they reach the level of human experts. These models are made accessible through the SynAPSeg framework, providing a user-friendly interface that combines segmentation, validation, and quantification into a single workflow. We demonstrate the scalability of this platform by applying SynAPSeg to generate the first comprehensive mapping of excitatory postsynaptic densities on inhibitory interneurons across the dorsal hippocampus through imaging of PSD95. Lastly, we demonstrate the utility of SynAPSeg for robust 3D quantification, replicating findings that PV intensity correlates with the number of excitatory synapses, as well as uncovering an age-related reduction in PSD95 puncta density along PV dendrites.

Results

Dataset creation and method valuation

In contrast to the extensive development of deep learning architectures for synapse detection [6,8,10,23,33,34], the field generally lacks publicly available data to train such models. To address this gap, we generated diverse training datasets containing pixel-precise manual annotations for over 4300 2D and 1200 3D putative synaptic puncta (Fig 1A). To provide broad utility, the datasets encompass different resolutions, sample preparations, synaptic markers, labeling methods, SNR, and morphologies (detailed in S1 Table). Given significant progress in spine detection, one focus of this dataset was improving the segmentation of *aspiny* postsynaptic densities, commonly referred to as “shaft synapses”. These punctate structures include GABAergic synapses as well as glutamatergic synapses received by inhibitory interneurons. Their dendritic localization and dense spatial clustering can pose detection challenges that spines do not face. By including puncta from both spine and aspiny synapses, as well as presynaptic markers, our datasets capture a wide range of morphologies and expand the potential for comprehensive synapse analysis. Throughout, we use the term synaptic puncta to refer generally to fluorescently labeled clusters of synaptic proteins whether they are presynaptic, postsynaptic, and either aspiny or located at dendritic spines. To generate ground truth labels for training the models, four human experts manually annotated synaptic puncta by assigning unique pixel value identifiers to each instance. Each expert labeled different images to mitigate systemic bias. Labeling was performed either within defined regions of interest (ROIs) or on the whole field of view (S1A Fig).

We then used this dataset to train 2D and 3D versions of the StarDist and Cellpose-SAM models. As they have fundamental differences in their implementations, the specific parameter values were individually tuned to elicit the best performance (see **Methods**). Models were trained for a sufficient number of epochs until loss plateaued while using a learning rate that decayed as validation loss decreased (S1B - S1E Fig). We noted a persistent generalization gap in the 3D model training curves, characteristic of training on limited data. We were able to force loss convergence by using smaller training patches to increase the number of examples, but this resulted in lower segmentation accuracy on our primary benchmark dataset. Consequently, we prioritized empirical performance for our target task by utilizing the more accurate model presented here, while making the alternative regularized weights publicly available. We compared these custom trained models to pre-trained versions as well as a pre-trained deepD3 model which is designed to detect synaptic spines. To establish a baseline of performance we compared these deep learning approaches to conventional image analysis methods: thresholding followed by watershed segmentation. Since thresholding-based methods are sensitive to the parameters used, we optimized the parameters (thresholding algorithm, gaussian blur, distance threshold) via grid search for each image using the ground truth labels as targets.

Each method was evaluated in 2D and 3D contexts, using the validation fold of the respective dataset. Performance was evaluated using intersection over union (IoU) scores which measures the spatial overlap of the model's predicted labels and the ground truth labels. We used a standard IoU threshold of 0.5 to classify detections as true positives, false positives, and false negatives and calculate recall, precision, and F1 scores. High recall indicates the model misses very few actual synapses. High precision indicates the model rarely mistakes background noise or artifacts for synapses (low false positives). F1 is the harmonic mean of precision and recall, providing a measure of overall performance.

For the 2D dataset, the custom Stardist and CellPose models similarly had the highest precision, recall, and F1 scores (Fig 1B and 1C). For the 3D dataset, the StarDist model performed significantly best on all metrics (Fig 1D and 1E, see S2A and S2B Fig for example segmentations). A known limitation of the IoU metric is its disproportionate penalization of under-segmentation relative to over-segmentation, an effect that is particularly pronounced for small structures such as synapses. We therefore evaluated segmentation performance using shape- and localization-based metrics, including Hausdorff distance and centroid distance, which provide more balanced sensitivity [35,36]. The trends remained the same as observed with the IoU scores (S2C and S2D Fig). Since the custom StarDist model performed best across both 2D and 3D datasets we used these models for subsequent analyses and testing.

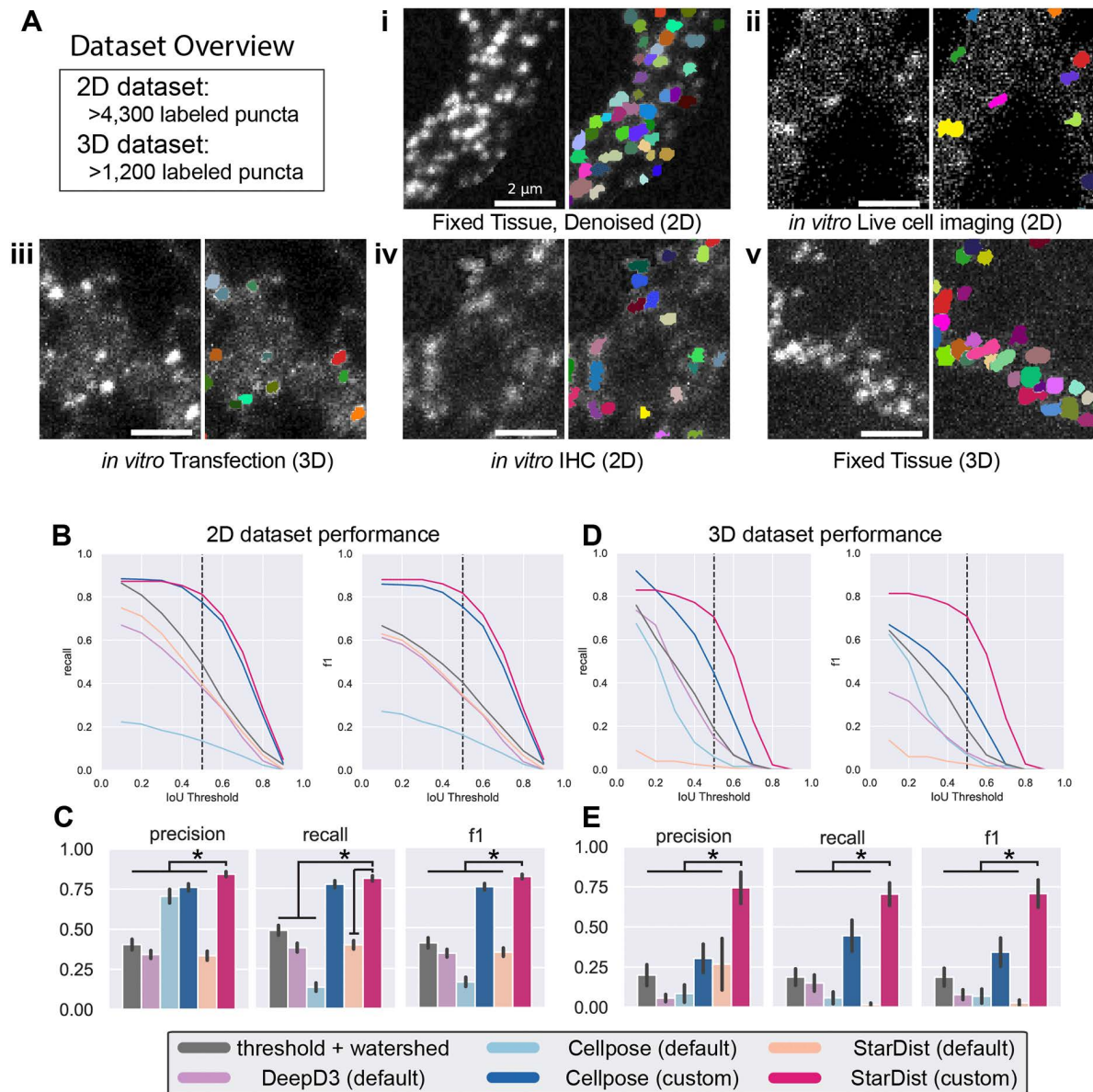


Fig 1. Models trained on the SynAPSeg dataset outperform pre-trained models and classical instance segmentation methods. A) Representative images from training dataset showing confocal maximum intensity z-projections (left half of each panel) and manually annotated instance segmentation labels (right half), with different colors reflecting unique synaptic puncta. Scale bars indicate 2 μ m. Panels show: **i**) PSD95-GFP labeled in fixed tissue from vGat-cre;PSD95-GFP mouse, signal was denoised using N2V³⁵; **ii**) putative inhibitory neuron from primary neuron culture expressing a PSD95-GFP FingR intrabody, image was acquired during live cell imaging; **iii**) putative excitatory neuron from fixed neuronal culture expressing PSD95-GFP FingR intrabody; **iv**) neuronal culture (fixed) where bassoon has been labeled through IHC; **v**) PSD95-GFP in tissue from vGat-cre;PSD95-GFP mouse. **iii** and **v** show images that were labeled in 3D. B) Evaluation of 2D segmentation methods on validation fold of the training data, recall (left) and F1 (right) scores were calculated at different intersection over union (IoU) thresholds. C) Bar plots show the mean precision, recall, and F1 scores across validation set at IoU = 0.5. Error bars display SEM. Significant differences between the custom-trained StarDist models and the other models is indicated by asterisk ($p < 0.05$, two-tailed t-test, FDR correction). D) Same as in B but for 3D data. E) Same as in C but for 3D data.

<https://doi.org/10.1371/journal.pcbi.1014571.g001>

Benchmarking model performance

To assess the custom StarDist models relative to human performance, we constructed three novel multi-rater benchmark datasets. These datasets consist of two 2D images and one 3D volume selected to represent visually distinct image contexts from the training distribution ([Fig 2A](#); see **Methods** for further details). Importantly, these images were held out entirely from both the training and validation sets. For each image in the benchmark datasets, five human experts provided independent, pixel-wise instance segmentation annotations.

Overall, the number, average intensity, and average size of model-predicted labels fell well within the range of human annotations ([S3A–S3C Fig](#)). For the 3D benchmark (Dataset 3), we noted that the model's predictions tended to be about 20% larger in size than the median expert-annotated object, though experts showed high variability and average sizes ranging from 65 to 125 voxels. To follow up on this, we performed one-sample t-tests over the benchmark datasets for count, average size, and average intensity of puncta detected by the model, relative to expert's annotations. These tests indicated no significant difference in count or size across the three datasets. However, there was a significant reduction in mean puncta intensity for dataset 3 (see [S2 Table](#)). To more quantitatively evaluate human and model performance, we calculated the IoU score between each expert and the model for each image ([S3D Fig](#)). The mean IoU score for the models was similar to the highest scoring raters across all datasets, indicating high overlap between experts and the model ([Fig 2B](#)). As human annotations of synaptic puncta have been shown to have high inter-rater variability ([S3E Fig](#), and see [\[6\]](#)), we sought to employ evaluation metrics that capture the expert consensus as a more robust performance target. To do this, we performed spatial clustering on the centroids of labeled objects to determine the convergence of expert agreement for individual synaptic puncta ([Fig 2C–2E](#) and see [S3F Fig](#) for pixel-wise segmentation masks from human annotators). This allowed us to quantify the percentage of objects the model identified relative to the number of humans that also identified that same object. Here we define consensus clusters as those identified by the majority of experts (i.e., $n > 2$). Overall, the model identified 95% (148/155) of consensus clusters ([Fig 2F–2H](#)). Only 3.5% (7/195) of clusters were uniquely identified by the model—a lower rate than the percentage of clusters with a single human annotator (8.2%; 16/195). To quantify reliability, we then evaluated human-to-human against human-to-model precision, recall, and F1 scores using consensus labels as ground truth ([Fig 2I–2K](#)). To statistically evaluate the model's performance relative to human annotators we conducted one-sample t-tests over the IoU, precision, recall, and F1 scores over the three benchmark datasets. No significant differences were observed except when comparing the precision and f1 score between the 2D model and the experts in benchmark dataset 1 ([S2 Table](#)). Importantly, while manual annotation of these datasets required upwards of an hour based on annotator self-reporting, the trained model performed complete segmentations in under 2 seconds ([S3 Table](#)). In summary, we show StarDist models trained on our datasets reach human-level performance in a significantly shorter amount of time, highlighting the scalability of this approach.

SynAPSeg enables fully automated deep learning-based synapse detection and quantification

To facilitate use and integration of deep learning-based tools for image analysis, we developed SynAPSeg as an open-source Python framework available across Windows, Mac, and Linux operating systems ([Fig 3](#)). Workflows in SynAPSeg are accessible through a graphical user interface ([S4A Fig](#)) and structured around three stages: segmentation, annotation, and quantification. The *Segmentation Module* provides a flexible interface for applying highly customizable deep learning-based pipelines. We include several different deep-learning models that users can chain together depending on their use case. This includes models for self-supervised denoising, like N2V2 [\[37\]](#), instance segmentation, like StarDist [\[26\]](#) or Cellpose [\[30\]](#), and semantic segmentation using the Segmentation-Models library [\[38\]](#) (e.g., for dendrite segmentation). The *Annotation Module* allows users to manually correct model outputs and define custom ROIs through a Napari [\[39\]](#) interface. In addition, The *Quantification Module* provides flexible plugins for feature extraction, colocalization analysis, and ROI assignment that allow users to quantify the density and morphological features of segmented objects, regardless of what they represent (e.g., synapses, dendrites, soma). To enable integration across existing tools and

