DeepFLaSH, a deep learning pipeline for segmentation of fluorescent labels in microscopy images

Dennis Segebarth[a,*], Matthias Griebel[b,*], Alexander Dürr[b], Cora R. von Collenberg[a], Corinna Martin[a], Dominik Fiedler[c], Lucas B. Comeras[d], Anupam Sah[e], Nikolai Stein[b], Rohini Gupta[a], Manju Sasi[a], Maren D. Lange[c], Ramon O. Tasan[d], Nicolas Singewald[e], Hans-Christian Pape[c], Michael Sendtner[a], Christoph M. Flath[b,**], Robert Blum[☆a,f,**]

[a] *Institute of Clinical Neurobiology, University Hospital Würzburg, 97078 Würzburg, Germany*
[b] *Department of Business and Economics, University of Würzburg, 97070 Würzburg, Germany*
[c] *Institute of Physiology I, Westfälische Wilhelms-University, 48149 Münster, Germany*
[d] *Department of Pharmacology, Medical University of Innsbruck, 6020 Innsbruck, Austria*
[e] *Department of Pharmacology and Toxicology, Institute of Pharmacy and Center for Molecular Biosciences Innsbruck, University of Innsbruck, 6020 Innsbruck, Austria*
[f] *Comprehensive Anxiety Center, 97078 Würzburg, Germany*

Abstract

Here we present and evaluate DeepFLaSH, a unique deep learning pipeline to automatize the segmentation of fluorescent labels in microscopy images. The pipeline allows training and validation of label-specific convolutional neural network (CNN) models that can be uploaded to an open-source CNN-model library. As there is no ground truth for fluorescent signal segmentation tasks, we evaluated the CNN with respect to inter-coding reliability. Similarity analysis showed that CNN-predictions highly correlated with segmentations by human experts. DeepFLaSH also allows adaptation of pretrained, label-specific CNN-models from our CNN-model library to new datasets by means of transfer learning. We show consistent model-performance on datasets from three independent laboratories after transfer learning, thus ensuring its objectivity and reproducibility. DeepFLaSH runs as a guided, hassle-free open-source tool on a cloud-based virtual notebook with free access to high computing power and requires no machine learning expertise.

Keywords: fluorescence microscopy, deep learning, image segmentation, convolutional neural network, open source, Jupyter notebook, cFOS, Parvalbumin, hippocampus, amygdala, paraventricular thalamus, infralimbic cortex, fear conditioning, plasticity

Introduction

Fluorescence labeling of cells in brain slices with antibodies is one of the most frequently used methods in neurobiology. Cell labels are routinely monitored by fluorescence microscopy techniques. Recent technical advances provide the possibility to automatically acquire large image datasets, even at high resolution and with high throughput [1, 2, 3]. However, the subsequent analysis of these images is very demanding. Typically, an expert evaluates fluorescent labels and categorizes them based on individual heuristic criteria, such as morphology, size or signal intensity, as background or as region of interest (ROI). This cognitive decision process is subjective and long known to potentially limit both objectivity and reproducibility [4, 5, 6]. Furthermore, manual image segmentation can become extremely time-consuming as the borders of thousands, if not tens of thousands of features need to be outlined. Ultimately, reaching pixel-wise accuracy manually in this task is virtually impossible.

In response to this challenge, computational

☆Lead contact
*These authors contributed equally
**These senior authors contributed equally

tools for semi-automatic or automatic segmentation of fluorescent labels have been deployed [5, 7, 8]. However, tools that use signal-thresholds to segment fluorescent labels depend on a high signal-to-noise ratio to correctly separate ROIs from background noise [9].Automatic segmentation is also hampered by variability in the section- and labeling-quality, often caused by batch-to-batch differences of reagents, or even variations between individual animals. Furthermore, standardization of image acquisition to differences in label qualities is difficult and thus impedes quantitative imaging. Due to all these drawbacks, heuristic analysis performed by an experimenter blinded to the experimental condition remained a gold standard to analyze fluorescent labels [10, 11, 12]. In recent years, deep learning and particularly convolutional neural networks (CNNs) have shown their remarkable capacities in image recognition tasks [13, 14]. Substantial progress has been made with deep neural networks for image feature recognition in biomedical imaging data. Deep learning has already been used for classifying types of skin cancer [15], to identify blinding retinal diseases [16] or to predict fluorescent labels from bright field images [17].). Also, segmentation of image features has been addressed with deep learning approaches and has recently been provided as a cloud-based segmentation tool, called CDeep3M [18]. However, the computations of CNNs in so-called hidden layers are incomprehensible, which makes it important to carefully validate CNNs before they can be used on research datasets. CNN validation is essential for analysis of fluorescent labels in particular, because there is no ultimate ground truth for the segmentation of fluorescent signals that can be used to train or evaluate deep learning networks. Therefore, objective analysis and reproducibility tests are key before CNNs can be used to accurately process research datasets. Here we present a deep learning pipeline, framed as DeepFLaSH (a Deep-learning pipeline for Fluorescent Label Segmentation that learns from Human experts), to create CNN-models for fluorescent signal segmentation. We hypothesized that deep learning can be used to improve objectivity in image segmentation, when CNNs are trained with equal input from multiple independent human experts. Here, we tested DeepFLaSH on brain slices with fluorescent labels of the neuronal plasticity marker cFOS or the calcium-binding protein Parvalbumin [19, 20]. We validated our approach in several steps: (1) We demonstrate that the resulting CNN-models reach expert-like performance on related imaging datasets. (2) We show that both models are suited to extract behavior-related changes pertaining to signal abundance and intensity of both proteins. (3) Ultimately, we show that CNN-models trained by DeepFLaSH can easily be adapted to new datasets acquired from three independent laboratories, while maintaining expert-like performance in image segmentation. The pipeline we designed and provide in this study includes generic pre- and post-processing of the image data, as well as training, evaluation and fine-tuning of label-specific CNNs. It can be run either in local facilities or - with virtually no requirements to both hardware and machine learning expertise in a cloud-based virtual notebook.

Results

To illustrate a strategy for automated segmentation of fluorescent images from histological samples, we used confocal microscopy images of anti-cFOS and/or anti-Parvalbumin labels in the hippocampus, following behavioral training of mice. In absence of a ground truth, we performed similarity analysis to show that segmentation maps based on the CNN-prediction correlate with segmentation maps created by multiple human experts. Finally, we implemented transfer learning to adapt the CNN-model to microscopy images from different laboratories and confirmed that the approach is suited to automatically process fully independent image datasets.

Contextual fear conditioning and image data acquisition

First, we prepared an image dataset that comprises three experimental groups: mice directly taken from their homecage as nave learning controls (HC), mice after retrieval of a previously explored training context as context controls (C-) and mice that underwent Pavlovian conditioning (fear/threat conditioning) in the same training context (C+, Figure 1A). Conditioned mice showed increased freezing behavior after fear acquisition and showed strong freezing responses when re-exposed to the training context 24 h later (Figure1B and 1C). Unconditioned control mice showed significantly lower amount of freezing behavior in the training context (Figure 1B and 1C).

Brain sections were prepared and proteins of interest the neuronal activity-related protein cFOS, the interneuron subpopulation marker Parvalbumin and the neuronal marker NeuN (Fox3) - were labelled by indirect immunofluorescence. NeuN labels served as counterstain for identification of the corresponding brain sub-regions in the hippocampus. Confocal microscopy images (x,y-z) were acquired and stored as maximum intensity projections (Figure 1D). Based on these raw 2D-images, five experts in neurobiology (throughout named as Expert 1 Expert 5) independently created expert-specific segmentation maps of regions-of-interest (ROI) of two fluorescent labels (cFOS and Parvalbumin) according to their individual heuristic criteria (Figure 1D). In case of Parvalbumin, the aim was to label the cell somata and not widely ramified neurites of these local interneurons. We also used a semi-automatic routine (see methods) to create threshold-based segmentation maps (Figure 1D). As indicated in Figure 1D, both, individual experts and threshold-based segmentation, reveal subjective differences in the interpretation of the fluorescent labels. As expected (Shuvaev et al., 2017)), the threshold-based segmentation showed a tendency to prefer high intensity labels (high bit-value), while ignoring weaker close-to-noise signals (Figure 1D).

Strategy to train a CNN for segmentation of fluorescent labels

Supervised deep learning uses a training dataset that consists of input pairs of instance (in our case microscopy images) and the corresponding label (here the segmentation maps) to iteratively optimize the computational weights in the hidden layers to minimize the deviation between predicted output and the ground truth input [14]. However, neither heuristic nor threshold-based segmentation maps provide an absolute ground truth in this particular case of fluorescent labels. To overcome this fundamental obstacle, we used for each input image multiple segmentation maps created by different experts. We then adapted the idea of inter-coder reliability to limit the subjectivity of a given ROI being classified as a relevant fluorescent label by pooling the segmentation data across these experts [21]. Inter-coding results in scaling of segments (or ROIs) such that marking by multiple coders increases the impact during training and thus ensures objectivity of the trained CNN-model. We selected a training dataset of 36 images, consisting of four randomly chosen images of each experimental condition (HC, C- and C+) for every analyzed hippocampal region (DG, CA3 and CA1; 4 x 3 x 3). The five experts served as inter-coders and processed the training dataset independently to create coder-specific segmentation maps. We used the raw images showing the fluorescent label as input for the CNN, and the corresponding coder-specific segmentation maps as the desired output (Figure 2A). However, CNNs, such as the one created for the present study, usually require large training datasets of several thousands of images to avoid an over-fitting of the network to the training data [22]. Given limited training data availability, we apply data augmentation using elastic deformations to the available training images. This allows the CNN to acquire robustness to such deformations, without the need to see these transformations in the annotated image corpus. This is particularly important in biomedical segmentation, since deformation is the most common variation in tissue and realistic deformations can be simulated efficiently. For data augmentation, both the original microscopy image and the segmentation maps were randomly rotated, shifted in x-y or flipped (Figure 2B). This created a large set of unique pairs from each pair of microscopy image and its corresponding segmentation map. For an efficient use and evaluation of the data, a 10-fold cross-validation during the training process was chosen [23]. For this validation method, the training dataset was split randomly in ten subsamples (Figure 2C). In each fold, the CNN was trained on nine subsamples while the remaining one was used for validation. This was repeated ten times, so that every subsample was ultimately used once for evaluation of the trained model (Figure 2C). This ensures an efficient use of the training data and allows an initial evaluation of the CNN based on the training data. The trained, feature specific CNN-model, can now be used to compute whole imaging datasets at superhuman speed (Figure 2D).