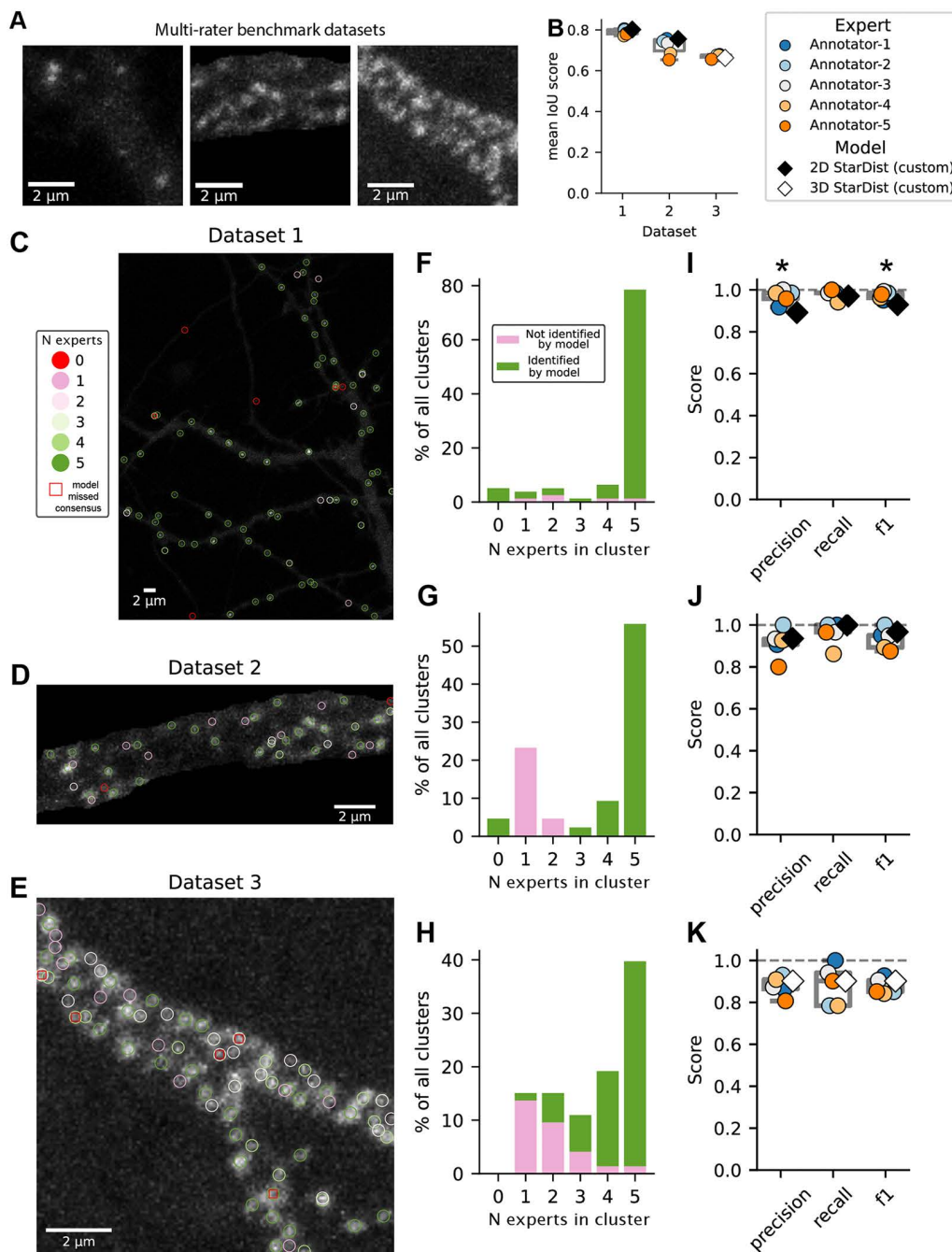


Fig 2. StarDist models trained on the SynAPSeg dataset reach human-level performance. A) Maximum intensity z-projection showing synaptic puncta in ROIs from each benchmark dataset acquired by confocal microscopy. Scale bars indicate 2 μ m. Dataset 1: z-projection showing *in vitro* preparation of primary neuronal culture transfected with plasmid encoding GFP-tagged PSD95-FingR. Dataset 2: z-projection showing *in vitro* preparation where VGluT1 puncta were labeled by immunohistochemistry. Dataset 3: volume acquired from vGat-cre:PSD95-GFP mouse brain tissue sections. B) Mean IoU score between each expert annotator and the StarDist models across each benchmark dataset. C-E) Maximum intensity z-projection showing full images of each dataset. Overlaid circles indicate expert-matched synaptic puncta aligned by spatial clustering. The circles are color-coded to reflect the number of experts in the cluster, except red circles which are detections uniquely identified by the model. Red squares indicate instances where the model did not identify a consensus cluster that the majority ($n > 2$) experts did. F-H) Stacked bar plots showing the percentage of experts that identified each cluster. The performance of the model relative to the human experts is indicated by the colored proportion of each bar. Green indicates the percentage of human-formed clusters that the model also identified, while pink reflects the percentage not identified by the model. The bar heights sum to

100% and reflect the total clusters identified by the experts and the model. **I-K**) Precision, accuracy, and recall scores calculated using expert consensus clusters as true positives, and false positives as clusters identified by a single expert or model. For all boxplots, the box extends from the 1st to 3rd quartiles, the center line reflects the median, and whiskers indicate full range. Model IoU, precision, recall, and F1 scores were compared against the expert annotators over the benchmark datasets using one-sample t-tests (** $p < 0.01$).

<https://doi.org/10.1371/journal.pcbi.1014571.g002>

platforms, SynAPSeg supports a range of image file formats through the Bio-Formats [40] library and utilizes an open file structure for Data Management so that external software can both read and contribute data. This interoperability allows using QuPath [41] and ABBA [42] for brain atlas registration and incorporating the region maps back into the SynAPSeg quantification pipeline. In addition, all steps generate human-readable metadata and configuration files to make analysis reproducible and transparent.

To demonstrate the broad utility of the SynAPSeg platform, we analyzed 3D confocal immunofluorescence data of a primary cultured GABAergic inhibitory interneuron with pre and postsynaptic markers. Bassoon (presynaptic) and PSD95 (postsynaptic) puncta were segmented using the custom 3D StarDist model, and ROIs were manually drawn to include primary and secondary dendrites (S4B and S4C Fig). The glutamatergic synapses were largely aspiny, as expected for a GABAergic inhibitory interneuron. We performed 3D object-based colocalization analysis (see **Methods**) to identify overlapping pre and postsynaptic puncta (S4D Fig). We considered co-localized puncta as putative synapses, and quantified the percentage of PSD95 that was synaptic, as well as the density of synapses within distinct dendritic compartments (S4E Fig). Beyond accurate density measurements, a key advantage of instance segmentation is the ability to extract features from individual puncta. This is particularly valuable given that the abundance of postsynaptic proteins, such as PSD95 and GluA1 measured by the intensity of their fluorescent signal, have been shown to positively correlate with synaptic strength [1,43]. We demonstrate this analysis by extracting features of each colocalized PSD95 puncta (S4F Fig). Collectively, these results demonstrate that SynAPSeg streamlines complex analytical pipelines, replacing labor-intensive manual steps with an automated workflow that enables testing fine-grained hypotheses at scale.

Large-scale analysis of cell type-specific synaptic organization in the hippocampus

While excitatory neurons have glutamatergic synapses that typically are located in dendritic spines that can be morphologically identified, GABAergic inhibitory interneuron (INs) have glutamatergic synapses that are largely located on the dendritic shaft. This makes it more of a challenge to study glutamatergic synapse number and properties in INs. Genetic strategies, where postsynaptic glutamatergic markers are fluorescently labelled, provide a means to visualize aspiny synapses. For example, PSD95 is typically used as a glutamatergic synapse marker, and can be endogenously labelled with a transgenic strategy [1,43,44] or with PSD95-targeting intrabodies [2,3,45,46]. Data from PSD95-GFP CreNABLED mice [44] were included in our training dataset (S1 Table, Fig 1A).

The density of INs is known to vary by subregion, and distinct anatomical projections are known to target INs in specific hippocampal subregions. However, the relative densities and features of IN glutamatergic synapses across the hippocampus is unknown. Previous studies have evaluated how glutamatergic synapses containing PSD95 and SAP102 in all neurons are distributed in the brain, including in the hippocampus [47]. Furthermore, they identified age-dependent changes in the synaptic architecture in aged animals [48,49]. We sought to characterize the organization of glutamatergic synapses in INs within the hippocampus and evaluate if this changes with aging. We used a genetic strategy where endogenous PSD95 was labelled with GFP selectively in INs (vGAT-Cre;PSD95-GFP mice; see **Methods**). We then used the SynAPSeg platform to quantify the properties of IN glutamatergic PSD95 puncta in the hippocampus of early and late adulthood mice. Coronal sections from vGAT-Cre;PSD95-GFP mice aged 3 or 12 months ($n = 5-6$) were stained with DAPI and 2D tile scan images of the dorsal hippocampus were acquired using confocal microscopy (Figs 4A and see S5A for zoomed in examples). We observed large autofluorescent puncta in the pyramidal and granular cell regions in older

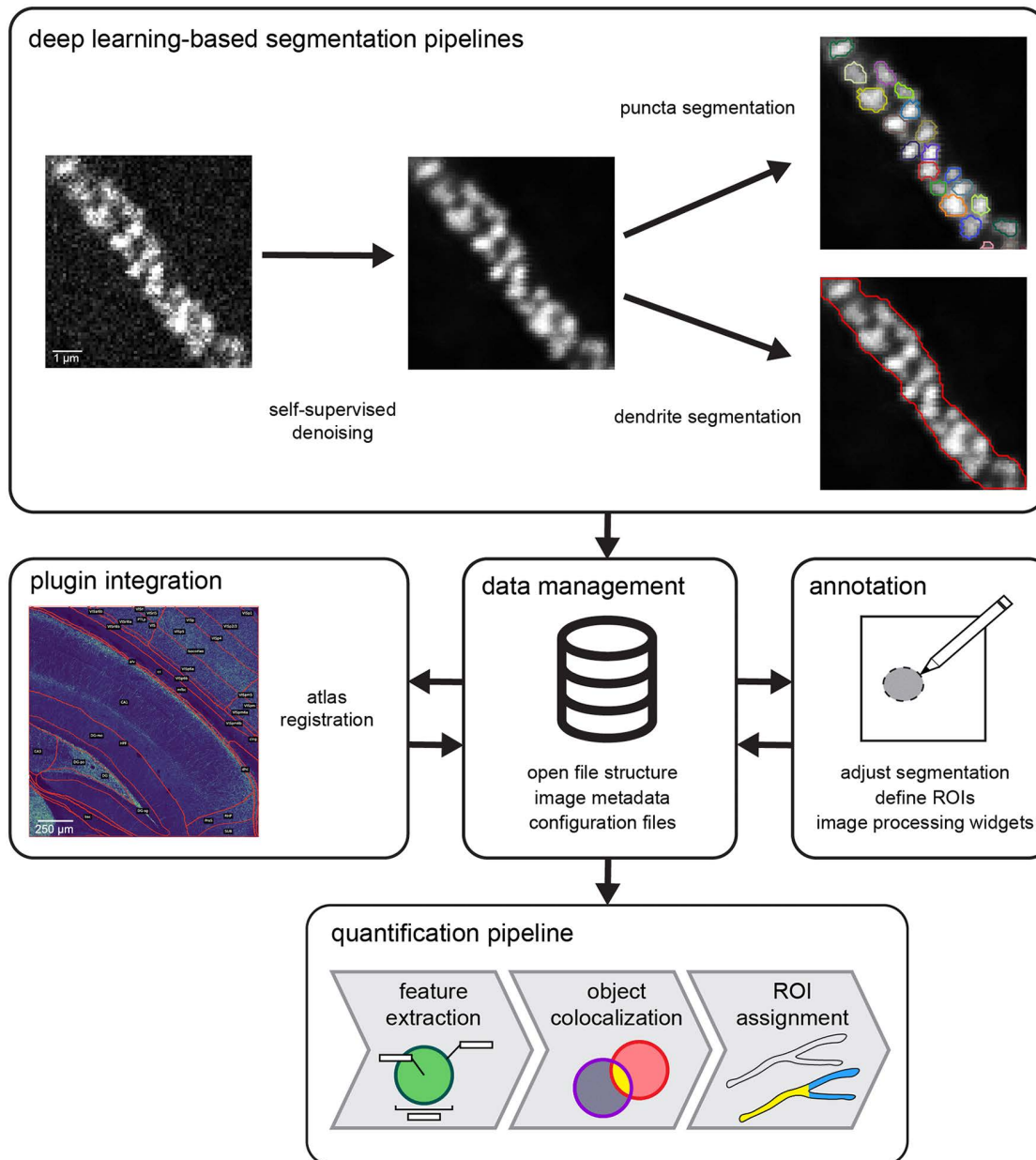


Fig 3. SynAPSeg provides an accessible framework for deep learning-based image analysis. The *Segmentation Module* allows users to compose flexible pipelines using state-of-the-art models for self-supervised denoising (e.g., Noise2Void via the CAREamics library) instance segmentation (e.g., StarDist, Cellpose). Semantic segmentation using custom models is implemented using the Segmentation Models library. *Data Management* is handled through an open file structure where images are stored as TIFF or OME-TIFF files with comprehensive metadata. SynAPSeg generates human-readable configuration files for every pipeline execution to ensure experimental reproducibility. The *Annotation Module* integrates a Napari-based GUI for manual refinement of segmentation results, definition of custom ROIs, and interactive image processing. SynAPSeg is designed for interoperability; QuPath projects can be initialized within the SynAPSeg directory structure to interface with ABBA (Aligning Big Brains & Atlases) for brain atlas registration. The *Quantification Module* provides a composable pipeline for extracting morphological and intensity features, performing multi-channel object colocalization, and assigning detected objects to semantic ROIs. All of these modules are integrated within a unified GUI, facilitating a streamlined end-to-end workflow from processing raw data to quantifying results.