Model Description

Our CNN (Figure S.1) performs a non-linear pixel-wise classification. The design of the deep neural network is inspired by the U-net architecture [22]. U-net like architectures are, at the current state, the most common architecture for biomedical image segmentation. Moreover, they yield outstanding results in relevant data-science

competitions (e.g. Kaggle Data Science Bowl 2018). The key principle of a U-net is that one computational path stays at the original scale, preserving the spatial information for the output, while the other computational path learns the specific features necessary for classification by applying convolutional filters and thus condensing information [22]. In order to benefit from current research and findings in the quickly emerging field of deep learning, we added two main components to our CNN. First, we attached batch normalization layers [24] and secondly, we replaced the standard 2D-convolutional layers with depthwise separable convolution [25] on the down part of the network. These changes significantly reduced trainable parameters (i.e., weights) and improved training speed while the level of performance was maintained. Our CNN takes greyscale microscopy images as an input and outputs a segmentation map. For each mask, the network outputs a probability distribution of belonging to the positive class (i.e., cFOS-positive nucleus) for each pixel. We implemented the network in Keras [25], a popular high-level open-source API for deep learning, with TensorFlow [26] in the backend. It was trained using the Adam optimizer [27], a commonly used gradient-based function optimizer included in TensorFlow.

Validation of CNN-based segmentation in absence of a ground truth

In absence of a ground truth, classical performance measures like precision or recall cannot be computed. Instead, we quantified the similarity between the segmentation maps of the individual human experts and our CNN to test, whether DeepFLaSH can reach expert-like performance in this segmentation task [28]. The validation of our CNN is representatively explained on the CNN-model for deep segmentation of cFOS labels. For similarity analysis, we calculated the Jaccard similarity (J) [29]. It represents the proportion of significantly overlapping features among all features in two segmentation maps (see methods). To illustrate this metric, three representative cFOS images and segmentation maps with their calculated Jaccard similarity are shown (Figure 3A). First, we calculated the Jaccard similarities of the segmentation maps for all 36 images between all human experts, a threshold-based approach and our CNN trained on cFOS labelled microscopic images (Figure 3B and Figure S.2). The median Jaccard similarity between two experts ranged from 0.33 (Expert 3 vs. Expert 5) to 0.61 (Expert 2 vs. Expert 3, Figure 3B), signifying the high inter-coder variability [6]. This was particularly reflected in the segmentations of Expert 5, who focused rather on high-intensity cFOS labels. Consequently, Expert 5 showed the highest Jaccard similarity compared to signal threshold computation (Figure 3B and Figure S.2). Notably, the segmentation maps created on base of the CNN predictions show an equal similarity to those of the human experts (0.43-0.60, Figure 3B) as they show among themselves (0.33-0.61, Figure 3B). This demonstrates that the similarity and variability of human expert analysis was successfully captured by the CNN-model. On the other side, as one human expert can behave very different from another expert, the similarity analysis indicates that inter-coding of training data may help to increase the objectivity of deep segmentation results. We suggest that CNN training profits from data created by multiple experts, as this may help to include diverse segmentation strategies into one CNN-model. To visualize the segmentation behavior of the individual coders, we computed error maps that show the deviations of each coder from an expert consensus. The expert consensus contains exclusively those pixels that are annotated in at least three of five expert segmentation maps (see methods, Figure 4A). The error maps indicate that the CNN, which was trained on the segmentations of all experts, is similar to the expert consensus, yet not identical (Figure 4B, green inlet). The error maps again point to the different heuristic behavior of the individual experts. For instance, Expert 1, 3 and 4 annotated more cFOS-positive ROIs than those present in the expert consensus (Figure 4C-4F). On the other hand, Expert 5 and the threshold-based approach segmented fewer cFOS-positive ROIs (Figure 4G and 4H). Together, these results demonstrate that our inter-coding approach was successful in training a CNN-model to segment cFOS-positive nuclei in the hippocampus and was able to reach expert-like performance on the initial dataset of 36 images. In order to further validate our CNN design and the generalizability of our approach, we evaluated our model on a new and larger image dataset. This ensures that the CNN-model is not over-fitted on our initial dataset of 36 images, but instead learned the desired features of the fluorescent labels. Therefore, we compared segmentations of the trained CNN to those of a human

expert (Expert 3) on 65 new images of cFOS fluorescent labels in the dorsal hippocampus. Jaccard similarity analysis confirmed that expert-like performance was maintained (Figure 5A and Figure S.3A). In addition, subsequent data analysis based on the segmentation maps of either coder shows equal effects on this dataset (Figure S.3B and S.3C). Therefore, we conclude that labeling experience of human experts was successfully incorporated in the CNN by our inter-coding training approach and that our trained model is suited to capture the complexity of cFOS immunolabels at human expert level.

DeepFLaSH performance in a complex image data analysis

After having established that our CNN-model correctly identifies and segments the desired fluorescent feature (cFOS-positive nuclei), we asked whether the validated model can also be used to analyze and detect changes in cFOS abundance in a complex image dataset. There is a large body of evidence that contextual memory processing leads to increased cFOS expression in the dorsal hippocampus [30, 31, 32, 33, 34]. Thus, the detection of these behavior-related changes in cFOS abundance appears to be a solid test system to evaluate whether our deep learning concept is also suitable for the subsequent image data analysis. Therefore, we used our cFOS-model for the deep segmentation of a large image dataset of anti-cFOS labeled hippocampi of mice after behavioral training (Figure 5B-5J). Based on these segmentation maps, we analyzed the number of cFOS-positive cells and the intensity of cFOS labels per cell in sub-regions of the dorsal hippocampus (CA1, CA3 and dentate gyrus) and compared our results to previous studies. In line with the principle of the 3Rs for animal research - the replacement, reduction and refinement of animal testing - microscopy images of immuno-labeled brain sections cannot be acquired in an unlimited fashion and existing datasets should therefore be used with maximal efficiency. Instead of excluding the images used to train the CNN from the final data analysis, we therefore suggest to re-use the manual analysis by a human expert on the training data and to combine it with the CNN-based analysis on the remaining images. This further facilitates the use of machine learning approaches to the limited dataset sizes in biomedical research. The deep segmentation based analysis of CA1 pyramidal neurons in the dorsal hippocampus revealed an increase in cFOS-positive neurons after retrieval of a previously seen context (C-), as well as after re-exposure to the conditioning context (C+, Figure 5B and 5C) [30, 34]. In CA3, we observed significantly more cFOS-positive cells after context re-exposure (C-), than after retrieval of the fear-associated context (Figure 5G-5H and Figure S.4A) [30, 35, 36, 34]. The dentate gyrus is organized in two blades, the suprapyramidal blade and the infrapyramidal blade, and the neurons of both blades are differently embedded in the hippocampal circuitry [37]. With our computed segmentation maps, we were able to detect an increase in cFOS abundance in the suprapyramidal blades of mice after retrieval of a contextual memory and in mice that retrieved the contextual fear memory (Figure 5G and Figure S.4B) [38, 39]. In contrast, the infrapyramidal blade of the dorsal dentate gyrus showed upregulation of cFOS only after retrieval of the conditioning context, but not after exploration of a previously seen neutral context (Figure 5I and Figure S.4B). Together, these data show that our cFOS-model was able to detect behavior-related changes in all examined regions. Data analysis based on DeepFLaSH segmentations are in line with previous studies. In our data, behavior-related changes were rather reflected in elevated numbers of cFOS-positive cells, than in changes of mean cFOS signal intensities (Figure 5C-5J). Consequently, this demonstrates that our approach to train a CNN with raw image material representing multiple brain regions, different experimental conditions, and with the input of multiple human experts at once was successful to create a model that learned the objective segmentation of the desired fluorescent feature and can be used to automatize image data analysis.

DeepFLaSH-based segmentation of cFOS in Parvalbumin-positive interneurons

To demonstrate the flexibility of our approach, we tested DeepFLaSH on a second fluorescent label. We trained another CNN-model to segment only the cell somata of Parvalbumin-positive (PV+) interneurons in the hippocampal regions CA1, CA3 and the dentate gyrus. Local PV+ interneurons are intensively ramified due to their function in soma-near inhibition of hundreds to thousands of neighboring excitatory neurons [40]. The calcium-binding protein Parvalbumin is distributed throughout the whole cytosol of

these interneurons, including their neurites. This makes analysis of fluorescent labels of Parvalbumin rather intricate compared to cFOS, as it can be quite complex to distinguish a PV+ soma from any PV+ neurite wrapping around a neighboring excitatory neuron. We used again 36 confocal microscopy images showing fluorescent labels for Parvalbumin and the corresponding segmentation maps of PV+ somata created by five human experts to train and evaluate the CNN-model. Notably, the CNN learned to segment also PV+ somata with expert-like performance (Figure S.5 and Figure S.6) and maintained this level of performance on new images as well (Figure 6A). To test whether also this CNN-model was suited for subsequent image data analysis, we used it to investigate behavior-related changes in PV+ interneurons in the hippocampus. As suggested, we again combined the analysis of a human expert (Expert 3) on the training dataset with the analysis of the remaining images by our PV CNN-model to maximize the yield of the limited image data. As expected, we found no differences in the overall number of detected PV+ somata in the hippocampus between the three experimental conditions (Figure S.7A-S.7C). Furthermore, we were not able to detect any context- or fear-dependent variations in the somatic Parvalbumin signal intensity in the dorsal hippocampus, when analyzing all PV+ interneurons (Figure S.7D-S.7F). In order to investigate the possibility of activity-related, rather than behavior-related changes in Parvalbumin-positive interneurons, we combined the two CNN-models for PV+ somata and cFOS-positive signal segmentation. This analysis revealed a higher proportion of cFOS+ PV+ of all PV+ interneurons in the regions CA3 and CA1 in context control (C-) mice, and in the CA1 region of context-conditioned (C+) mice (Figure 6B, 6D and 6E and Figure S.8A) [10, 41]. We did not observe such effects in the dentate gyrus of the dorsal hippocampus (Figure 6C and Figure S.8B). We then calculated for each image the ratio of Parvalbumin signal intensity of cFOS-positive PV+ somata to that of cFOS-negative ones. In this ratio, values greater than one represent a higher Parvalbumin signal intensity in cFOS-positive PV+ somata. Notably, we did not detect any changes in this ratio in the individual regions (Figure S.9A S.9C), nor after pooling the images by conditions (Figure S.9D), by regions (Figure S.9E) or all together (Figure 6F). With these data, we show that our CNN-based analysis revealed behavior-related changes also in the population of Parvalbumin-positive interneurons in the dorsal hippocampus after retrieval of a contextual memory [42]. However, in our dataset, these changes were reflected again in an increased proportion of cFOS+ PV+ interneurons (Figures 6D and 6E), rather than in changes of Parvalbumin signal intensities (Figures 6F, Figure S.7D-S.7F and Figure S.9). Taken together, these results demonstrate that DeepFLaSH can create CNN-models as per individual needs - here the segmentations of cFOS+ nuclei and Parvalbumin-positive somata - and that these models can be used either individually or in conjunction with each other to fully automatize the segmentation of individual image datasets.

Transfer of a label-specific CNN to new datasets with minimal training data

Fluorescent labeling is a standard technique in neuroscience and many groups analyze the same fluorescent features, like cFOS-positive nuclei. Therefore we aimed at testing the generalizability of our cFOS-model to fully independent image datasets that show anti-cFOS labels in different brain regions and that were acquired with different microscopy techniques. Transfer learning is a computational strategy to adapt feature interpretation in one set of data to new datasets with similar features. It allows the adaptation of a pretrained model to similar images with only little extra training data and is hence particularly advantageous for the use of machine learning approaches on datasets with limited sizes, like in biomedical research [17]. To test whether our cFOS-model can be adapted to similar images while maintaining its expert-like performance, we used microscopy images showing cFOS fluorescent labels from three different laboratories (marked as *Lab-Mue*, *Lab-Inns1* and *Lab-Inns2*). These datasets were created in a fully independent manner and included behavioral analysis, brain sectioning, fluorescent labeling, microscopy and image analysis according to lab-specific protocols (see methods: *Lab-Mue*, *Lab-Inns1*, *Lab-Inns2*). First, we generated lab-specific training datasets consisting of only five microscopy images and corresponding manually prepared segmentation maps each (according to Figure 2A). After augmentation of these image pairs (according to Figure 2B), we adopted the cFOS-model, which was pretrained on 36 images of *Lab-Wue*, to create lab-

specific models by means of transfer learning (Figure 7A). Finally, we used these fine-tuned models for the deep segmentation of the corresponding, lab-specific cFOS imaging dataset (Figure 7B). For one dataset (*Lab-Mue*), mice experienced restraint stress and subsequent Pavlovian fear conditioning (cue-conditioning, tone-footshock association). The number of cFOS-positive cells in the paraventricular thalamus (PVT) was compared between early (eRet) and late (lRET) phases of fear memory retrieval. The analyses by two experts of *Lab-Mue* and the fine-tuned CNN-Mue, all revealed a significantly higher number of cFOS-positive cells in the PVT of mice 24h after fear conditioning (lRET; Figure 7C). In *Lab-Inns1*, mice underwent Pavlovian fear conditioning (cue-conditioning, tone-footshock association) and extinction in the same context. Again, we compared the analysis of the experimenter from that lab (*Lab-Inns1*) with deep segmentation using the CNN-Inns1, which resulted from transfer learning solely on five images. Both, the human expert and the CNN found an increased number of cFOS-positive cells in the basolateral amygdala (BLA) after extinction of a previously learned fear (Figure 7D). A third image dataset was provided by *Lab-Inns2* and shows cFOS immunoreactivity in the infralimbic cortex (IL) following fear renewal (return of extinguished fear in a context different from the extinction training context). In 129S1/SvlmJ mice, which display impaired fear extinction acquisition and extinction consolidation, we have previously shown that enhancing dopaminergic signaling by L-DOPA (L-3,4-Dihydroxyphenylalanine) treatment rescued deficient fear extinction and co-administration of a cognitive enhancer (MS-275) rendered this effect enduring and context-independent [43, 44]. In the present study, we replicated our findings that L-DOPA/MS-275 reduces fear renewal (see methods) and now show for the first time that this reduction of fear renewal is associated with an increased number of cFOS-positive cells in the infralimbic cortex. Furthermore, since heterogeneity in this behavioral response was observed, mice were classified as responders or non-responders, based on freezing responses, which was more than 2 x Std. deviations from the average of the responders. Both, the human expert and the corresponding CNN-Inns2 found increased numbers of cFOS-positive cells only in the infralimbic cortex of the L-DOPA responders, compared to controls as well as non-responders (Figure 7E). These analyses confirmed that a CNN-model trained with DeepFLaSH on one dataset can be adapted to similar images, albeit cFOS labeling and image acquisition was done differently (see methods) and different brain regions (here PVT, BLA and IL) were investigated. This gives experimental proof that pretrained CNN-models can easily be adapted to data from different laboratories, with a very limited number of training images (here five images) and can perform on the new data with expertise.

Discussion

Here, we introduce DeepFLaSH, an open-source deep-learning pipeline for fluorescent label segmentation that learns from human-experts. DeepFLaSH is open to be used in local computing facilities, but is also implemented in a computational notebook and runs as an interactive web tool in a computer cloud (Jupyter notebook in Google Colab). The deep learning network we developed can be fed with pairs of raw images showing fluorescent labels and the corresponding manually created segmentation maps. Once trained on some image-segmentation pairs, the label-specific convolutional neural network (CNN) can be exploited to analyze independent microscopy image datasets of similar labels, while maintaining expert-like performance. This adaption of pretrained models can dramatically reduce the time and effort required for CNN training. Consequently, this urges the creation of open-access CNN-model libraries that allow the quick adaptation of a suitable model to individual demands.

Inter-coding reliability and objectivity of CNNs for fluorescent feature segmentation

Deep learning segmentation approaches, such as DeepEM3D [8], its cloud-version CDeep3M [18] or DeepFLaSH (this paper), share the potential to be applicable for a wide range of segmentation tasks. However, deep learning requires ground truth information, meaning data objectively being true. Ground truthing for image segmentation tasks can be done automatically or manually, but both depend on heuristic information [18, 45, 8] and are therefore never fully objective. This is a minor problem, when fluorescent signals provide a high signal-to-noise ratio, as seen for cell nucleus stains with DNA binding dyes [18, 45]. In such experiments, objective measures are comparable to

subjective assessments. However, when fluorescent signals are not homogenous and are of relevance over a wide signal-to-noise spectrum, manual segmentation becomes rather demanding and subjectivity increases [5, 6] and this study). Anyhow, manual, heuristic segmentation is still the most common method to generate training data for segmentation tasks [18, 45]. In the present study, we trained our CNN-model by means of multiple segmentation maps created by different experts for each training image. This approach assumes no ground truth per se but pools the input from multiple independent experts who were unaware of the segmentations done by the other experts. Subsequently, we used similarity analysis to compare the segmentations of the individual coders (experts, CNN-models and thresholding). We found that our CNN-models showed higher similarities to four of the five experts (Experts 1-4) and less to Expert 5, which was in accordance with higher similarities within Experts 1-4 compared to Expert 5. It indicates that the CNN is capable of learning the congruent information present in the input of multiple experts and can balance individual behavior. Based on our data, we therefore suggest to always use input from multiple experts to train a CNN. This helps to improve reproducibility and objectivity. In contrast, CNNs based on single expert training data may be rather subjective and might create data that are not easy to reproduce. To facilitate the validation of CNN-predictions, we implemented similarity analysis in the pipeline of DeepFLaSH, so that the user can easily compare the CNN output with manual segmentations.

CNN-model libraries for feature segmentation

DeepFLaSH is designed to be used already with limited training data. We showed, that for two fluorescent labels in brain slices, the neuronal plasticity marker cFOS and the calcium-binding protein Parvalbumin, training on just 36 images was sufficient to create label-specific CNNs that behaved like human experts. Furthermore, we show that a pretrained CNN-model could easily be adapted to new and independent datasets with only five images. Consequently, storage of validated CNN-models in open-access libraries offers great opportunities. For example, a cFOS label is in its signature indistinguishable from a variety of other fluorescent labels, like those of transcription factors (CREB, phospho-CREB, Pax6, NeuroG2 or Brain3a), cell division markers (phospho-histone H3), apoptosis markers (Caspase-3) and multiple others. Therefore, we surmise that fast transfer learning will allow the adaptation of our pretrained cFOS-model as a general tool for nucleosomatic fluorescent label segmentation in brain slices. This highlights the great potential of DeepFLaSH and the creation of CNN-model libraries for the life science community.