<https://doi.org/10.1371/journal.pcbi.1014571.g003>

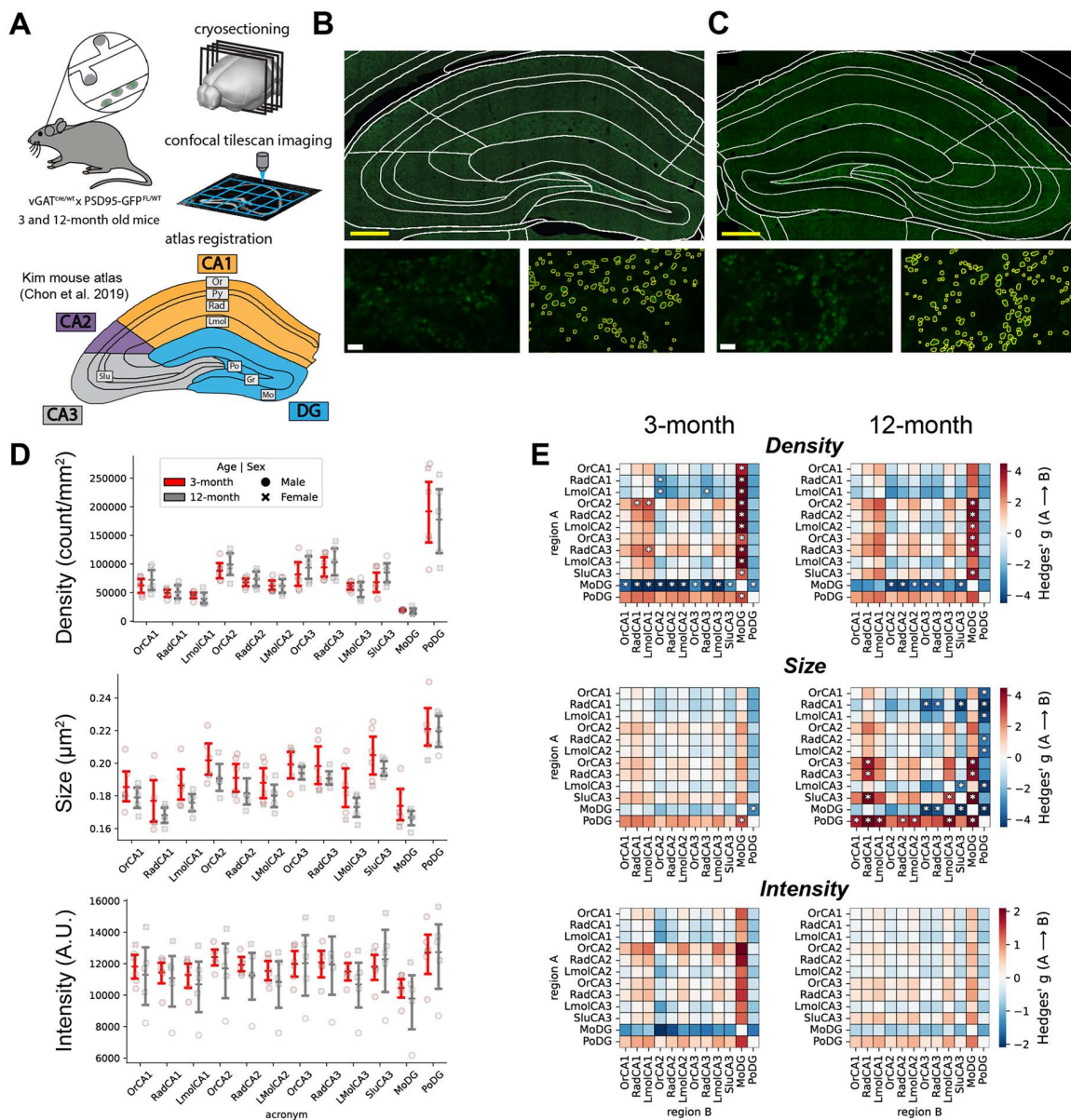


Fig 4. Large-scale 2D analysis of inhibitory neuron glutamatergic inputs in the dorsal hippocampus reveals conserved sub-region-specific density over adulthood **A) Overview of approach.** Coronal sections from vGAT-Cre;PSD95-GFP mice aged 3 and 12 months were imaged using a confocal microscope in tile scan mode to acquire 63x magnification images of the dorsal hippocampus. Images were registered to the Kim Mouse Atlas using ABBA. SynAPSeg was used to then segment PSD95 puncta with the custom 2D StarDist model. Puncta features (density, intensity, size) were extracted and localized to distinct hippocampal subregions using the Quantification Module. **B)** Representative image showing full FOVs acquired at 63X magnification for the 3-month group (top panel; scalebar: 250 μ m), enlarged ROI showing dense glutamatergic innervation of inhibitory interneurons in the polymorphic layer of the dentate gyrus (bottom left panel; scalebar: 2 μ m), and overlaid segmentation outlines (bottom right panel). **C)** same as in B but for the 12-month group. **D)** Quantitative analysis of PSD95 puncta density, mean fluorescence intensity, and puncta size across hippocampal subregions. Each data point represents an individual animal, with markers color-coded by age: 3 months (red) and 12 months (gray). Biological sex is indicated by marker style, with circles representing male subjects and 'X' symbols representing female mice. Data are summarized by error bars reflecting the mean and 95% confidence interval for each age group within each subregion. Two-way ANOVAs revealed no significant age x region interactions. Intensity did however show a significant effect of age alone ($p=0.0003$). **E)** Heatmaps show pair-wise comparisons of puncta features between regions within each age group. Values represent Hedges' g of effect size and were calculated using the regions on the left axis relative to regions along the bottom axis. Unpaired T-tests were performed, and p-values were adjusted using Šidák's multiple comparisons. Asterisks indicate significance based on adjusted p-values (q-values): * $q<0.05$. Hippocampal subregion abbreviations are defined as follows: Field CA1-3 (CA1-CA3), Oriens layer (Or),

Pyramidal layer (Py), Radiatum layer (Rad), Lacunosum moleculare (Lmol), Stratum lucidum (Slu), Dentate gyrus (DG), Granule cell layer (Gr), Molecular layer (Mo), and Polymorph layer (Po).

<https://doi.org/10.1371/journal.pcbi.1014571.g004>

animals which we interpreted as lipofuscin-related autofluorescence, so these regions were excluded from analyses. Synapses were segmented using the custom 2D StarDist model described above. We utilized QuPath and ABBA to manually register images to the Kim Mouse Atlas [50] (Fig 4B and 4C). PSD95-GFP density, area, and intensity were extracted and localized to distinct hippocampal subregions using the object detection and ROI assignment plugins from the Quantification Module in SynAPSeg (S5B Fig). In total, this dataset comprises over 3.8 million puncta localized to 16 distinct hippocampal subregions. Though both sexes were included in the present study, the final cohorts yielded an uneven distribution which failed to reach sufficient statistical power to infer an overall sex-age interaction. We make exploratory statistical comparisons where technically possible: all F vs all M, 3-mo M vs 3-mo F, 3-mo F vs 12-mo F (see S2 Table for summary statistics and supporting information for full model outputs). Overall, this analysis did not reveal clearly significant sex-dependent effects.

We observed discrete patterns of synaptic density, size, and intensity localized to specific hippocampal subregions (Fig 4D). Across these measurements, the poDG showed the largest and most dense puncta of all hippocampal subregions, followed by CA3 subregions. These patterns were largely consistent across age groups. Two-way ANOVA of synaptic density indicated 72% of variance was explained by region with no significant effect from age or age x region interaction. The same was observed for intensity, however area did show an overall significant age effect. However, since the interactions were not significant we did not investigate this further (See S2 Table for full statistical details). To identify larger trends in these properties, we then quantified inter-region differences within age groups using pair-wise multiple comparisons and are expressed as Hedges' *g* of effect size (Fig 4E). In both age groups, we saw some significant differences between inter-regional PSD95-GFP properties. A greater number of regions showed significant changes in puncta size in the 12-month group, while the 3-month group showed slightly more inter-regional density differences. No significant trends in PSD95 intensity were observed. The vast majority of all differences occurred between the moDG and other brain regions, it notably showed the lowest absolute values across all features. Overall, these data indicate patterns of synaptic density were largely conserved with age and demonstrate the utility of SynAPSeg for fully automated segmentation and quantification of synaptic puncta in large datasets.

Analysis of PSD95 puncta on dendrites of CA1 PV Interneurons in 3- and 12-month mice

INs represent just 10–15% of all neurons in the hippocampus but are incredibly diverse with respect to their molecular profiles, electrophysiological properties, and their roles in neuronal circuits [51–54]. In the hippocampus, parvalbumin positive INs (PV-INs) make up roughly 25% of all INs [55], and are important in CA1 for the generation of gamma oscillations which have been shown to play a critical role in information processing and memory encoding [56–59]. Aging is associated with a cognitive decline, and reductions in gamma power [60,61]. While we did not see broad changes in the organization of glutamatergic synapses in all INs with aging (Fig 4), it remains possible that specific subtypes of INs are impacted with differences masked when looking at all INs. Because of their link to circuit and cognitive function, we chose to evaluate glutamatergic synapse organization specifically in PV-INs within the CA1 with aging.

To address this, coronal sections from vGAT-cre;PSD95-GFP mice aged 3- and 12-months were sectioned and PV-INs identified with IHC. Volumetric confocal images were acquired along the radiatum and oriens layers in the CA1 (Fig 5A). PV-IN dendrites were manually drawn using SynAPSeg's annotation module to generate 3D ROIs. The mean intensity of PV immunoreactivity was extracted using the 3D dendrite ROIs, averaged within animals, and compared between age groups (Fig 5B and 5C). The 12-month group showed significantly lower parvalbumin intensity compared to 3-month animals (Fig 5C). The size and length of PV-IN dendrites analyzed were not significantly different between groups (S6A

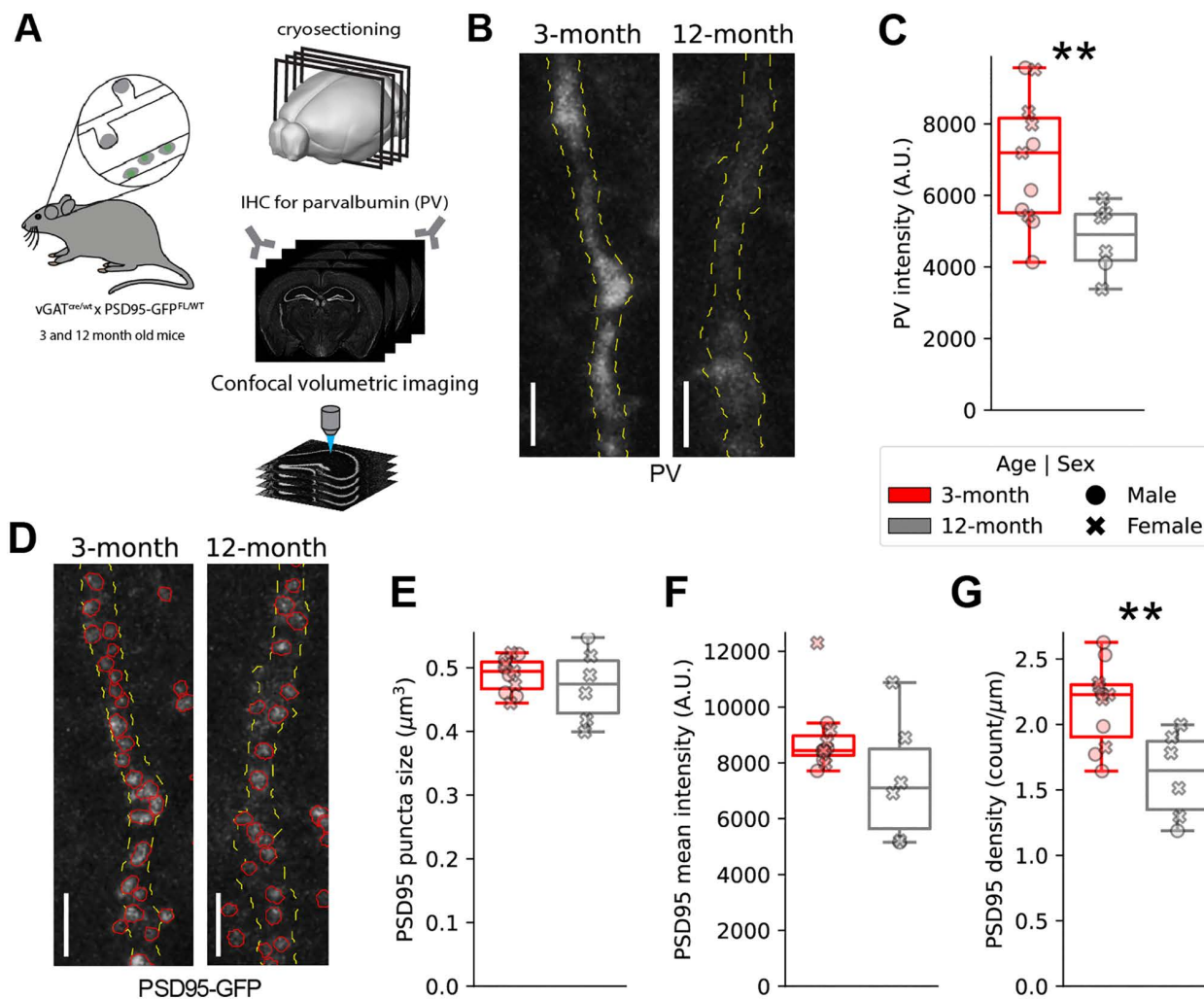


Fig 5. Reduced PV immunoreactivity and PSD95 density along dendrites in CA1 of aged animals. A) Experimental workflow: Coronal sections from 3- and 12-month-old vGAT-cre;PSD95-GFP mice were processed via IHC for PV. PV-positive dendrites were imaged in the radiatum and oriens layers of CA1 using volumetric confocal microscopy. B) Representative z-projections of PV-positive dendrites and manual 3D reconstructions (yellow dashed lines) in 3-month (left) and 12-month (right). Scale bars indicate 2 μm . C) Boxplots compare mean parvalbumin fluorescence intensity within 3D dendritic ROIs. Data points indicate the mean value of several ($N=5-11$; average = 7) dendrites for each animal. PV expression is significantly reduced in 12-month animals (** $p < 0.01$, Welch's T-test). D) Representative images of PSD95-GFP puncta along the same dendrites from B. Red outlines indicate segmented puncta detected by the custom 3D StarDist model. Scale bars: 2 μm . E-G) Quantitative assessment of PSD95-GFP puncta properties. E) puncta size (μm^3), F) mean fluorescence intensity (A.U.), and G) linear density (count/ μm). Linear density is reduced in aged animals (** $p < 0.01$, Welch's T-test), while intensity and size remain stable. For all plots, circles represent males and 'X' represents females; red indicates 3-month and gray indicates 12-month animals. $N=11$ 3-month ($N=5$ females, 6 males) and 6 12-month: (5 females, 1 male).

<https://doi.org/10.1371/journal.pcbi.1014571.g005>

Fig). Next, we used SynAPSeg to quantitatively assess changes in PV-IN glutamatergic synapse number and properties with age, using PSD95-GFP as a proxy for glutamatergic synapses. To do this we used a segmentation pipeline consisting of a N2V2 self-supervised denoising and our custom 3D StarDist model to detect PSD95-GFP puncta (Fig 5D). The segmented PSD95-GFP puncta were assigned to PV dendrites using the 3D quantification module in SynAPSeg. The properties of PSD95 puncta along PV dendrites were then analyzed and features including the density, area, and intensity were extracted. The PSD95-GFP linear density was positively correlated with the intensity of parvalbumin expression in

both age groups ([S6B Fig](#)), consistent with previous reports [62]. While the intensity ([Fig 5E](#)) and size ([Fig 5F](#)) were not significantly different, the linear density (count/ μm) of PSD95-GFP puncta was significantly reduced in the 12-month group ([Fig 5G](#)). Linear regression analysis was performed to isolate the effects of sex and age on PV metrics ([S2 Table](#)). Within the sufficiently powered 3-month cohort (6 males, 5 females), we observed no significant sex differences in density, size, soma intensity, or PV dendrite intensity. Conversely, an age-based comparison restricted to females confirmed our aggregate findings, showing significant changes in both density ($p=0.017$) and PV dendrite intensity ($p=0.01$) at 12 months compared to 3 months.

To address whether the reduction in PSD95 puncta density represents a true loss of synapses or simply altered baseline expression, we incorporated PSD95 intensity, within the dendrite ROI defined by PV expression, as a covariate in a linear regression model to evaluate how PSD95 density is affected by age while controlling for PSD95 intensity. Crucially, the main effect of age on PSD95 density remained significant ([S2 Table](#)).