Accessibility of DeepFLaSH

Deep learning algorithms require high computing power (graphics processing unit, GPUs or tensor processing units, TPUs) and artificial intelligence expertise that are still rarely found at biomedical research facilities. To enable facilitated access for the life science community, we implemented DeepFLaSH as an open-source tool that is easily accessible in a cloud-based environment. It allows out-of-the-box segmentation and adaption of pretrained models as per individual demands. Using the free service and computational power of Google Colab (http://colab.research.google.com) in a Jupyter Notebook, DeepFLaSH is very user-intuitive and ensures that no compatibility problems can arise. Jupyter became the computational notebook of choice for data scientists [46] and allows interactive guidance through DeepFLaSH, also for non-AI experts.

Conclusion

We highly recommend creating open source libraries with label-specific CNN-models for broad use in the neuroscience community. Label-specific CNN-models, validated on base of inter-coding approaches may become a new benchmark for feature segmentation in neuroscience. These models will allow transferring expert performance in image feature analysis from one lab to any other. Deep segmentation can better interpret feature-to-noise borders, can work on the whole dynamic range of bit-values and exhibits consistent performance. This should increase both, objectivity and reproducibility of image feature analysis. DeepFLaSH is suited to create CNN-models for high-throughput microscopy techniques and allows automatic analysis of large image datasets with expert-like performance and at super-human speed. DeepFLaSH is easy to use, can run in a computer cloud and is provided as an interactive web tool known as open source computational notebook.

Acknowlegments

The following researchers were supported by the Deutsche Forschungsgemeinschaft SFB-TRR58 Fear, Anxiety and Anxiety Disorders: R.B. and M.Se. project A10, H.C.P. project A03 and M.D.L. project B08. R.G. and M.Sa. were supported by fellowships of the Graduate School of Life Sciences (GSLS) Wrzburg. R.O.T. was supported by the Austrian Science Fund (FWF) P29952 & P25851. N.Si. was supported by the FWF (Austrian Science Fund, I2433-B26, DK W-1206 and SFB F4410).

Author Contributions

Conceptualization: D.S., M.G., A.D., N.St., C.M.F., and R.B.; Methodology: D.S., M.G., A.D., C.R.v.C., C.M., D.F., L.B.C., A.S., N.St., R.G., M.Sa., M.D.L., R.O.T., N.Si., H.-C.P., M.Se., C.M.F., and R.B.; Software: D.S., M.G., A.D., and C.M.F.; Validation: D.S., M.G., A.D., C.R.v.C., C.M., D.F., L.B.C., A.S., R.G., M.Sa., M.D.L., C.M.F., and R.B.; Formal Analysis: D.S., M.G., A.D.; Investigation: D.S., M.G., A.D., C.R.v.C., C.M., D.F., L.B.C., A.S., R.G., M.Sa., and M.D.L.; Resources: M.D.L., R.O.T., N.Si., H.-C.P., M.Se., C.M.F., and R.B.; Data Curation: D.S., M.G., and A.D.; Writing Original Draft: D.S., M.G., A.D., C.M.F., and R.B.; Writing Review & Editing: D.S., M.G., C.M.F., and R.B.; Visualization: D.S., M.G., and A.D.; Supervision: C.M.F. and R.B.; Project Administration: D.S., M.G., A.D., M.D.L., R.O.T., N.Si., H.-C.P., M.Se., C.M.F., and R.B.; Funding Acquisition: R.G., M.Sa., M.D.L., R.O.T., N.Si., H.-C.P., M.Se., C.M.F., and R.B.

Declaration of Interests

All authors declare no conflict of interest.

References

[1] A. Li, H. Gong, B. Zhang, Q. Wang, C. Yan, J. Wu, Q. Liu, S. Zeng, Q. Luo, Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain, Science 330 (2010) 1404–1408.

[2] K. McDole, L. Guignard, F. Amat, A. Berger, G. Malandain, L. A. Royer, S. C. Turaga, K. Branson, P. J. Keller, In toto imaging and reconstruction of post-implantation mouse development at the single-cell level, Cell (2018).

[3] P. Osten, T. W. Margrie, Mapping brain circuitry with a light microscope, Nat Methods 10 (2013) 515–23.

[4] D. C. Collier, S. S. Burnett, M. Amin, S. Bilton, C. Brooks, A. Ryan, D. Roniger, D. Tran, G. Starkschall, Assessment of consistency in contouring of normal-tissue anatomic structures, J Appl Clin Med Phys 4 (2003) 17–24.

[5] C. J. Niedworok, A. P. Brown, M. Jorge Cardoso, P. Osten, S. Ourselin, M. Modat, T. W. Margrie, amap is a validated pipeline for registration and segmentation of high-resolution mouse brain data, Nat Commun 7 (2016) 11879.

[6] C. Schmitz, H. Korr, H. Heinsen, Design-based counting techniques: the real problems, Trends Neurosci 22 (1999) 345–6.

[7] C. A. Schneider, W. S. Rasband, K. W. Eliceiri, Nih image to imagej: 25 years of image analysis, Nat Methods 9 (2012) 671–5.

[8] T. Zeng, B. Wu, S. Ji, Deepem3d: approaching human-level performance on 3d anisotropic em image segmentation, Bioinformatics 33 (2017) 2555–2562.

[9] S. A. Shuvaev, A. A. Lazutkin, A. V. Kedrov, K. V. Anokhin, G. N. Enikolopov, A. A. Koulakov, Dalmatian: An algorithm for automatic cell detection and counting in 3d, Front Neuroanat 11 (2017) 117.

[10] N. Guo, M. E. Soden, C. Herber, M. T. Kim, A. Besnard, P. Lin, X. Ma, C. L. Cepko, L. S. Zweifel, A. Sahay, Dentate granule cell recruitment of feedforward inhibition governs engram maintenance and remote memory generalization, Nat Med 24 (2018) 438–449.

[11] L. Li, X. Feng, Z. Zhou, H. Zhang, Q. Shi, Z. Lei, P. Shen, Q. Yang, B. Zhao, S. Chen, L. Li, Y. Zhang, P. Wen, Z. Lu, X. Li, F. Xu, L. Wang, Stress accelerates defensive responses to looming in mice and involves a locus coeruleus-superior colliculus projection, Curr Biol 28 (2018) 859–871 e5.

[12] J. D. Zaremba, A. Diamantopoulou, N. B. Danielson, A. D. Grosmark, P. W. Kaifosh, J. C. Bowler, Z. Liao, F. T. Sparks, J. A. Gogos, A. Losonczy, Impaired hippocampal place cell dynamics in a mouse model of the 22q11.2 deletion, Nat Neurosci 20 (2017) 1612–1623.

[13] J. C. Caicedo, S. Cooper, F. Heigwer, S. Warchal, P. Qiu, C. Molnar, A. S. Vasilevich, J. D. Barry, H. S. Bansal, O. Kraus, M. Wawer, L. Paavolainen, M. D. Herrmann, M. Rohban, J. Hung, H. Hennig, J. Concannon, I. Smith, P. A. Clemons, S. Singh, P. Rees, P. Horvath, R. G. Linington, A. E. Carpenter, Data-analysis strategies for image-based cell profiling, Nat Methods 14 (2017) 849–863.

[14] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–44.

[15] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118.

[16] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. N. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, K. Zhang, Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (2018) 1122–1131 e9.

[17] E. M. Christiansen, S. J. Yang, D. M. Ando, A. Javaherian, G. Skibinski, S. Lipnick, E. Mount, A. O'Neil, K. Shah, A. K. Lee, P. Goyal, W. Fedus, R. Poplin, A. Esteva, M. Berndl, L. L. Rubin, P. Nelson, S. Finkbeiner, In

silico labeling: Predicting fluorescent labels in unlabeled images, Cell 173 (2018) 792–803 e19.

[18] M. G. Haberl, C. Churas, L. Tindall, D. Boassa, S. Phan, E. A. Bushong, M. Madany, R. Akay, T. J. Deerinck, S. T. Peltier, M. H. Ellisman, Cdeep3m-plug-and-play cloud-based deep learning for image segmentation, Nat Methods 15 (2018) 677–680.

[19] J. I. Morgan, D. R. Cohen, J. L. Hempstead, T. Curran, Mapping patterns of c-fos expression in the central nervous system after seizure, Science 237 (1987) 192–7.

[20] J. F. Pechere, Muscular parvalbumins as homologous proteins, Comp Biochem Physiol 24 (1968) 289–95.

[21] J. Victoroff, W. J. Mack, S. T. Grafton, S. S. Schreiber, H. C. Chui, A method to improve interrater reliability of visual inspection of brain mri scans in dementia, Neurology 44 (1994) 2267.

[22] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, arXiv:1505.04597 (2015).

[23] R. Kohavi, A study of cross validation and bootstrap for accuracy estimation and model selection, International Joint Conference on Artificial Intelligence (1995).

[24] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv:1502.03167 (2015).

[25] F. Chollet, Keras, https://keras.io (2015).

[26] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, Tensorflow: Large-scale machine learning on heterogeneous distributed systems, arXiv:1603.04467 (2016).

[27] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980 (2014).

[28] S. K. Warfield, K. H. Zou, W. M. Wells, Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation, IEEE Trans Med Imaging 23 (2004) 903–21.

[29] P. Jaccard, The distribution of the flora in the alpine zone, The new phytologist 11 (1912) 37–50.

[30] S. Campeau, W. A. Falls, W. E. Cullinan, D. L. Helmreich, M. Davis, S. J. Watson, Elicitation and reduction of fear: behavioural and neuroendocrine indices and brain induction of the immediate-early gene c-fos, Neuroscience 78 (1997) 1087–1104.

[31] N. C. Huff, M. Frank, K. Wright-Hardesty, D. Sprunger, P. Matus-Amat, E. Higgins, J. W. Rudy, Amygdala regulation of immediate-early gene expression in the hippocampus induced by contextual fear conditioning, J Neurosci 26 (2006) 1616–23.

[32] X. Liu, S. Ramirez, P. T. Pang, C. B. Puryear, A. Govindarajan, K. Deisseroth, S. Tonegawa, Optogenetic stimulation of a hippocampal engram activates fear memory recall, Nature 484 (2012) 381–5.