Previous studies have demonstrated that PV-IN PSD95 puncta exhibit reduced synapse-to-synapse variability compared to those in pyramidal neurons [43]. To investigate the spatial relationships between these clusters, we performed a nearest-neighbor analysis ([S6C Fig](#)). For each PSD95 puncta (“self”) we identified the three closest puncta on the same dendrite (“neighborhood”) and calculated the difference between the ‘self’ intensity and the ‘neighborhood’ average ([S6D Fig](#)). To account for inter-dendritic variance, intensity values were normalized within each dendrite. In both age groups, we observed a significant correlation between self and neighborhood intensities ([S6E Fig](#)). Notably, no correlation was observed when neighbors were randomly shuffled within a dendrite ([S6F Fig](#)). Taken together, these data indicate that high-intensity PSD95 puncta are spatially clustered, and that this relationship remains stable with aging.

Because synapse density onto PV interneurons has been shown to vary along the proximal-distal axis in other brain regions [63], we analyzed whether a spatial gradient existed in our dataset. This was performed on a subset (76% of the original dendrites) where the PV-positive dendrite was visibly attached to a soma. Since our images were centered on the CA1 pyramidal layer ([S7A Fig](#)), this dataset comprises dendrites within the SO and SR subregions. On average, the origin and end of each dendrite was located 33.76 and 62.81 μm from the soma, respectively, and were similar between age groups ([S7B Fig](#)). To analyze density with respect to distance from soma, dendrites were broken up into discrete 2 μm dendritic segments. We observed no correlation between distance from soma and PSD95 puncta density for 3-month animals while a weak but significant positive correlation was observed in the 12-month group ([S7C Fig](#)). The correlations were significantly different. Since 96% of the analyzed dendritic length fell within 100 μm ([S7D Fig](#)), we interpret this dataset to be primarily comprised of proximal/intermediate dendritic segments. However, we analyzed PSD95 density within subsets of the data where distal segments were defined as over 68 μm from the soma (25% of the data), proximal as under 40 μm (36% of the data), and intermediate segments in between (39% of the data). Consistent with what we show for the whole dataset ([Fig 5](#)) we observed a significant reduction in PSD95 density in the 12-month group in the proximal and intermediate dendrites ([S7E Fig](#)).

To investigate the spatial relationship between PV expression and excitatory synapses along individual dendrites, we examined localized PV intensity and PSD95 puncta characteristics in 3-month and 12-month animals ([S8A](#) and [S8B Fig](#)). Correlation analysis using a sliding window to analyze properties at points along the length of individual dendrites demonstrated a potential relationship between PV intensity and PSD95 puncta properties ([S8C](#) and [S8D Fig](#)). To quantify this relationship, dendrites were divided into discrete 2 μm dendritic segments and stratified into ‘high’ and ‘low’ PV expression bins ([S8E Fig](#)). We found that high PV intensity bins also had increased PSD95 density and PSD95 intensity relative to low bins ([S8F](#) and [S8G Fig](#)). Furthermore, repeated measures ANOVA revealed significant main effects of age, with 3-month animals generally exhibiting higher overall PV intensity, PSD95 density, and PSD95 intensity compared to 12-month animals ([S2 Table](#)). Notably, while PSD95 density and intensity co-varied with PV levels and showed a significant age effect, PSD95 puncta size showed only a small but significant effect of bin but not age ([S8H Fig](#)).

Discussion

Machine learning methods have revolutionized biological image segmentation, yet the development of deep learning models for synapse detection remains hindered by the scarcity of high-quality, diverse training data. Generating such datasets is notoriously time-consuming, particularly in domains like synaptic analysis where ground truth subjectivity is high even among experts. Existing resources primarily focus on dendritic spines and lack the instance-level labels necessary to resolve individual synapses in tissue preparations with dense labels [6]. To address these gaps, we present the first comprehensive publicly available instance segmentation dataset of fluorescent synaptic puncta. We evaluated several state-of-the-art architectures and trained custom versions on this dataset. To make these tools available to the broader community we developed SynAPSeg, a Python-based framework which integrates segmentation, annotation, and quantification modules through an accessible interactive interface. All of this is freely available through our GitHub, where we provide pre-trained models as well as the SynAPSeg training and benchmarking datasets.

Deep learning models benefit from exposure to diverse types of data during training, and as such, we generated a comprehensive training dataset containing confocal images spanning multiple imaging conditions, resolutions, synaptic morphologies, and labeling techniques. Though we have tried to be expansive in synaptic markers used, there is a bias for postsynaptic markers, PSD95 in particular. This bias is mitigated through the incorporation of diverse training examples which span experimental conditions, labeling techniques, and sample preparations. Our view is that these factors have a greater impact on the signal characteristics than the marker identity and thus do not impede generalizability. By making these datasets publicly available we aim to empower others to further expand the diversity of open-access synaptic data, train the next generation of models, and broaden the applications of deep learning tools.

These datasets were used to evaluate whether current deep learning tools could be trained to outperform thresholding-based approaches. Custom models significantly outperformed all other methods, with StarDist architectures achieving the highest performance across both 2D and 3D datasets. The lower performance of the custom Cellpose model was unexpected given its more recent development (2025 vs. 2018). This discrepancy likely stems from a combination of merging errors and label boundary overestimation, as reflected by a marked increase in the 3D Hausdorff distance. We suspect the strategy of training 3D Cellpose models, utilizing orthogonal 2D planes to learn 3D representations, is likely ill-suited for puncta segmentation due to the compounding effect of small object size and the inherent anisotropy of this data. The low performance of the pre-trained models in this segmentation task emphasizes that context-specific training is essential to developing robust deep learning models. In a similar vein, the low performance of DeepD3 was not surprising given its training context. However, it does highlight that models trained to detect synaptic spines do not necessarily generalize to the detection of synaptic puncta. The classical thresholding-based approach, even when optimized per image, failed to capture puncta morphology and highlights limitations in conditions of variable SNR. These results underscore that robust detection of dense synaptic puncta requires deep learning models with domain-specific training.

Consistent with previous findings in spine segmentation tasks [4,6,21], our analysis of inter-rater reliability confirmed high variability among human experts. We find that human experts show lower agreement on 3D datasets, highlighting the difficulty of obtaining accurate instance segmentation labels in 3D contexts. Given this subjectivity, we evaluated the custom StarDist models against a human expert consensus rather than diverging ground truth signals provided by individual raters. Across all benchmark datasets, the model's predictions showed strong agreement with the expert consensus, reaching performance levels indistinguishable from human experts. We did, however, observe a subtle, but statistically significant, overall reduction in puncta intensity using our model compared to expert annotators on the 3D dataset (dataset 3). However, we believe this concern is minimized by the fact that feature value disagreements would be applied systematically and therefore should not bias overall group differences. Overall, one of the most striking differences is the time required to generate these labels. While human expert annotations took between 10–60 minutes, depending on the dataset, the model achieved comparable results in under 10 seconds. Together this evidence suggests that properly

trained deep learning approaches offer similar performance to human experts while reducing human bias and enabling large-scale synaptic analysis.

The SynAPSeg framework was developed to make advanced deep-learning models, such as those trained in this study, more accessible to the wider research community. SynAPSeg provides a unified GUI through which segmentation and quantification pipelines can be dynamically composed. We designed this platform to be extensible; the plugin-style architecture enables developers to incorporate new models or quantification techniques with minimal code. The framework's open data management allows external tools to easily interface with the platform, enabling users to integrate custom analytical steps into the existing quantification pipeline. Importantly, SynAPSeg enhances experimental reproducibility by automatically logging configuration parameters for every run. While we have designed this platform with synaptic analysis in mind, it supports a wide range of microscopy image formats, dimensions, and operating systems—making it a general-purpose tool for large-scale image analysis. SynAPSeg represents a significant step forward in providing end-to-end image analysis that reduces labor-intensive manual steps while maintaining the flexibility required for complex biological datasets.

To demonstrate the utility of SynAPSeg, we performed a proof-of-concept large-scale mapping of excitatory postsynaptic densities on INs across the dorsal hippocampus. Unlike pyramidal neurons, many hippocampal INs are largely aspiny (but see [64]) which poses unique detection challenges. Overcoming these challenges is critical to studying their synaptic organization relevant for neuronal circuit function. While we did not observe broad age-dependent differences in synapse organization across the general IN population, we found consistent inter-regional differences in PSD95 density and size. This analysis encompassing millions of PSDs is only addressable through the scalable, automated quantification presented here, and enabled us to present the first large-scale mapping of IN glutamatergic synapses in the hippocampus. One of the most striking trends was the high density of PSD95 puncta in the poDG, which to the best of our knowledge has not been previously reported as most studies have focused on IN inputs in the CA1 region or more general counts of IN soma. In addition, we observed larger puncta size in the CA3 and poDG, likely reflecting Mossy Fiber inputs on INs in these regions [65], as this aligns with known larger terminals received by CA3 pyramidal neurons from DG mossy fibers [66–68]. This synapse type is known to be plastic [69–71] and could represent a synaptic population to focus on in studies of learning related synaptic changes. Given that PSD95 is a reliable marker of glutamatergic synapses and their strength [1,43], these findings provide high-resolution data useful for computational models of hippocampal function.

Due to sample size imbalances across groups, a fully powered age-by-sex interaction model could not be conducted. Exploratory comparisons revealed no significant sex differences within the 3-month cohort, nor significant age differences within the female cohort. The most consistent significant effects were seen in a handful of anatomical regions, most notably the MoDG, PoDG, and RadCA3. Collectively, these data suggest that differences are primarily due to region-specific architecture and baseline individual variation. However, it is important to reiterate that due to unbalanced sample sizes we are unable to conclusively address sex-dependent interactions.

While the total density of IN PSD95 puncta appeared unchanged with age, we speculated that specific subtypes of INs could be impacted. We focused on PV-INs, which make up around 25% of hippocampal INs in the CA1 [55], play an essential role in generating gamma oscillations [56,72], and have been reported to be susceptible to age-related changes [61,73]. We replicated the previously reported decrease in PV immunoreactivity with age [61] and confirmed that PV expression intensity correlates with the density of glutamatergic synapses [62]. Interestingly, our analysis revealed a significant reduction in the density of PSD95 puncta in 12-month animals. This reduced synaptic input from excitatory neurons to PV-INs may indicate a diminished capacity for these cells to regulate circuit timing, potentially contributing to age-related cognitive decline before overt neuronal loss occurs. Our results contrast with a previous report [73], but there were several notable differences including their use of a presynaptic marker to quantify glutamatergic synapses onto PV-INs, as well as their analysis being in the cortex. Future studies utilizing additional age points will be essential to map the exact progression of these synaptic changes. Furthermore, it would be interesting to extend these analyses to other

brain regions, and in neurodegenerative mouse models where PV-IN dysfunction has been reported [58,74,75]. Curiously, we observed hotspots of high and low PV expression along the dendrite which, in an age-independent manner, showed varied PSD95 puncta density and intensity. Our analysis showed there are dendritic hotspots of glutamatergic inputs, in line with our nearest neighbor analysis which indicated synapses close in space are more similar to one another. The functional implication of this spatial relationship and how it relates to PV expression levels remains unexplored.

In general, fluorescent imaging and automatic puncta detection offers a promising method to address questions such as synapse density. However, we would like to highlight general limitations of this method, wherein decreased fluorescent signal could lead to detectability issues. Therefore, when possible, orthogonal verification of synapse changes with electron microscopy or electrophysiology will always be advantageous. We acknowledge that we are using PSD95 as a proxy for glutamatergic synapse number. Previous work has suggested that some glutamatergic synapse might lack PSD95 [47], consequently it is possible that in the old animals there are greater number of PSD95-negative synapses that we are unable to quantify.

In summary, the SynAPSeg framework provides a robust, end-to-end pipeline that improves the rigor and reproducibility of synaptic analysis. While our study utilized a genetic strategy with transgenic mice, the platform is equally compatible with viral vectors, genetically encoded intrabodies, and traditional IHC. By making our custom models and training data publicly available, we provide a reliable tool for the research community. The modular, open-source nature of the framework enables it to be easily extended to other brain regions or synaptic markers, ultimately helping expand the purview of comprehensive synaptic analysis.

Methods

Ethics statement

All experiments were conducted with approval from Tufts Institutional Biosafety Committee (IBC protocol 2022-B33). Animal research protocols were reviewed, approved and consented by Tufts Institutional Animal Care and Use committee (protocol number B2025-50).

SynAPSeg dataset generation

The training dataset is comprised of 41 confocal images containing 2,400 and 1,200 manually annotated synaptic puncta in 2D and 3D, respectively. This data was generated from several different experiments and encompasses different markers as well as acquisition settings, labeling strategies, and sample preparations as detailed in [S1 Table](#). All images were acquired using the same confocal microscope (Zeiss LSM980) and software (Zeiss Zen). The raw intensity images were converted from Carl Zeiss Image (proprietary file format) to the tagged image file format (TIFF) to increase the accessibility of these datasets. To generate the labels for the training dataset, pixel-wise manual annotation was conducted by four human experts. To ensure high fidelity, an iterative review process was employed: initial annotations were generated by a primary expert and subsequently subjected to at least two rounds of quality control review by independent annotators, ensuring that every image was verified a minimum of two times. Individual object instances (synaptic puncta) were labeled by assigning unique integer identifiers to pixels using SynAPSeg's Napari-based Annotation Module. Two distinct annotation strategies were utilized ([S1 Fig](#)): "dense" annotation, where all labels within a full field of view were segmented, and "sparse" annotation, restricted to masked regions of interest, such as along individual dendritic segments.

Training data preparation

2D images were generated by creating maximum intensity projections of confocal z-stacks. To ensure a consistent dynamic range across heterogeneous imaging conditions and minimize pixel saturation, images were normalized between the 1st and 99.9th percentile for 2D data, while 3D data was normalized between the 1st and 99.99th percentiles. For

sparsely annotated images, pixels outside the annotated ROI were replaced with Gaussian noise centered on the mean intensity of the image. To standardize inputs for neural network training, images and their corresponding labels were broken up into non-overlapping patches (**S1**). Based on preliminary optimization indicating that larger fields of view improved model performance, patch dimensions were maximized to utilize the available GPU memory. 2D models were trained on patches of (256, 256) pixels. For 3D models, a patch size of (32, 128, 128) was used. Patches smaller than these dimensions were zero-padded to maintain consistent input shapes. To maximize data utility, individual 2D optical sections extracted from the 3D training volumes were incorporated into the 2D training and validation sets. This greatly increased the number of training labels and the performance of the 2D models. In optimization experiments, this training approach was more effective than training on small z projections generated from contiguous optical sections. Prior to training, label masks were preprocessed to remove artifacts smaller than 5 pixels. Then dilation was applied to close small holes within labeled objects. Finally, the data was filtered to exclude patches containing less than five distinct synaptic labels. The dataset was then partitioned into training and validation folds using an 85/15 split. This split was applied to distribute the different image modalities and maximize the heterogeneity in the training and validation sets. This resulted in 586 2D patches (287 patches came from 3D slices) and 63 3D patches available for training. The same training and validation sets were utilized for all models. The complete dataset is available in both whole-volume and patched formats.