[33] K. Z. Tanaka, H. He, A. Tomar, K. Niisato, A. J. Y. Huang, T. J. McHugh, The hippocampal engram maps experience but not place, Science 361 (2018) 392–397.

[34] K. K. Tayler, K. Z. Tanaka, L. G. Reijmers, B. J. Wiltgen, Reactivation of neural ensembles during the retrieval of recent and remote memory, Curr Biol 23 (2013) 99–106.

[35] C. A. Denny, M. A. Kheirbek, E. L. Alba, K. F. Tanaka, R. A. Brachman, K. B. Laughman, N. K. Tomm, G. F. Turi, A. Losonczy, R. Hen, Hippocampal memory traces are differentially modulated by experience, time, and adult neurogenesis, Neuron 83 (2014) 189–201.

[36] I. Lee, R. P. Kesner, Differential contributions of dorsal hippocampal subregions to memory acquisition and retrieval in contextual fear-conditioning, Hippocampus 14 (2004) 301–10.

[37] B. Schmidt, D. F. Marrone, E. J. Markus, Disambiguating the similar: the dentate gyrus and pattern separation, Behav Brain Res 226 (2012) 56–65.

[38] M. K. Chawla, J. F. Guzowski, V. Ramirez-Amaya, P. Lipa, K. L. Hoffman, L. K. Marriott, P. F. Worley, B. L. McNaughton, C. A. Barnes, Sparse, environmentally selective expression of arc rna in the upper blade of the rodent fascia dentata by brief spatial experience, Hippocampus 15 (2005) 579–86.

[39] E. Satvat, B. Schmidt, M. Argraves, D. F. Marrone, E. J. Markus, Changes in task demands alter the pattern of zif268 expression in the dentate gyrus, J Neurosci 31 (2011) 7163–7.

[40] H. Hu, J. Gan, P. Jonas, Interneurons. fast-spiking, parvalbumin(+) gabaergic interneurons: from cellular design to microcircuit function, Science 345 (2014) 1255263.

[41] N. Ognjanovski, S. Schaeffer, J. Wu, S. Mofakham, D. Maruyama, M. Zochowski, S. J. Aton, Parvalbumin-expressing interneurons coordinate hippocampal network dynamics required for memory consolidation, Nat Commun 8 (2017) 15039.

[42] S. Karunakaran, A. Chowdhury, F. Donato, C. Quairiaux, C. M. Michel, P. Caroni, Pv plasticity sustained through d1/5 dopamine signaling required for long-term memory consolidation, Nat Neurosci 19 (2016) 454–64.

[43] J. Haaker, S. Gaburro, A. Sah, N. Gartmann, T. B. Lonsdorf, K. Meier, N. Singewald, H. C. Pape, F. Morellini, R. Kalisch, Single dose of l-dopa makes extinction memories context-independent and prevents the return of fear, Proc Natl Acad Sci U S A 110 (2013) E2428–36.

[44] N. Whittle, V. Maurer, C. Murphy, J. Rainer, D. Bindreither, M. Hauschild, A. Scharinger, M. Oberhauser, T. Keil, C. Brehm, T. Valovka, J. Striessnig, N. Singewald, Enhancing dopaminergic signaling and histone acetylation promotes long-term rescue of deficient fear extinction, Transl Psychiatry 6 (2016) e974.

[45] D. A. Van Valen, T. Kudo, K. M. Lane, D. N. Macklin, N. T. Quach, M. M. DeFelice, I. Maayan, Y. Tanouchi, E. A. Ashley, M. W. Covert, Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments, PLoS Comput Biol 12 (2016) e1005177.

[46] J. M. Perkel, By jupyter, it all makes sense, Nature 563 (2018) 145–146.

[47] M. S. Fanselow, Conditioned and unconditional components of post-shock freezing, Pavlov J Biol Sci 15 (1980) 177–82.

[48] F. Chauveau, M. D. Lange, K. Jungling, J. Lesting, T. Seidenbecher, H.-C. Pape, Prevention of stress-impaired fear extinction through neuropeptide s action in the lateral amygdala, Neuropsychopharmacology 37 (2012) 1588–1599.

[49] C. P. Murphy, X. Li, V. Maurer, M. Oberhauser, R. Gstir, L. E. Wearick-Silva, T. W. Viola, S. Schafferer, R. Grassi-Oliveira, N. Whittle, A. Huttenhofer, T. W. Bredy, N. Singewald, Microrna-mediated rescue of fear extinction memory by mir-144-3p in extinction-impaired mice,

Biol Psychiatry 81 (2017) 979–989.

[50] T. M. Gruene, K. Flick, A. Stefano, S. D. Shea, R. M. Shansky, Sexually divergent expression of active and passive conditioned fear responses in rats, Elife 4 (2015).

[51] G. Paxinos, K. B. J. Franklin, The mouse brain in stereotaxic coordinates, Elsevier Academic Press, Amsterdam ; Boston, compact 2nd edition, 2004.

[52] K. B. J. Franklin, G. Paxinos, The mouse brain in stereotaxic coordinates, Elsevier, Amsterdam [u.a.], compact 3. edition, 2008.

[53] H. J. Van De Werd, G. Rajkowska, P. Evers, H. B. Uylings, Cytoarchitectonic and chemoarchitectonic characterization of the prefrontal cortical areas in the mouse, Brain Struct Funct 214 (2010) 339–53.

[54] P. J. Fitzgerald, N. Whittle, S. M. Flynn, C. Graybeal, C. R. Pinard, O. Gunduz-Cinar, A. V. Kravitz, N. Singewald, A. Holmes, Prefrontal single-unit firing associated with deficient extinction in mice, Neurobiol Learn Mem (2013).

[55] N. Whittle, M. Hauschild, G. Lubec, A. Holmes, N. Singewald, Rescue of impaired fear extinction and normalization of cortico-amygdala circuit dysfunction in a genetic mouse model by dietary zinc restriction, J Neurosci 30 (2010) 13586–96.

[56] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, 2016 Fourth International Conference on 3D Vision (3DV) (2016) 565–571.

[57] W. R. Crum, O. Camara, D. L. Hill, Generalized overlap measures for evaluation and validation in medical image analysis, IEEE Trans Med Imaging 25 (2006) 1451–61.

[58] R. Cardenes, R. de Luis-Garcia, M. Bach-Cuadra, A multidimensional segmentation evaluation for medical image data, Comput Methods Programs Biomed 96 (2009) 108–24.

[59] C. Molnar, I. H. Jermyn, Z. Kato, V. Rahkama, P. Ostling, P. Mikkonen, V. Pietiainen, P. Horvath, Accurate morphology preserving segmentation of overlapping cells based on active contours, Sci Rep 6 (2016) 32412.

[60] V. Ulman, M. Maska, K. E. G. Magnusson, O. Ronneberger, C. Haubold, N. Harder, P. Matula, P. Matula, D. Svoboda, M. Radojevic, I. Smal, K. Rohr, J. Jalden, H. M. Blau, O. Dzyubachyk, B. Lelieveldt, P. Xiao, Y. Li, S. Y. Cho, A. C. Dufour, J. C. Olivo-Marin, C. C. Reyes-Aldasoro, J. A. Solis-Lemus, R. Bensch, T. Brox, J. Stegmaier, R. Mikut, S. Wolf, F. A. Hamprecht, T. Esteves, P. Quelhas, O. Demirel, L. Malmstrom, F. Jug, P. Tomancak, E. Meijering, A. Munoz-Barrutia, M. Kozubek, C. Ortiz-de Solorzano, An objective comparison of cell-tracking algorithms, Nat Methods 14 (2017) 1141–1152.

[61] M. Maska, V. Ulman, D. Svoboda, P. Matula, P. Matula, C. Ederra, A. Urbiola, T. Espana, S. Venkatesan, D. M. Balak, P. Karas, T. Bolckova, M. Streitova, C. Carthel, S. Coraluppi, N. Harder, K. Rohr, K. E. Magnusson, J. Jalden, H. M. Blau, O. Dzyubachyk, P. Krizek, G. M. Hagen, D. Pastor-Escuredo, D. Jimenez-Carretero, M. J. Ledesma-Carbayo, A. Munoz-Barrutia, E. Meijering, M. Kozubek, C. Ortiz-de Solorzano, A benchmark for comparison of cell tracking algorithms, Bioinformatics 30 (2014) 1609–17.

[62] C. T. Rueden, J. Schindelin, M. C. Hiner, B. E. DeZonia, A. E. Walter, E. T. Arena, K. W. Eliceiri, ImageJ2: ImageJ for the next generation of scientific image data, BMC Bioinformatics 18 (2017) 1–26.

[63] E. Jones, T. Oliphant, P. Peterson, et al., SciPy: Open source scientific tools for Python, http://www.scipy.org/ (2001).