Selection of segmentation methods for comparison

To benchmark performance, segmentation methods were selected based on two primary constraints: the capability to perform instance segmentation (separating touching objects) and the ability to handle both 2D and 3D images. We evaluated DeepD3, a framework for synaptic spine segmentation, as well as two state-of-the-art “generalist” deep learning frameworks, StarDist and Cellpose. While we compare pre-trained and custom models for the generalist architectures, we were only able to evaluate a pre-trained DeepD3 model as training custom models requires paired spine-dendrite segmentations which are not present in our dataset. Our rationale for including DeepD3 was to understand whether architectures designed and trained for spine segmentation could translate to puncta segmentation. For our evaluation of the pretrained models, we used DeepD3’s “DeepD3_32F”, Stardist’s “2D Versatile Fluo” and “3D Demo,” and the “cpsam” weights for Cellpose. To establish a baseline for performance we evaluated these models alongside an automated thresholding and watershed segmentation approach. To provide a rigorous comparison over the diverse imaging modalities, the parameters (threshold, gaussian blur, and distance threshold) for the thresholding-based approach were optimized based on the IoU score for each image and its corresponding ground truth labels. For each image, we first generated binary masks using different automated thresholding algorithms (Otsu, Yen, Li, Isodata, Triangle, Mean filters) from the scikit-image library [76]. The binary mask with the highest IoU score was then used for watershed segmentation to separate connected components. A Euclidean distance transform was applied to the binary mask, followed by Gaussian smoothing (tested sigmas: 0.25, 0.5, 1.0, 1.5, 2.0). Local maxima were then detected using a minimum distance threshold (tested distances: 1, 3, 5, 8 pixels) and utilized as seeds for the watershed transform. Artifacts smaller than five pixels were removed.

Parameters for training the custom Stardist and Cellpose models

The StarDist and Cellpose architectures employ fundamentally distinct algorithms and different training strategies for 3D segmentation. StarDist utilizes a deep neural network to predict object geometries as star-convex polygons by estimating radial distances from a central pixel to the object’s boundary, followed by a non-maximum suppression step to resolve individual instances. For 3D StarDist models, training was performed directly on 3D volumetric patches. In contrast, Cellpose employs a “topological flow” approach where the model predicts a vector field that points toward the spatial center of each object, allowing instances to be defined by tracking these gradients to a common centroid, but limits it to using 2D slices to learn 3D representations. The recent Cellpose-SAM iteration incorporates a pre-trained Vision Transformer to enhance generalization across spatial scales and imaging conditions. For 3D segmentation, Cellpose models were trained

on orthogonal 2D planes (XY, YZ, and XZ) sampled from each training volume, following the guidance outlined in the documentation.

Given the significant differences between these architectures, we selected the parameters used for training each model individually over the course of several optimization experiments. We generally found the default parameters provided by each architecture performed best. The most relevant non-default values are listed below; however, we make available the full training and hyperparameter specifications, along with the model weights (see **Data Availability**). Models were trained using learning rate scheduling that decayed as validation loss stabilized. The number of training epochs was experimentally derived based on when loss plateaued, 120 for StarDist 3D, 100 for Cellpose 3D, and 1000 epochs for the 2D models. We used learning rate of 3E-5 with decay factor of 0.5 for StarDist models. The Cellpose models used a learning rate of 5E-5 and decay of 0.1 and learning rate of 3E-5 and decay rate of 1E-5 for 2D and 3D, respectively. We employed train-time data augmentation strategies, which differed slightly by framework, and tested several different augmentation steps and parameters. We ultimately found that the simplest procedures produced the best results. For StarDist this included random flips, rotations, and Gaussian noise which was implemented using the Python libraries Albumentations [77] for 2D and Volumentations [38] for 3D augmentation. Since external augmentation pipelines are not supported in Cellpose's high-level API, we utilized their default internal augmentations except for scaling, thus the addition of random Gaussian noise was not applied for the Cellpose models. Models were trained using a Nvidia RTX 4090 graphics card on Windows in Conda environment running python 3.10.16, stardist 0.8.4, TensorFlow 2.10.1, cellpose 4.0.6, and PyTorch 2.9.1 (full environment builds for windows, linux, and mac operating systems can be found on our GitHub repository).

Evaluation of segmentation methods

To quantify the performance of these different algorithms we evaluated segmentation accuracy using standard computer vision metrics, including Intersection over Union (IoU), accuracy, recall, and F1 scores. All methods were evaluated independently on the validation folds of the 2D and 3D datasets. The IoU score was utilized as the primary metric for comparison due to its prevalence as a standard benchmark in the machine learning field, despite recognized limitations regarding small-object segmentation [6]. Predicted synaptic puncta were classified as True Positives (TP), False Positives (FP), or False Negatives (FN) based on the IoU overlap with ground truth masks. We calculated mean recall and F1 scores across a range of IoU thresholds (Fig 1B and 1D). Objects with an IoU score > 0.5 were considered TPs for the purpose of selecting the best-performing model (Fig 1C and 1E). To provide a complementary assessment of boundary precision and localization accuracy, we also calculated the Hausdorff distance and centroid distance (S2C and S2D Fig), which offer more egalitarian penalization of over- and under-segmentation compared to overlap-based metrics [36]. Both metrics were applied on matched pairs. Hausdorff distance was calculated using Scipy's implementation [78]. Centroid distance was calculated as the Euclidean distance between the centroids of the ground truth label and nearest matched predicted label.

Benchmark dataset generation and evaluation

The multi-rater benchmark dataset consists of three distinct images (two 2D images and a 3D volume), selected to represent the diversity of imaging modalities present in the broader dataset (see S1 Table for detailed description of different modalities). Critically, these images were held out from the training set. Each image was manually annotated by five experts using the protocols described above. We chose to only evaluate the custom StarDist models against these multi-rater benchmark datasets, as this architecture yielded the highest performance in the initial validation.

We quantified model performance using two approaches. First, we evaluated pixel wise similarity by calculating the mean IoU for every unique pair of human raters and compared these to the human-model IoU scores to determine if the model fell within the range of human variability. While this addresses pixel level agreement, utilizing a varying target as ground truth is problematic, as others have described. Therefore, we also utilized a distance-based clustering approach to establish a human expert-defined "consensus" that accounts for subjectivity in synaptic puncta identification. This was

implemented following the approach as previously described [6]. Centroids of annotated objects from all experts were extracted and spatially overlapping annotations were identified using DBSCAN to cluster annotations within a proximity threshold of three pixels in all dimensions. If this resulted in a cluster containing multiple annotations from the same expert, we then used K-means to separate it into distinct clusters. Consensus objects were defined as clusters identified by a majority of experts ($n > 2$). We stratified individual annotations into Confusion Matrix Components: TPs were assigned detections that identified a consensus cluster, FNs were assigned when a consensus cluster was missed, and FPs were clusters identified by a lone annotator (whether this was a human expert or the model). These components were then used to calculate classification performance metrics (accuracy, recall, and F1 scores) and jointly quantify model and expert human performance. It is important to emphasize this strategy inflates the performance of human experts as the models have no influence in determining what constitutes consensus. The confusion matrix components were also used for measuring inter-rater reliability, which was assessed through Cohen's Kappa score using the scikit-learn libraries implementation [79].

SynAPSeg design

The SynAPSeg framework is structured around three steps: segmentation, annotation, and quantification. This core workflow is accessible through a graphical user interface that facilitates batch processing by allowing the user to define a set of parameters which will be automatically applied to all images in the project and allows settings to be applied to other projects with minimal adjustments. *Segmentation*. To initiate the pipeline, users generate a configuration file that defines data input paths, pre-processing steps, and the specific deep learning models to be deployed. This process yields an “Example” folder for each source image, which serves as a self-contained data package storing: a copy of the raw image data, the generated segmentation masks, and a metadata file recording the exact segmentation parameters and metadata extracted from the source image. In the initial segmentation stage, raw microscopy files are read using the AICSImageIO [80] and Bio-Formats [40] libraries, ensuring compatibility with a wide range of proprietary and open-standard imaging formats. *Annotation*. The annotation module serves as a data quality control step, providing a Napari window to visualize the raw data and segmentation results. The module incorporates interactive widgets and annotation tools that allow users to refine the segmentation results by applying thresholding or by drawing directly on the image. These tools can also be used to define custom ROIs to constrain or categorize detected objects in subsequent quantification steps. Any adjustments to the segmentation masks or generated ROIs can then be exported and saved within the example folder. *Quantification*. The quantification module allows users to define custom pipelines for feature extraction, object colocalization across image channels, and ROI handling. The pipeline outputs data at two levels of granularity to accommodate different analytical needs. Low-level object data is stored in comprehensive spreadsheets containing extracted features, ROI assignments, and colocalization metrics for individual detected objects and ROIs. A high-level spreadsheet integrates these low-level sources with user-defined grouping variables (e.g., sample ID, treatment group, etc.) to generate a summary of the object level data for each initial source image. This summary is formatted for immediate integration into external graphing software.

While the GUI supports these core features, the underlying library contains several advanced modules that are currently accessible only through standalone Python scripts. These include specialized analysis tools for unsupervised object feature clustering and nearest-neighbor spatial analysis. Although the GUI does not currently support training custom deep-learning models (except for the Careamics denoising suite), we provide comprehensive training scripts via our GitHub repository. We also direct users toward dedicated platforms such as BiaPy [81] and ZeroCostDL4Mic [82] which offer accessible interfaces for custom model development. As a framework, SynAPSeg is designed to be extensible—each segmentation model and quantification stage is implemented using a plugin-style architecture making it simple to incorporate new features. Any externally trained model supported by a SynAPSeg plugin can be integrated into the segmentation pipeline.

Animals used and tissue preparation

All experiments were approved by Tufts IACUC protocol. To express GFP-tagged PSD95 in inhibitory neurons, Psd95-GFP CreN-ABLED mice [1,43,44] were crossed with vGat-Cre mice (Jax #028862). All mice used in the aging experiments (Figs 4 and 5) were heterozygous for both alleles. Genotyping was performed by TransnetYX. Preparation of mouse brain tissue was performed uniformly across dataset generation and aging experiments. Animals were deeply anesthetized using intraperitoneal injection of pentobarbital and transcardially perfused with 35ml of ice-cold 1X phosphate-buffered saline (PBS) followed by fixative solution (4% PFA (Electron Microscopy Solutions), in PBS, pH 7.4). Brains were extracted and post-fixed for 2 hours in fixative solution at 4°C, rinsed in PBS, then put in 30% sucrose at 4°C until sunk. Brains were then transferred to Tissue-Tek O.C.T. Compound (Sakura #4583) and flash frozen in a dry ice-ethanol bath and stored at -80°C. Floating coronal sections were acquired using a Leica CM1900 cryostat. All tissue sections used in aging experiments were sectioned at 25 µm. Varying thickness (20–40 µm) were acquired for generation of the training data. Sections were stored in cryoprotectant solution (1% polyvinyl-pyrrolidone (Millipore Sigma #PVP-40), 30% sucrose, 30% ethylene glycol (Millipore Sigma #102466, 1X PBS) at -20°C.

Immunohistochemistry

Various IF/IHC protocols were utilized to generate the training dataset depending on the antibodies used (see S1 Table). All immunohistochemistry was performed on free-floating sections. All incubation and wash steps were conducted at room temperature under constant agitation unless otherwise specified. For tissue used in Fig 5, sections were first washed three times for 10 minutes in 1X PBS, followed by 2 hours in blocking solution (10% normal goat serum (Vector Labs #S-1000–20), 0.3% Triton X-100, 1X PBS). Tissue was then incubated with rabbit anti-parvalbumin primary antibody (1:1000, Swant #PV27a) diluted in 1X PBS containing 5% normal goat serum and 0.15% Triton X-100 for 24 hours at 4°C. Following primary incubation, sections were washed and incubated with goat anti-rabbit Alexa Fluor 594 secondary antibody (1:1000, ThermoFisher #A-11012) in 1X PBS containing 5% normal goat serum and 0.15% Triton X-100 for 2 hours. After a final wash series, sections were placed in 0.1 M phosphate buffer for mounting onto SuperFrost Plus glass slides (Fisher Scientific #22-037-246). Once excess moisture had evaporated, sections were coverslipped (Millipore Sigma #CLS2980245) using 200 µL of ProLong Glass Antifade Mountant (ThermoFisher #P36984). Sections used in Fig 4 were washed, stained with DAPI (1:5000; ThermoFisher #62248) for 10 minutes, and washed again prior to mounting.

Culture preparation and immunocytochemistry

Timed pregnant Sprague-Dawley rats were purchased (Charles River) and dissected at embryonic day 18. Hippocampi of 4–5 embryos were dissected into ice-cold dissection media (10x HBSS (Gibco), Penicillin-Streptomycin (Gibco), sodium pyruvate (Gibco), 10mM HEPES (Gibco), 30mM Glucose (Sigma) and Milli-Q water) prior to chemical dissociation via papain (Worthington) and subsequent trituration via P1000 pipette. Cells were plated at 100,000/well onto glass coverslips coated with Poly-L-Lysine (Sigma) of a 12-well plate (for immunofluorescence). Neurons were grown in Neurobasal media (Gibco) supplemented with Penicillin-Streptomycin (Gibco), 2mM Glutamax (Gibco), 5% horse serum (Gibco), and 2% B27 (Gibco). A full media change was performed two hours after plating cells. One day after plating cells (DIV1), 70% of media was replaced with serum-free Neurobasal media. Cells were fed once a week and maintained at 37°C with 5% CO₂. To express plasmids, neurons were transfected at DIV12 with Lipofectamine 2000 (ThermoFisher #11668027) mixed with plasmid DNA for 45 minutes followed by a full media change containing 50% media retained prior to transfection and 50% freshly prepared Neurobasal supplemented with B27.