Figures and legends:

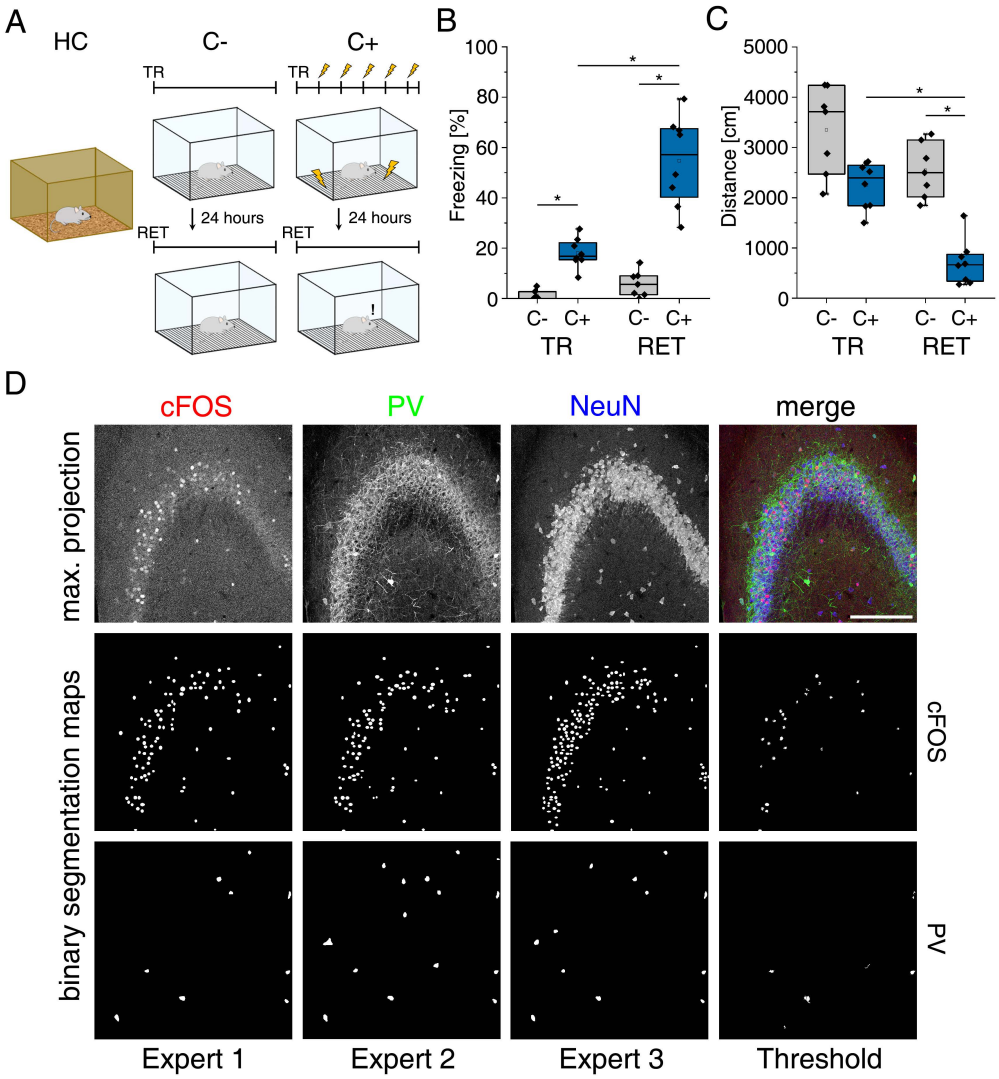

Figure 1: **Contextual fear conditioning, immunolabeling and manual segmentation of fluorescent labels in imaging data**
(A) Mouse behavior. Three experimental groups were investigated: Mice kept in their homecage (HC), mice that were trained to a context, but did not experience an electric foot shock (C-), and mice exposed to five foot shocks in the training context (C+). 24 hours after the initial training (TR), mice were re-exposed to the training context for memory retrieval (RET).
(B) Fear acquisition was observed in conditioned mice (C+), while unconditioned controls (C-) did not show freezing behavior during initial context exposure (TR). In the memory retrieval session (RET), conditioned mice showed strong freezing behavior, while unconditioned mice did not freeze in response to the training context. ($X^2(3) = 25.330$, $p < 0.0001$, $N_{TR\ C-} = 7$, $N_{TR\ C+} = 8$, $N_{RET\ C-} = 7$, $N_{RET\ C+} = 8$, Kruskal-Wallis ANOVA followed by pairwise Mann-Whitney tests with Bonferroni correction, *P < 0.05).
(C) Distance travelled in the training context is reduced in fear conditioned mice. (X^2 (3)= 19.988, $p < 0.01$, $N_{TR\ C-} = 7$, $N_{TR\ C+} = 8$, $N_{RET\ C-} = 7$, $N_{RET\ C+} = 8$, Kruskal-Wallis ANOVA followed by pairwise Mann-Whitney tests with Bonferroni correction, *P < 0.05).
(D) Representative confocal microscopy image (maximum intensity projection) showing the CA3 region in the dorsal hippocampus of a context conditioned (C+) mouse. Triple-label for cFOS, Parvalbumin (PV) and NeuN. Merge is the overlay of all three fluorescent labels. Binary segmentation maps show annotated ROIs for cFOS+ nuclei or PV+ somata. The segmentation maps created by three human experts indicate the variability in heuristic analysis of the same image. Threshold-based segmentation reliably annotates ROIs with strong fluorescent labels. Scale bar: 200 µm.

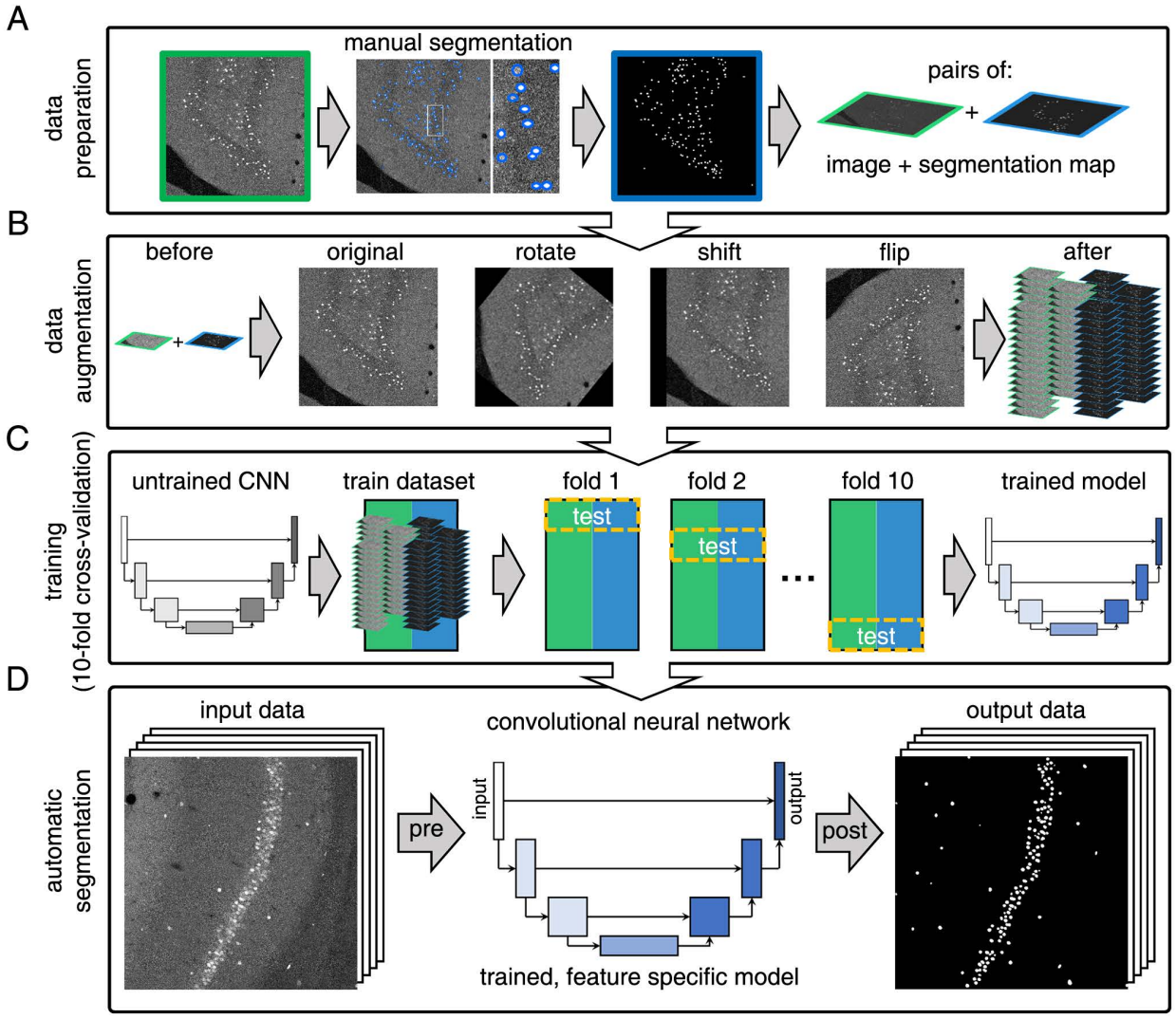

Figure 2: **Workflow for DeepFLaSH**
(A) Data preparation. Data consist of fluorescence microscopy images of immunolabeled brain sections. To create a training dataset, a microscopy image (green frame) is manually segmented by one or multiple human experts. The segmentation data are used to compute a binary image showing the segmented fluorescent features (ROIs) in white and unrelated image elements in black (segmentation map, marked by blue frame). Each image and its corresponding segmentation map form a pair of images.
(B) Computational data augmentation. The image pairs are augmented to artificially increase the number of training data images. For this, images are randomly rotated, shifted in x, y, or flipped to create unique pairs of images.
(C) CNN-Training. The augmented dataset is then used to train a convolutional neural network (CNN; U-net). The training process includes a 10-fold cross-validation procedure. For this, the training dataset is split randomly into ten subsamples and in each fold, nine are used to train the algorithm, while the remaining one is used for evaluation. After ten folds, each subsample is used once for validation. This reduces overfitting and allows the user to perform a preliminary CNN validation already on the training data set. The trained CNN-model is specific for a learned fluorescent feature. To increase the objectivity of a CNN-model, we highly recommend pooling the training data generated by multiple independent experts on base of the same images (inter-coding approach).
(D) Automatic segmentation. After generic pre-processing (pre) of the microscope images, the CNN-model can be used to segment fluorescent labels in image datasets at super-human speed. In post-processing (post), the predicted segments are used to generate binary segmentation maps (ROI masks) for the subsequent data analysis.

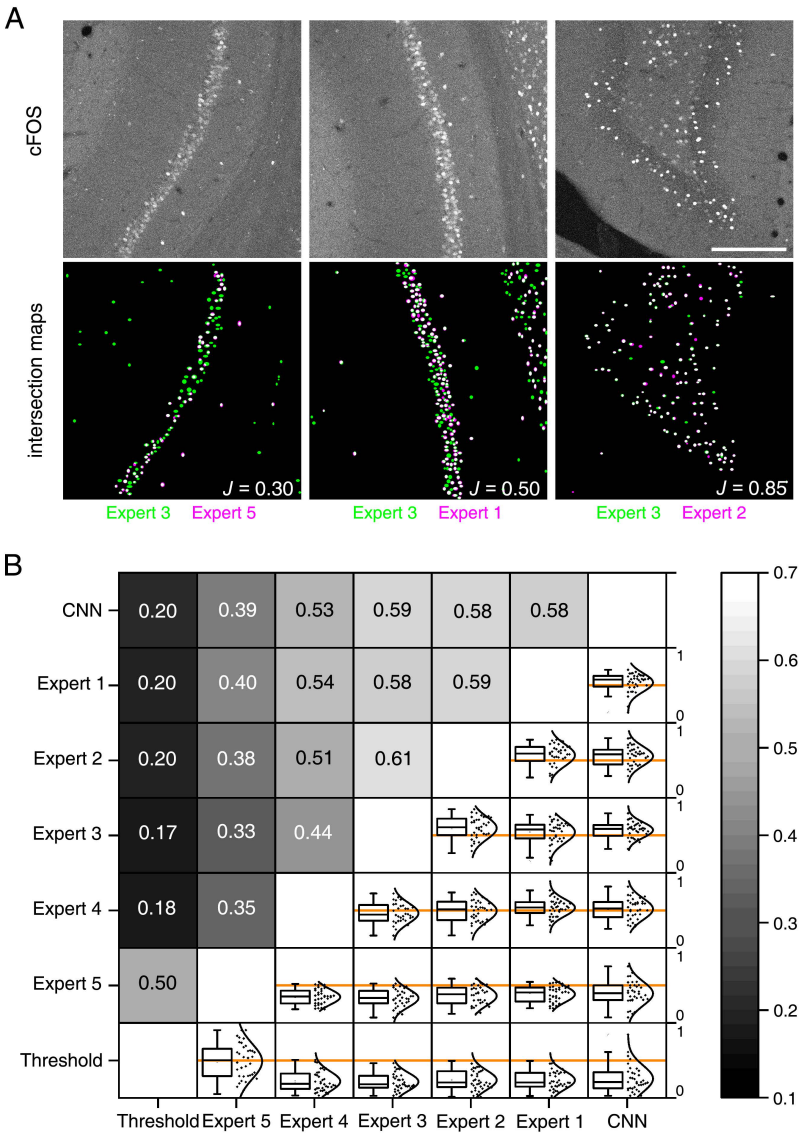

Figure 3: **Automatic segmentation of cFOS immunolabels by a trained CNN-model shows expert-like performance**
Five experts (Expert 1 – 5) in neurobiology independently segmented cFOS-positive nuclei in microscopy images to create expert-specific segmentation maps. Manual segmentation cannot create an ultimate ground truth. To evaluate the individual segmentation performance in absence of a ground truth, we therefore calculated the similarities between the segmentation maps of all experts, those created by the CNN and a threshold-based segmentation tool. The figure shows the performance of the cFOS-model, which was trained in an inter-coding approach, in comparison to human experts.
(A) Microscopy images showing cFOS immunoreactivity (upper row) and the corresponding segmentation maps of cFOS+ nuclei by two representative experts (one in green, one in magenta) and the overlay of both maps (intersection in white; lower row). The Jaccard similarity between two segmentation maps (ROI-level) is indicated. The three examples visualize three representative Jaccard similarities of 0.3 (low), 0.5 (medium), and 0.85 (high). The Jaccard similarity of 0.85 represents the best overlap between segmentation maps created by human experts. Scale bar: 200 µm.
(B) Heat map showing the Jaccard similarity (ROI-level) between corresponding segmentation maps. Compared are the segmentation maps of the five human experts, a semi-automated signal thresholding approach and the trained CNN-model. The Jaccard similarities are shown as median value (color-coded, upper-left half) and as boxplot, together with the individual data points in the normal distribution curve (lower-right half). The orange lines mark the value of 0.5 (n=36 for each comparison between two coders). The data shows that the CNN-based segmentations are as similar to those of human experts, as they are among themselves (inter-coder variability).

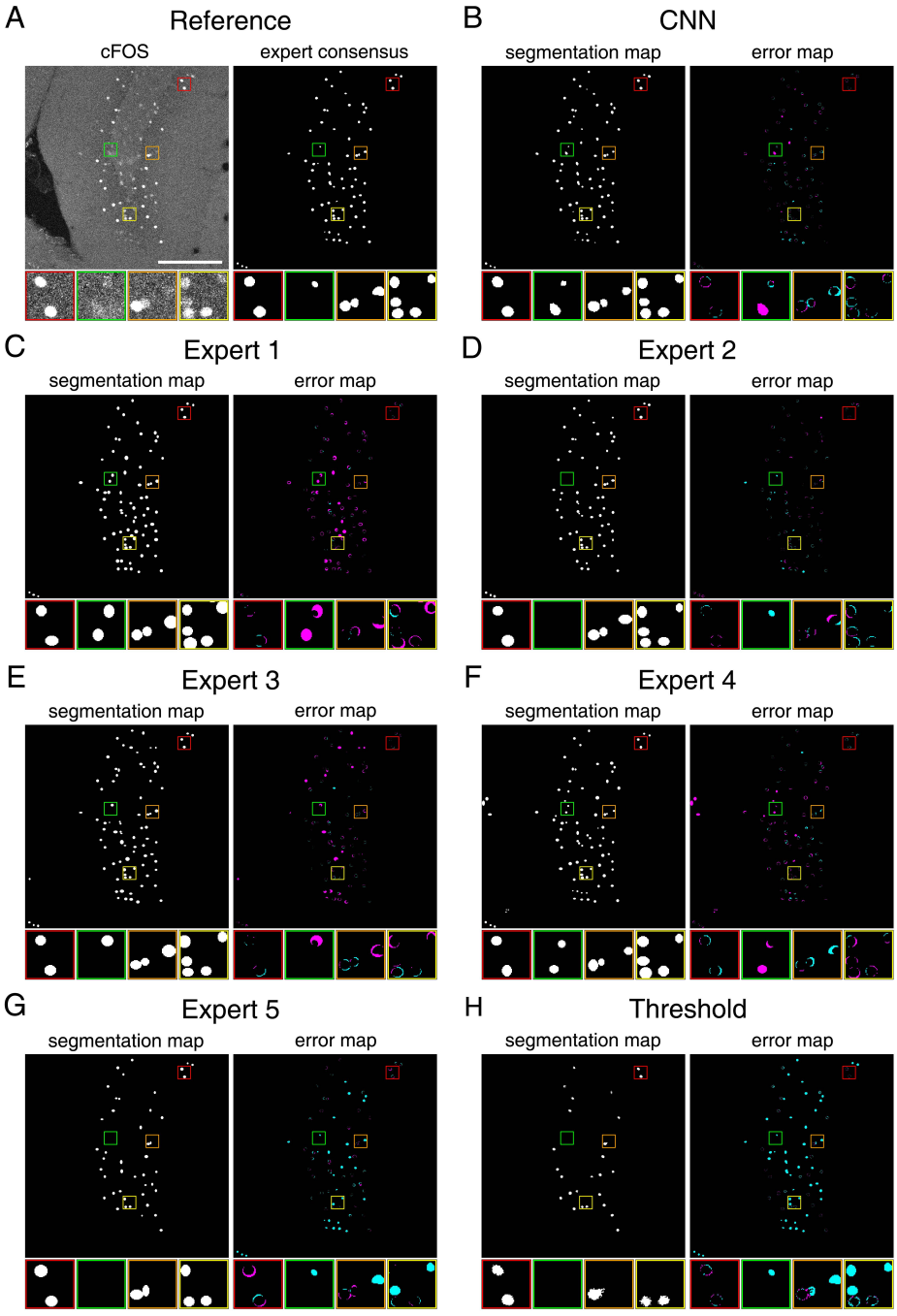

Figure 4: **Accuracy of segmentation of cFOS-positive nuclei at pixel-wise resolution**
Pixels that overlapped in the segmentation data of at least three of the five experts were used to create a segmentation map that represents the expert consensus. This consensus was used to compute error maps that visualize the deviation of each coders segmentation behavior from the expert consensus.
(A) Comparison of cFOS microscopy image with expert consensus segmentation map. Four inlets are selected to highlight the variability of typical anti-cFOS fluorescent labels.
(B-H) Segmentation maps and computed error maps are shown for the indicated coder. The error maps are pixel-wise comparisons of the corresponding coder segmentation to the expert consensus map. In cyan: pixels exclusively present in the expert consensus; in magenta: pixels that were exclusively labeled by the indicated coder. Individual behavior of experts is best seen in the high similarity between Expert 5 and the threshold-based approach and between the CNN and Experts 1-4.