Immunocytochemistry

Prior to fixation, cells were washed once with sterile room temperature PBS. Cells were then fixed with 4% paraformaldehyde and 4% sucrose in PBS for 15 minutes at room temperature then washed three times with PBS. Prior to antibody

labeling, fixed cells were permeabilized with 0.2% Triton-X in PBS for 20 minutes, then washed 3 times with PBS before blocking in 10% Normal Goat Serum and 5% Bovine Serum Albumin (Sigma-Aldrich) in PBS for 1 hour. Primary antibodies were added in the same blocking buffer and left overnight at 4°C. The next day, cells were washed 3 times with PBS before adding secondaries for 1.5 hours at room temperature, in the same buffer as primary and blocking. See [S1 Table](#) for a list of antibodies and concentrations used in this study. Following secondary incubation, cells were washed three times in PBS followed by a single rinse in distilled water prior to mounting on SuperFrost Plus glass slides with ProLong Glass Antifade mounting media. Slides were maintained in the dark at 4°C. For [S4 Fig](#), the following antibodies were used: FluoTag-X2 abberior STAR 635P anti-PSD95 (1:5000; NanoTag Biotechnologies #N3702-Ab635P-L), guinea pig anti-bassoon (1:1000; Synaptic Systems #141318), mouse IgG2A anti-GAD67 (1:500; Millipore #MAB5406), DyLight 405 anti-guinea pig (Jackson Immuno Research #106-475-003), Alexa Fluor 488 anti-mouse IgG2A (1:1000; ThermoFisher #A21131).

Confocal imaging and data analysis using SynAPSeg

All imaging was performed on a Zeiss LSM 980 confocal microscope using a 63× objective and Zen Software (Zeiss). The confocal image used in [S4 Fig](#) was acquired using a pixel spacing of 0.500 x 0.071 x 0.071 μm (Z, Y, X dimensions), sequential laser scans, and excitation wavelengths of 353, 493 and 653 nm. Analysis was performed in SynAPSeg. Bassoon and X2-PSD95 puncta were segmented using the custom StarDist 3D model. A small GAD67-positive segment of primary and secondary dendrite was manually annotated to generate 3D ROIs. For quantification, puncta size, intensity, and density were extracted and assigned to primary or secondary dendrites using the 'Overlap3D' ROI assignment algorithm. Colocalization was performed on the bassoon and PSD95 puncta using an IoU threshold of 0 so that if objects overlapped by one voxel, they would be considered putative synapses. The percentage of colocalized PSD95 was calculated by dividing the number of bassoon-positive PSD95 puncta by the total number of PSD95 puncta. The linear density (puncta/μm) was quantified by dividing the number of PSD95 puncta on each dendrite segment by the dendrite's length.

Imaging and analysis of tissue sections from vGAT-Cre:PSD95-GFP mice was performed blind to age group (3-month or 12-month). For the experiments in [Fig 4](#), each image was acquired over the entire dorsal hippocampus on one hemisphere using Zen's tile scan mode and pixel size of 0.071 μm. A rectangular ROI, consisting of several tiles, was drawn over the imaging region and focus points were marked at the brightest Z position at each corner to acquire a single optical section over the entire imaging region. This was done to accommodate any unevenness in the tissue from mounting. DAPI and endogenous PSD95-GFP signal were acquired using sequential scans, 2x averaging, and excitation wavelengths of 353 and 493 nm. The laser intensity and gain for PSD95-GFP was optimized on control tissue. Images were excluded if there was Z-axis drift. 2–3 images were acquired for each animal. Acquired images ranged in size from 14514 to 46110 pixels along any one dimension, with median dimensions of 20352 x 33230 pixels (Y, X).

For the tile scan experiments in [Fig 4](#), Carl Zeiss Image files were segmented in SynAPSeg using the 2D StarDist model to detect PSD95 puncta. The "output_image_pyramid" setting was checked to save the intensity images as multi-resolution image pyramids for compatibility with ABBA and a QuPath project was created for each image. Registrations were then performed in ABBA using the "kim_mouse_10um" atlas which contains brain region annotations at a finer parcellation level (e.g., oriens layer of CA1) than is represented in the most commonly used Allen CCFv3. For each brain section we performed a manual affine "interactive transform" to roughly position the slice, followed by a manual spline registration to position the finer structures using landmarks such as the granule cell, pyramidal, and slm layers. The atlas registrations were exported as .geojson files using a custom QuPath groovy script (available on our GitHub) and processed in SynAPSeg along with the puncta segmentation masks and intensity image using the ABBA Quantification Module. Each .geojson file stores the atlas regions as collections of polygons that have been deformed to the images. To assign objects to particular brain regions, the centroid of each puncta is used to determine the regions that contain it using an optimized implementation of the ray casting algorithm. The density, size, and intensity of PSD95 puncta in each brain

region was aggregated across sections from the same animal by weighted averaging. Inter-regional differences were calculated within age groups using Hedges' *g* of effect size. Multiple paired T-tests using Sidak's correction were performed to evaluate significance.

To generate the images used in Fig 5, we imaged the hippocampal CA1 region of tissue from vGAT-Cre:PSD95-GFP mice that had been labeled for PV as described above. One confocal z-stack (23 x 2639 x 2639 voxels) with voxel size of 0.460 x 0.085 x 0.085 μm (Z, Y, X) was acquired for each animal. PV (AF-647) and endogenous PSD95 (GFP) signal were acquired using sequential scans, 2x averaging, and excitation wavelengths of 653 and 493 nm.

For the 3D analysis of PSD95 puncta along PV dendrites, a SynAPSeg segmentation pipeline consisting of the following steps was applied to each image: first a Careamics N2V2 denoising model [37] was trained on a randomly cropped image patch (ZYX dim size: 16, 1024, 1024) for 25 epochs using a patch size of (16, 64, 64) and otherwise default parameters (careamics v0.0.11). After training, the model performed inference over the whole volume and the resulting denoised image was processed by the custom StarDist 3D model to segment puncta. A separate N2V2 model was trained for each image to ensure optimal performance across the whole dataset and improve the segmentation of the dimmer PSD95 signal in the older samples. PV-positive primary dendrites were then manually annotated using SynAPSeg's Annotation Module to generate 3D ROI masks. Each segment was demarcated using a unique pixel value. For quantification of PV dendrites, the size, intensity, and dendrite length features were extracted. For PSD95 quantification, puncta were first localized to individual dendrites using the 'Overlap3D' ROI assignment algorithm, then size and intensity features were extracted. The intensity was calculated using the raw source image, and not the denoised image. The linear density (puncta/ μm) was quantified by dividing the number of PSD95 puncta on each dendrite segment by the dendrite's length. The density, size, and intensity of PSD95 puncta was averaged over several dendrites per animal. For the nearest neighbor analysis in S6 Fig, we examined how the intensity of each punctum ("self") compared to its three nearest neighbors ("neighborhood") along each dendrite. The puncta intensity values were z score normalized within each dendrite. The "neighborhood delta" was calculated as averaged neighborhood intensity minus self intensity. For the correlation analysis, we compared the self intensity to the mean intensity of its neighbors and relative to a permuted control where neighbors were randomly drawn from the pool of all puncta on that dendrite.

To investigate spatial relationships between the properties of PSD95 puncta along dendrites, we analyzed a subset of the data where the dendrites were visibly connected to PV-positive soma. Overall, this subset represented 76% of the original data set. Dendritic length was extracted using skeletonization as described above and PSD95 puncta locations were mapped along the dendritic shaft. For analysis of PSD95 density in proximal, intermediate, and distal segments, we used a distance threshold based on the distribution of distances from soma after dividing each dendrite up into 2 μm segments. We set the 25th and 75th percentiles of this distribution as thresholds to classify segments as being proximal or distal, 40 and 68 μm respectively. We then analyzed each dendrite independently, calculating the density in these proximal, intermediate, or distal segments.

For analysis of PV covariation with PSD95 properties, the same subset of data described above was used. To generate continuous spatial profiles of PV intensity, PSD95 puncta density, and PSD95 puncta intensity, these properties were extracted using a 2 μm sliding window along the length of the dendrite. The resulting individual traces were smoothed using a Gaussian kernel (sigma=3.0). For quantitative categorical analysis, data were binned into discrete 2 μm segments based on PV expression levels, categorized as either 'low' or 'high' PV intensity. To ensure uniform sampling across varying dendrite lengths, one high PV bin and one low PV bin were sampled for every 10 μm of dendrite. These binned values were then averaged within each individual dendrite. Statistical comparisons of synaptic parameters (PV intensity, PSD95 density, PSD95 intensity, and PSD95 size) were conducted using repeated measures ANOVAs to assess differences between high and low bins, between age groups (3-month vs. 12-month), and the interaction between bin and age.

Linear mixed-effects models (LMMs) were employed to evaluate the effects of age, sex, and anatomical region on PSD95 synaptic puncta metrics (density, size, and intensity) across hippocampal subregions (Fig 4). Since multiple

anatomical regions were sampled from the same animals, and we observed correlated within-subject baseline measurements (e.g., density's intraclass correlation coefficient was 0.430), individual animal subjects were designated as a random intercept to account for within-subject biological variability. Due to unbalanced sample sizes within the cohorts, the dataset was insufficiently powered to evaluate a full age x sex interaction. Consequently, the LMMs assessed sex, age, and region as additive fixed effects where possible. To investigate sex-dependent effects related to the analysis of PV dendrites (Fig 5), a standard linear regression model was used. In addition to puncta density, size, and intensity, PV dendrite intensity was included as a dependent variable. Similar to the regional analyses, group imbalances precluded a fully powered interaction model. Age and sex were assessed as additive fixed effects."

Software

SynAPSeg is developed as a community tool. All code is open-source and available at our GitHub repository <https://github.com/pascalschamber/SynAPSeg> which also contains documentation and several demos. The code uses several open-source Python libraries, which are cited throughout.

Data availability

The training datasets, benchmark datasets, and the custom trained models have deposited at the following DOI: <https://doi.org/10.5281/zenodo.18988899>. All code is open-source and available at our GitHub repository <https://github.com/pascalschamber/SynAPSeg>.

Statistical analysis

All statistical analysis was performed using Python Scipy and Pingouin [83] libraries. Figures were generated using Python Matplotlib and Seaborn libraries. Custom scripts used to analyze data and generate figures are available on our GitHub. N represents the number of animals used in each experiment, unless otherwise noted. $P < 0.05$ was considered statistically significant and is presented as: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Details of all statistical tests (n numbers, age, and sex of animals) can be found in S2 Table.

Supporting information

S1 Fig. Dataset preparation and model training. A) Two approaches were used to convert annotated images into patches for training the models. Sparse annotations (upper panels) refer to images where only puncta within a region of interest (ROI) were labeled. Each ROI was processed independently, first masking the field of view outside the ROI, before splitting up into non-overlapping patches. All pixels outside of the ROI mask were filled with random noise drawn from a distribution centered on the mean intensity of the patch. For densely annotated images (lower panels), the entire field of view was annotated, and non-overlapping patches were extracted using a regularly spaced grid. Before training, images were processed to fill small holes and patches containing less than 5 labels were discarded. B) Combined training and validation loss (left panel) and IoU (right panel) during training of the 2D StarDist model on the SynAPSeg dataset. C) Same as in B but for the 3D StarDist model. D) Training loss for the 2D Cellpose model. E) Same as in D but for the 3D Cellpose model. Note that the models use different loss functions and values are not directly comparable. Additionally, the Cellpose library's training module does not calculate IoU metrics and thus are not shown. (TIF)

S2 Fig. Extended comparison of instance segmentation techniques. A) Representative images selected from the 2D validation dataset comparing segmentation results of the different methods. Black box denotes the result with the highest F1 score. Scale bars indicate 2 μ m. B) Same as in A but for the 3D dataset. C) Shape/boundary (left) and centroid-based (right) metrics computed over the 2D validation dataset. Lower values indicate higher spatial precision. Hausdorff distance

was calculated by comparing the distance between the sets of foreground pixels for each matched pair of ground truth and predicted instances, defined as the maximum of the shortest distances between any two points in the respective sets. Centroid distance was calculated as the Euclidean distance between the center of mass of the predicted mask and its matched ground truth. Error bars display SEM. D) Same as in C but for the 3D dataset.

(TIF)

S3 Fig. Extended benchmarking of expert annotations and performance of the custom trained StarDist models.

A-C) Boxplots comparing morphological and intensity features of segmented objects identified by experts and the StarDist model. Features include A) total puncta count, B) mean size (in pixels for datasets 1 and 2; voxels for dataset 3), C) and mean fluorescence intensity. The 2D and 3D StarDist models are indicated by filled and unfilled diamonds, respectively. The count, size, and intensity of objects generated by the model were compared against the expert annotators over the benchmark datasets using one-sample t-tests ($p < 0.05$). D) Heatmaps show IoU score computed between each of the five expert annotators and the custom-trained StarDist model across the three benchmark datasets. E) Inter-rater reliability was determined using the experts' consensus clusters as ground truth and is measured by Cohen's Kappa score. Each datapoint reflects the score between each of the expert annotations and the consensus labels. Values range from -1 (complete disagreement) to +1 (perfect agreement), while 0 indicates agreement no better than random chance. F) Images show pixel-wise segmentation masks for each of the human expert annotations that have been converted to binary pixel maps and summed to reflect the consensus.

(TIF)

S4 Fig. Implementation of the SynAPSeg framework for 3D colocalization analysis of pre- and postsynaptic markers.

A) Maximum intensity z-projection of a Gad67 (cyan)-expressing putative inhibitory neuron in neuronal culture where bassoon (green) and PSD95 (magenta) were labeled by immunocytochemistry (left). The Gad67 channel was used to define dendritic ROIs (right). Blue and red outlines indicate manually defined primary and secondary dendritic segments, respectively. B) High-magnification views of segmented bassoon (left) and PSD95 (right) puncta. C) Visualization of colocalized puncta representing putative synapses identified via 3D object-based colocalization in SynAPSeg. D) Barplots (left) show the percentage of PSD95 puncta colocalized with bassoon (relative to the total number of PSD95 puncta) and the density of synapses (right) for primary and secondary dendrites. E) Demonstration of capability to extract features from individual synaptic objects, illustrated by the distribution of colocalized PSD95 puncta size (total number of voxels) across dendritic ROIs.

(TIF)

S5 Fig. PSD95 puncta in the hippocampus. Representative images showing PSD95 puncta (green) and DAPI (blue) in subregions of CA1, CA3, and DG (top to bottom), illustrating qualitative differences in synaptic distribution and density. Scalebar: 50 μm . Hippocampal subregion abbreviations are defined as follows: Field CA1–3 (CA1–CA3), Oriens layer (Or), Pyramidal layer (Py), Radiatum layer (Rad), Lacunosum moleculare (Lmol), Stratum lucidum (Slu), Dentate gyrus (DG), Granule cell layer (Gr), Molecular layer (Mo), and Polymorphic layer (Po).

(TIF)

S6 Fig. Extended morphological and spatial analysis of PV-IN dendrites and PSD95 puncta.

A) Morphological properties of PV-IN dendrites analyzed. Boxplots show the total ROI area (μm^3 , left) and length (μm , right) for 3-month (red) and 12-month (gray) age groups. Data points represent animals. N = 11 3-month (N = 5 females, 6 males) and 6 12-month: (5 females, 1 male). B) Correlation between PV-IN linear PSD95-GFP density and intensity of PV expression in dendritic segments. Individual points represent single dendritic segments; a positive correlation is observed in both age groups (Pearson; $r = 0.32$, $p < 0.01$ 3-month; $r = 0.51$, $p < 0.01$ 12-month). The correlations were not significantly different (Fisher r-to-z; $z = 1.07$, $p = 0.29$). C-F) Synaptic Neighborhood analysis. C) Schematic of the spatial

nearest-neighbor analysis. The average z-scored intensity of the 3 closest synaptic neighbors was used to calculate “neighbor delta” (self-intensity minus average neighbor intensity) for each PSD95 puncta along individual dendrites. For all analyses raw intensity values of PSD95 puncta were z-score normalized within each dendrite independently D) Histogram showing the distribution of neighbor delta values; the y-axis displays the percentage of total puncta within each bin. E) Correlation between self-intensity and the average intensity of the surrounding synaptic neighborhood. A significant positive correlation is maintained across both age groups. (Pearson; $r = 0.70$, $p < 0.001$ 3-month; $r = 0.71$, $p < 0.001$ 12-month). The correlations were not significantly different (Fisher r-to-z; $z = 0.59$, $p = 0.55$) F) No correlation was observed between self-intensity and randomly shuffled neighbors (Pearson; $r = 0.01$, $p = 0.49$, 3-month; $r = -0.03$, $p = 0.24$ 12-month).

(TIF)

S7 Fig. PSD95 puncta density along proximal-distal axis of PV CA1 dendrites. A) Representative maximum intensity z-projection of whole volumes showing PV (magenta) dendrites and PSD95 (green) puncta in 3-month (left) and 12-month (right) animals. Scale bar represents 20 μm . B) The origin and end of dendrites were similar in distance from soma between 3-month (red) and 12-month (gray) age groups. C) Correlation between distance from soma and PSD95 puncta density. Data points represent discrete 2 μm dendritic segments. No correlation was observed for 3-month animals while a weak but significant positive correlation was observed in 12-month group (Pearson; $r = 0.0$, $p = 0.92$ 3-month; $r = 0.25$, $p < 0.001$ 12-month). The correlations were significantly different (Fisher r-to-z; $z = 4.44$, $p < 0.001$). D) Histogram showing the distribution of distance from soma for the dendrites analyzed. E) Bar plots show puncta density in proximal, intermediate, and distal dendritic segments. Significant differences in PSD95 density (Mann-Whitney U test) were observed between age groups in the proximal ($*p = 0.0103$) and intermediate ($**p = 0.0071$) segments. Data points represent per-animal means.

(TIF)

S8 Fig. PSD95 puncta density and intensity co-vary with PV intensity. A) Confocal maximum intensity z-projection of a PV positive dendrite from 3-month animal. White circles indicate the positions of PSD95 puncta. Scale bar indicates 5 μm . B) Same as A but from a 12-month animal. C) Line plots showing PV intensity (magenta), PSD95 puncta density (black), and PSD95 puncta intensity (green) extracted within a 2 μm sliding window along the length of the dendrite in A. Individual traces were smoothed using a gaussian kernel ($\sigma = 3.0$). D) Same as C but for the dendrite in B. E-H) Quantitative analysis of synaptic parameters stratified by PV expression level (low vs. high bins). Data were binned within discrete 2 μm segments, sampled such that one high and one low PV bin were taken for every 10 μm of dendrite, and averaged within dendrites. Mean binned values are compared for E) PV intensity, F) PSD95 density, G) PSD95 intensity, and H) PSD95 size across 3-month (red) and 12-month (grey) dendrites. Statistical comparisons were made between high and low bins within each age group as well as between group comparisons within high or low bins using repeated measure ANOVAs. Line plots represent individual dendrites.

(TIF)

S1 Table. Details of the SynAPSeg training dataset. Table with the details related to sample preparation of imaging data used for training and validation of models.

(DOCX)

S2 Table. Details of statistical tests. Table with statistical details for all tests performed and information about the N for each experiment.

(DOCX)

S3 Table. Model inference times. The StarDist models were used to generate predictions over the benchmark datasets as well as average-sized images from the experimental data presented in Fig 4 (TilesCan) and Fig 5 (Volume).

Image size are provided in pixels (Y, X) for 2D data or voxels (Z, Y, X) for 3D volumes. Inference time reflects number of seconds to generate a segmentation, averaged over ten iterations. Note: The reported times encompass the complete end-to-end pipeline utilized by the SynAPSeg framework, including all requisite pre- and post-processing steps. Therefore, these metrics reflect practical, real-world user speeds rather than the isolated inference time of the base StarDist architecture.

(DOCX)

S1 Data. Supplementary data tables related to Figs 1 and S2.

(XLSX)

S2 Data. Statistical tests related to model performance comparisons in Fig 1.

(XLSX)

S3 Data. Supplementary data tables related to Figs 2 and S3.

(XLSX)

S4 Data. Statistical tests related to Figs 2 and S3.

(XLSX)

S5 Data. Supplementary data tables and statistical tests related to Fig 4.

(XLSX)

S6 Data. Statistical models related to age x region interactions in Fig 4.

(XLSX)

S7 Data. Supplementary ANOVAs related to Fig 4.

(XLSX)

S8 Data. Statistical models related to age x sex x region interactions in Fig 4.

(XLSX)

S9 Data. Supplementary data tables related to Figs 5 and S6.

(XLSX)

S10 Data. Statistical tests for covariation analysis of PSD95 density and intensity in Fig 5.

(XLSX)

S11 Data. Statistical models related to age x sex interactions in Fig 5.

(XLSX)

S12 Data. Supplementary data tables related to S7 Fig.

(XLSX)

S13 Data. Supplementary data tables related to S8 Fig.

(XLSX)

S14 Data. Supplementary ANOVAs related to S8 Fig.

(XLSX)

Acknowledgments

We are grateful for support from the Tufts University CMS staff for help with mouse breeding and husbandry.

Author contributions

Conceptualization: Pascal Schamber, Alexei Bygrave.

Data curation: Pascal Schamber, Sahana Darbhamulla, Molly Boyer, Madison Pelletier, Helene Hartman, Olivia Friedman, Shiyu Zhang, Allison Blais, Seyun Oh, Alexei Bygrave.

Formal analysis: Pascal Schamber.

Funding acquisition: Haining Zhong, Alexei Bygrave.

Investigation: Pascal Schamber, Molly Boyer, Alexei Bygrave.

Methodology: Pascal Schamber, Alexei Bygrave.

Project administration: Alexei Bygrave.

Resources: Haining Zhong, Alexei Bygrave.

Software: Pascal Schamber.

Supervision: Alexei Bygrave.

Validation: Pascal Schamber, Sahana Darbhamulla, Molly Boyer, Madison Pelletier, Helene Hartman, Olivia Friedman, Shiyu Zhang, Allison Blais, Seyun Oh, Alexei Bygrave.

Visualization: Pascal Schamber.

Writing – original draft: Pascal Schamber, Alexei Bygrave.

Writing – review & editing: Pascal Schamber, Sahana Darbhamulla, Molly Boyer, Madison Pelletier, Helene Hartman, Olivia Friedman, Shiyu Zhang, Allison Blais, Seyun Oh, Haining Zhong, Alexei Bygrave.

References

- Fortin DA, Tillo SE, Yang G, Rah J-C, Melander JB, Bai S, et al. Live imaging of endogenous PSD-95 using ENABLED: a conditional strategy to fluorescently label endogenous proteins. *J Neurosci*. 2014;34(50):16698–712. <https://doi.org/10.1523/JNEUROSCI.3888-14.2014> PMID: 25505322
- Gross GG, Junge JA, Mora RJ, Kwon H-B, Olson CA, Takahashi TT, et al. Recombinant probes for visualizing endogenous synaptic proteins in living neurons. *Neuron*. 2013;78(6):971–85. <https://doi.org/10.1016/j.neuron.2013.04.017> PMID: 23791193
- Rimbault C, Breillat C, Compans B, Toulmé E, Vicente FN, Fernandez-Monreal M, et al. Engineering paralog-specific PSD-95 recombinant binders as minimally interfering multimodal probes for advanced imaging techniques. *Elife*. 2024;13:e69620. <https://doi.org/10.7554/eLife.69620> PMID: 38167295
- Graves AR, Roth RH, Tan HL, Zhu Q, Bygrave AM, Lopez-Ortega E, et al. Visualizing synaptic plasticity in vivo by large-scale imaging of endogenous AMPA receptors. *Elife*. 2021;10:e66809. <https://doi.org/10.7554/eLife.66809> PMID: 34658338
- Rodriguez A, Ehlenberger DB, Dickstein DL, Hof PR, Wearne SL. Automated three-dimensional detection and shape classification of dendritic spines from fluorescence microscopy images. *PLoS ONE* 3, e1997 (2008).
- Fernholz MHP, Guggiana Nilo DA, Bonhoeffer T, Kist AM. DeepD3, an open framework for automated quantification of dendritic spines. *PLoS Comput Biol*. 2024;20(2):e1011774. <https://doi.org/10.1371/journal.pcbi.1011774> PMID: 38422112
- Majeed M. Toolkits for detailed and high-throughput interrogation of synapses in *C. elegans*. *eLife*. 2024; 12, RP91775.
- Bernal-Garcia S, Schlotter AP, Pereira DB, Recupero AJ, Polleux F, Hammond LA. A deep learning pipeline for accurate and automated restoration, segmentation, and quantification of dendritic spines. *Cell Rep Methods*. 2025;5(10):101179. <https://doi.org/10.1016/j.crmeth.2025.101179> PMID: 40972567
- Simhal AK, Aguerreberre C, Collman F, Vogelstein JT, Micheva KD, Weinberg RJ, et al. Probabilistic fluorescence-based synapse detection. *PLoS Comput Biol*. 2017;13(4):e1005493. <https://doi.org/10.1371/journal.pcbi.1005493> PMID: 28414801
- Kulikov V, Guo S-M, Stone M, Goodman A, Carpenter A, Bathe M, et al. DoGNet: a deep architecture for synapse detection in multiplexed fluorescence images. *PLoS Comput Biol*. 2019;15(5):e1007012. <https://doi.org/10.1371/journal.pcbi.1007012> PMID: 31083649
- Fantuzzo JA, Mirabella VR, Hamod AH, Hart RP, Zahn JD, Pang ZP. Intellicount: high-throughput quantification of fluorescent synaptic protein puncta by machine learning. *eNeuro*. 2017;4(6):ENEURO.0219-17.2017. <https://doi.org/10.1523/ENEURO.0219-17.2017> PMID: 29218324
- Wang Y, Wang C, Ranefall P, Broussard GJ, Wang Y, Shi G, et al. SynQuant: an automatic tool to quantify synapses from microscopy images. *Bioinformatics*. 2020;36(5):1599–606. <https://doi.org/10.1093/bioinformatics/btz760> PMID: 31596456

13. Savage JT, Ramirez JJ, Risher WC, Wang Y, Irala D, Eroglu C. SynBot is an open-source image analysis software for automated quantification of synapses. *Cell Rep Methods*. 2024;4(9):100861. <https://doi.org/10.1016/j.crmeth.2024.100861> PMID: [39255792](#)
14. Manrique JFM. SynapseJ: an automated, synapse identification macro for ImageJ. *Front. Neural Circuits*. 2021; 15: 731333.
15. Danielson E, Lee SH. SynPAnal: software for rapid quantification of the density and intensity of protein puncta from fluorescence microscopy images of neurons. *PLoS One*. 2014;9(12):e115298. <https://doi.org/10.1371/journal.pone.0115298> PMID: [25531531](#)
16. Attarpour A, Osmann J, Rinaldi A, Qi T, Lal N, Patel S, et al. A deep learning pipeline for three-dimensional brain-wide mapping of local neuronal ensembles in teravoxel light-sheet microscopy. *Nat Methods*. 2025;22(3):600–11. <https://doi.org/10.1038/s41592-024-02583-1> PMID: [39870865](#)
17. Wu H, Souedet N, Jan C, Clouchoux C, Delzescaux T. A general deep learning framework for neuron instance segmentation based on efficient UNet and morphological post-processing. *bioRxiv*.
18. Ma J, Xie R, Ayyadury S, Ge C, Gupta A, Gupta R, et al. The multimodality cell segmentation challenge: toward universal solutions. *Nat Methods*. 2024;21(6):1103–13. <https://doi.org/10.1038/s41592-024-02233-6> PMID: [38532015](#)
19. Chen R, Liu M, Chen W, Wang Y, Meijering E. Deep learning in mesoscale brain image analysis: a review. *Comput Biol Med*. 2023;167:107617. <https://doi.org/10.1016/j.combiomed.2023.107617> PMID: [37918261](#)
20. Kar A, Petit M, Refahi Y, Cerutti G, Godin C, Traas J. Benchmarking of deep learning algorithms for 3D instance segmentation of confocal image datasets. *PLoS Comput Biol*. 2022;18(4):e1009879. <https://doi.org/10.1371/journal.pcbi.1009879> PMID: [35421081](#)
21. Xu YKT, Graves AR, Coste GI, Hugarir RL, Bergles DE, Charles AS, et al. Cross-modality supervised image restoration enables nanoscale tracking of synaptic plasticity in living mice. *Nat Methods*. 2023;20(6):935–44. <https://doi.org/10.1038/s41592-023-01871-6> PMID: [37169928](#)
22. Kawaguchi Y, Karube F, Kubota Y. Dendritic branch typing and spine expression patterns in cortical nonpyramidal cells. *Cereb Cortex*. 2006;16(5):696–711. <https://doi.org/10.1093/cercor/bhj015> PMID: [16107588](#)
23. Vidaurre-Gallart I, Feraud-Espinosa I, Cosmin-Toader N, Talavera-Martinez L, Martin-Abadal M, Benavides-Piccione R, et al. A deep learning-based workflow for dendritic spine segmentation. *Front Neuroanat*. 2022;16:817903. <https://doi.org/10.3389/fnana.2022.817903> PMID: [35370569](#)
24. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. A review of deep learning based methods for medical image multi-organ segmentation. *Phys Med*. 2021;85:107–22. <https://doi.org/10.1016/j.ejmp.2021.05.003> PMID: [33992856](#)
25. Schmidt U, Weigert M, Broaddus C, Myers G. Medical image computing and computer assisted intervention – MICCAI 2018, 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II. *Lect. Notes Comput. Sci.* 2018; 265–73. https://doi.org/10.1007/978-3-030-00934-2_30
26. Weigert M, Schmidt U, Haase R, Sugawara K, Myers G. Star-convex Polyhedra for 3D object detection and segmentation in microscopy. 2020 IEEE Winter Conf. Appl. Comput. Vis. (WACV). 2020; 3655–62.
27. Weigert M, Schmidt U. Nuclei instance segmentation and classification in histopathology images with StarDist. *arXiv*. 2022. <https://doi.org/10.48550/arxiv.2203.02284>
28. Stringer C, Wang T, Michaelos M, Pachitariu M. Cellpose: a generalist algorithm for cellular segmentation. *Nat Methods*. 2021;18(1):100–6. <https://doi.org/10.1038/s41592-020-01018-x> PMID: [33318659](#)
29. Pachitariu M, Stringer C. Cellpose 2.0: how to train your own model. *Nat Methods*. 2022;19(12):1634–41. <https://doi.org/10.1038/s41592-022-01663-4> PMID: [36344832](#)
30. Pachitariu M, Rariden M, Stringer C. Cellpose-SAM: superhuman generalization for cellular segmentation. *bioRxiv*. 2025. <https://doi.org/10.1101/2025.04.28.651001>
31. Chen Z. Automatic detection of fluorescently labeled synapses in volumetric in vivo imaging data. *bioRxiv*. 2025. <https://doi.org/10.1101/2025.01.22.634278>
32. Zhang J, Vaidya R, Pendharkar R, Chung HJ, Selvin PR. Revealing synaptic nanostructure distribution through automatic dendritic spine segmentation and single-molecule localization microscopy. *bioRxiv*. 2025. <https://doi.org/10.1101/2024.07.15.603616>
33. Vogel FW, Alipek S, Eppler J-B, Osuna-Vargas P, Triesch J, Bissen D, et al. Utilizing 2D-region-based CNNs for automatic dendritic spine detection in 3D live cell imaging. *Sci Rep*. 2023;13(1):20497. <https://doi.org/10.1038/s41598-023-47070-3> PMID: [37993550](#)
34. Yan Y, Qiu Z, Zhang Q. Segmentation of synapses in fluorescent images using U-Net++ and gabor-based anisotropic diffusion. In: 2021 IEEE International Conference on Medical Imaging Physics and Engineering (ICMIPE), 2021. 1–5. <https://doi.org/10.1109/icmipe53131.2021.9698901>
35. Krull A, Buchholz T-O, Jug F. Noise2Void - Learning Denoising from Single Noisy Images. *arXiv*. 2018. <https://doi.org/10.48550/arxiv.1811.10980>
36. Aydin OU, Taha AA, Hilbert A, Khalil AA, Galinovic I, Fiebach JB, et al. On the usage of average Hausdorff distance for segmentation performance assessment: hidden error when used for ranking. *Eur Radiol Exp*. 2021;5(1):4. <https://doi.org/10.1186/s41747-020-00200-2> PMID: [33474675](#)
37. Höck E, Buchholz T-O, Brachmann A, Jug F, Freytag A. N2V2 -- fixing noise2void checkerboard artifacts with modified sampling strategies and a tweaked network architecture. *arXiv*. 2022. <https://doi.org/10.48550/arxiv.2211.08512>
38. Solovyyev R, Kalinin AA, Gabruseva T. 3D convolutional neural networks for stalled brain capillary detection. *Comput Biol Med*. 2022;141:105089. <https://doi.org/10.1016/j.combiomed.2021.105089> PMID: [34920160](#)
39. Windhager J, Zanotelli VRT, Schulz D, Meyer L, Daniel M, Bodenmiller B, et al. An end-to-end workflow for multiplexed image processing and analysis. *Nat Protoc*. 2023;18(11):3565–613. <https://doi.org/10.1038/s41596-023-00881-0> PMID: [37816904](#)

40. Linkert M, Rueden CT, Allan C, Burel J-M, Moore W, Patterson A, et al. Metadata matters: access to image data in the real world. *J Cell Biol.* 2010;189(5):777–82. <https://doi.org/10.1083/jcb.201004104> PMID: 20513764
41. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: open source software for digital pathology image analysis. *Sci Rep.* 2017;7(1):16878. <https://doi.org/10.1038/s41598-017-17204-5> PMID: 29203879
42. Chiaruttini N, Castoldi C, Reque LM, Camarena-Delgado C, Dal Bianco B, Gräff J, et al. ABBA+Brain, an integrated suite for whole-brain mapping, reveals brain-wide differences in immediate-early genes induction upon learning. *Cell Rep.* 2025;44(7):115876. <https://doi.org/10.1016/j.celrep.2025.115876> PMID: 40553651
43. Melander JB. Distinct in vivo dynamics of excitatory synapses onto cortical pyramidal neurons and parvalbumin-positive interneurons. *Cell Rep.* 2021; 37:109972.
44. Bygrave AM, Sengupta A, Jackert EP, Ahmed M, Adenuga B, Nelson E, et al. Btd11 supports cell-type-specific synaptic function. *Cell Rep.* 2023;42(6):112591. <https://doi.org/10.1016/j.celrep.2023.112591> PMID: 37261953
45. Mosso MB, Zhu M, Ma X, Park E, Barth AL. Long-lasting, subtype-specific regulation of somatostatin interneurons during sensory learning. *Sci Adv.* 2025;11(33):eadt8956. <https://doi.org/10.1126/sciadv.adt8956> PMID: 40815647
46. Okur Z, Schlauri N, Bitsikas V, Panopoulou M, Ortiz R, Schwaiger M, et al. Control of neuronal excitation–inhibition balance by BMP–SMAD1 signalling. *Nature.* 2024;629(8011):402–9. <https://doi.org/10.1038/s41586-024-07317-z>
47. Zhu F. Architecture of the mouse brain synaptome. *Neuron.* 2018; 99: 781–99.
48. Cizeron M, Qiu Z, Koniaris B, Gokhale R, Komiyama NH, Fransén E, et al. A brainwide atlas of synapses across the mouse life span. *Science.* 2020;369(6501):270–5. <https://doi.org/10.1126/science.aba3163> PMID: 32527927
49. Bulovaite E, Qiu Z, Kratschke M, Zgraj A, Fricker DG, Tuck EJ, et al. A brain atlas of synapse protein lifetime across the mouse lifespan. *Neuron.* 2022;110(24):4057–4073.e8. <https://doi.org/10.1016/j.neuron.2022.09.009> PMID: 36202095
50. Chon U, Vanselow DJ, Cheng KC, Kim Y. Enhanced and unified anatomical labeling for a common mouse brain atlas. *Nat Commun.* 2019;10(1):5067. <https://doi.org/10.1038/s41467-019-13057-w> PMID: 31699990
51. Sun Y, Nguyen AQ, Nguyen JP, Le L, Saur D, Choi J, et al. Cell-type-specific circuit connectivity of hippocampal CA1 revealed through Cre-dependent rabies tracing. *Cell Rep.* 2014;7(1):269–80. <https://doi.org/10.1016/j.celrep.2014.02.030> PMID: 24656815
52. Tyan L, Chamberland S, Magnin E, Camiré O, Francavilla R, David LS, et al. Dendritic inhibition provided by interneuron-specific cells controls the firing rate and timing of the hippocampal feedback inhibitory circuitry. *J Neurosci.* 2014;34(13):4534–47. <https://doi.org/10.1523/JNEUROSCI.3813-13.2014> PMID: 24671999
53. Hu H, Gan J, Jonas P. Interneurons. Fast-spiking, parvalbumin⁺ GABAergic interneurons: from cellular design to microcircuit function. *Science.* 2014;345(6196):1255263. <https://doi.org/10.1126/science.1255263> PMID: 25082707
54. Pelkey KA. Hippocampal GABAergic Inhibitory Interneurons. *Physiol Rev.* 2017; 97:1619–747.
55. Bezaire MJ, Soltesz I. Quantitative assessment of CA1 local circuits: knowledge base for interneuron-pyramidal cell connectivity. *Hippocampus.* 2013;23(9):751–85. <https://doi.org/10.1002/hipo.22141> PMID: 23674373
56. Carlén M, Meletis K, Siegle JH, Cardin JA, Futai K, Vierling-Claassen D, et al. A critical role for NMDA receptors in parvalbumin interneurons for gamma rhythm induction and behavior. *Mol Psychiatry.* 2012;17(5):537–48. <https://doi.org/10.1038/mp.2011.31> PMID: 21468034
57. Davis P, Zaki Y, Maguire J, Reijmers LG. Cellular and oscillatory substrates of fear extinction learning. *Nat Neurosci.* 2017;20(11):1624–33. <https://doi.org/10.1038/nn.4651> PMID: 28967909
58. Barnes SA, Pinto-Duarte A, Kappe A, Zembrzycki A, Metzler A, Mukamel EA, et al. Disruption of mGluR5 in parvalbumin-positive interneurons induces core features of neurodevelopmental disorders. *Mol Psychiatry.* 2015;20(10):1161–72. <https://doi.org/10.1038/mp.2015.113> PMID: 26260494
59. Tziliavaki A, Tukker JJ, Maier N, Poirazi P, Sammons RP, Schmitz D. Hippocampal GABAergic interneurons and memory. *Neuron.* 2023;111(20):3154–75. <https://doi.org/10.1016/j.neuron.2023.06.016> PMID: 37467748
60. McQuail JA, Frazier CJ, Bizon JL. Molecular aspects of age-related cognitive decline: the role of GABA signaling. *Trends Mol Med.* 2015;21(7):450–60. <https://doi.org/10.1016/j.molmed.2015.05.002> PMID: 26070271
61. Ueno H, Takao K, Suemitsu S, Murakami S, Kitamura N, Wani K, et al. Age-dependent and region-specific alteration of parvalbumin neurons and perineuronal nets in the mouse cerebral cortex. *Neurochem Int.* 2018;112:59–70. <https://doi.org/10.1016/j.neuint.2017.11.001> PMID: 29126935
62. Donato F, Rompani SB, Caroni P. Parvalbumin-expressing basket-cell network plasticity induced by experience regulates adult learning. *Nature.* 2013;504(7479):272–6. <https://doi.org/10.1038/nature12866> PMID: 24336286
63. Morabito A, Zerlaut Y, Dhanasobhon D, Berthaux E, Pinho CM, Tziliavaki A, et al. Distinct dendritic integration strategies control dynamics of inhibition in the neocortex. *Neuron.* 2025;113(18):2962–2978.e10. <https://doi.org/10.1016/j.neuron.2025.05.029> PMID: 40592329
64. Kaufhold D, Maristany de Las Casas E, Ocaña-Fernández MDÁ, Cazala A, Yuan M, Kulik A, et al. Spine plasticity of dentate gyrus parvalbumin-positive interneurons is regulated by experience. *Cell Rep.* 2024;43(3):113806. <https://doi.org/10.1016/j.celrep.2024.113806> PMID: 38377001
65. Frotscher M, Soriano E, Misgeld U. Divergence of hippocampal mossy fibers. *Synapse.* 1994;16(2):148–60. <https://doi.org/10.1002/syn.890160208> PMID: 8197576

66. Amaral DG, Dent JA. Development of the mossy fibers of the dentate gyrus: I. A light and electron microscopic study of the mossy fibers and their expansions. *J Comp Neurol*. 1981;195(1):51–86. <https://doi.org/10.1002/cne.901950106> PMID: [7204652](#)
67. Blackstad TW, Kjaerheim A. Special axo-dendritic synapses in the hippocampal cortex: electron and light microscopic studies on the layer of mossy fibers. *J Comp Neurol*. 1961;117:133–59. <https://doi.org/10.1002/cne.901170202> PMID: [13869693](#)
68. Rollenhagen A. Structural determinants of transmission at large hippocampal mossy fiber synapses. *J. Neurosci*. 2007; 27:10434–44.
69. Vyleta NP, Borges-Merjane C, Jonas P. Plasticity-dependent, full detonation at hippocampal mossy fiber-CA3 pyramidal neuron synapses. *Elife*. 2016;5:e17977. <https://doi.org/10.7554/eLife.17977> PMID: [27780032](#)
70. Lee KJ, Queenan BN, Rozeboom AM, Bellmore R, Lim ST, Vicini S, et al. Mossy fiber-CA3 synapses mediate homeostatic plasticity in mature hippocampal neurons. *Neuron*. 2013;77(1):99–114. <https://doi.org/10.1016/j.neuron.2012.10.033> PMID: [23312519](#)
71. Alle H, Jonas P, Geiger JR. PTP and LTP at a hippocampal mossy fiber-interneuron synapse. *Proc Natl Acad Sci U S A*. 2001;98(25):14708–13. <https://doi.org/10.1073/pnas.251610898> PMID: [11734656](#)
72. Hadler MD, Tziliivaki A, Schmitz D, Alle H, Geiger JRP. Gamma oscillation plasticity is mediated via parvalbumin interneurons. *Sci Adv*. 2024;10(5):eadj7427. <https://doi.org/10.1126/sciadv.adj7427> PMID: [38295164](#)
73. Eavri R, Shepherd J, Welsh CA, Flanders GH, Bear MF, Nedivi E. Interneuron simplification and loss of structural plasticity as markers of aging-related functional decline. *J Neurosci*. 2018;38(39):8421–32. <https://doi.org/10.1523/JNEUROSCI.0808-18.2018> PMID: [30108129](#)
74. Hadler MD, Alle H, Geiger JRP. Parvalbumin interneuron cell-to-network plasticity: mechanisms and therapeutic avenues. *Trends Pharmacol Sci*. 2024;45(7):586–601. <https://doi.org/10.1016/j.tips.2024.04.003> PMID: [38763836](#)
75. Ma D, Gu C. Discovering functional interactions among schizophrenia-risk genes by combining behavioral genetics with cell biology. *Neurosci Biobehav Rev*. 2024;167:105897. <https://doi.org/10.1016/j.neubiorev.2024.105897> PMID: [39278606](#)
76. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. *PeerJ*. 2014;2:e453. <https://doi.org/10.7717/peerj.453> PMID: [25024921](#)
77. Buslaev A. Albumentations: fast and flexible image augmentations. *Information*. 2020.
78. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261–72. <https://doi.org/10.1038/s41592-019-0686-2> PMID: [32015543](#)
79. Pedregosa F. Scikit-learn: Machine Learning in Python. *arXiv*. 2012. <https://doi.org/10.48550/arxiv.1201.0490>
80. Brown EM, Toloudis D, Sherman J, Swain-Bowden M, Lambert T. AllenCellModeling/aicsimageio: Image reading, metadata conversion, and image writing for microscopy images in python. 2021. <https://github.com/AllenCellModeling/aicsimageio>
81. Franco-Barranco D, Andrés-San Román JA, Hidalgo-Cenalmor I, Backová L, González-Marfil A, Caporal C, et al. BiaPy: accessible deep learning on bioimages. *Nat Methods*. 2025;22(6):1124–6. <https://doi.org/10.1038/s41592-025-02699-y> PMID: [40301624](#)
82. von Chamier L, Laine RF, Jukkala J, Spahn C, Krentzel D, Nehme E, et al. Democratising deep learning for microscopy with ZeroCostDL4Mic. *Nat Commun*. 2021;12(1):2276. <https://doi.org/10.1038/s41467-021-22518-0> PMID: [33859193](#)
83. Vallat R. Pingouin: statistics in Python. *J. Open Source Softw*. 2018; 3:1026.