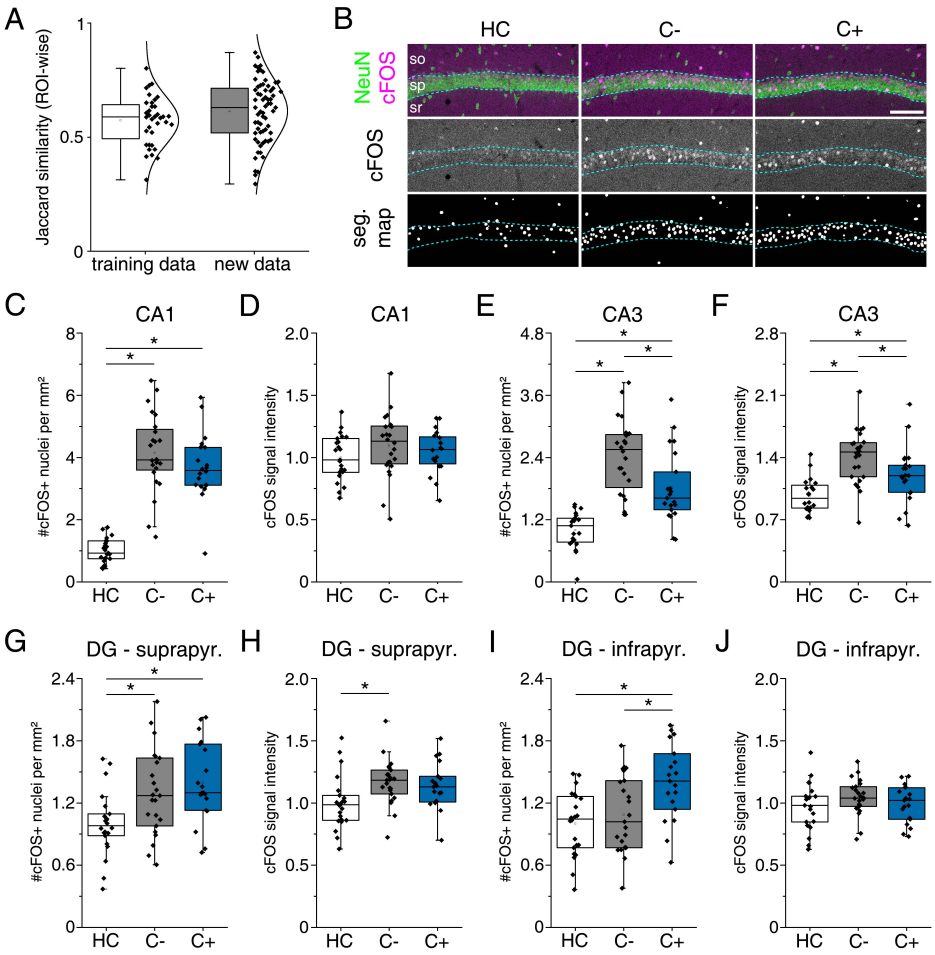

Figure 5: **Image analysis with a DeepFLaSH-model for cFOS fluorescent label segmentation**
Behavior-induced changes in hippocampal cFOS were analyzed with DeepFLaSH-computed segmentation maps. The deep segmentation maps (161 images) and manual segmentation data (36 images) were pooled for the statistical analysis to ensure the use of all data.
(A) Comparison of ROI-wise Jaccard similarities between the CNN and Expert 3 across the set of 36 images used for training (white box) and 65 new images (grey box). The trained cFOS-model maintains expert-like performance on new images ($p > 0.05$, Welchs two-sample t-test, $n_{training\ data} = 36$, new data: $n_{new\ data} = 65$).
(B) Representative images showing cFOS labels in CA1 of the three experimental groups (HC, C-, C+). A dashed line marks the analyzed area in CA1. The segmentation maps are based on CNN predictions. Green: NeuN, magenta: cFOS; so: stratum oriens, sp: stratum pyramidale, sr: stratum radiatum. Scale bar: 100 µm.
(C-J) Analysis of cFOS fluorescent labels in stratum pyramidale (CA1 and CA3) or granule cell layer (DG blades). All values are normalized to the mean of the respective homecage control (for all regions: $N_{HC} = 4$, $N_{C-} = 5$, $N_{C+} = 4$).
(in C, E, G and I) Analysis of cFOS+ nuclei per mm^2 in indicated regions. Data reveal context-dependent increase in the number of cFOS+ nuclei in CA1, CA3, and the suprapyramidal blade of the DG. (CA1: $X^2(2) = 43.630$, $p<0.001$, $n_{HC} = 23$, $n_{C-} = 24$, $n_{C+} = 20$, Kruskal-Wallis ANOVA followed by pairwise Mann-Whitney tests with Bonferroni correction, $^*P<0.05$; CA3: $X^2(2) = 37.032$, $p<0.001$, $n_{HC} = 21$, $n_{C-} = 24$, $n_{C+} = 21$, Kruskal-Wallis ANOVA followed by pairwise Mann-Whitney tests with Bonferroni correction, $^*P<0.05$; DG infrapyr.: $F_{(2,\ 61)} = 7.272$, $p<0.01$, $n_{HC} = 22$, $n_{C-} = 23$, $n_{C+} = 19$, one-way ANOVA with post-hoc pairwise comparisons with Bonferroni correction; DG suprapyr.: $F_{(2,\ 61)} = 6.443$, $p<0.01$, $n_{HC} = 22$, $n_{C-} = 23$, $n_{C+} = 19$, one-way ANOVA with post-hoc pairwise comparisons with Bonferroni's correction, $^*P<0.05$).
(D, F, H and J) Analysis of mean cFOS signal intensities in indicated regions. No context- or threat-dependent effects on cFOS signal intensities were found for CA1 or in the DG. In CA3, signal intensities differed significantly between the experimental groups (CA1: $F_{(2,\ 64)} = 1.280$, $p = 0.285$, $n_{HC} = 23$, $n_{C-} = 24$, $n_{C+} = 20$, one-way ANOVA; CA3: $F_{(2,\ 62} = 13.719$, $p<0.0001$, $n_{HC} = 21$, $n_{C-} = 24$, $n_{C+} = 21$, one-way ANOVA with post-hoc pairwise comparisons with Bonferroni's correction, $^*P<0.05$; DG infrapyr.: $F_{(2,\ 61)} = 0.595$, $p = 0.555$, $n_{HC} = 22$, $n_{C-} = 23$, $n_{C+} = 19$, one-way ANOVA; DG suprapyr.: $F_{(2,\ 61)} = 4.520$, $p<0.05$, $n_{HC} = 22$, $n_{C-} = 23$, $n_{C+} = 19$, one-way ANOVA with post-hoc pairwise comparisons with Bonferroni's correction, $^*P<0.05$).

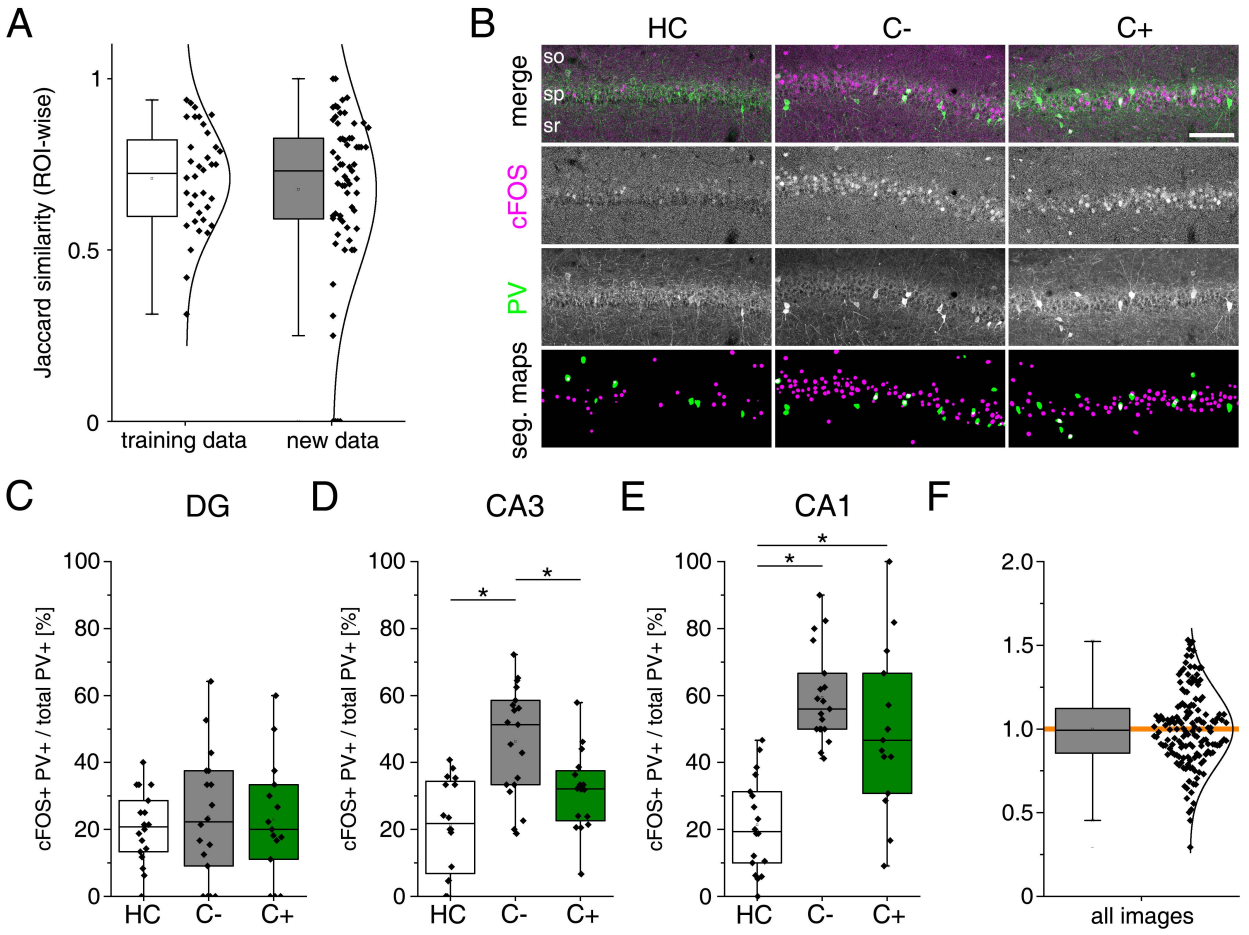

Figure 6: **Two different DeepFLaSH-models are combined to analyze activity-related changes in Parvalbumin-positive interneurons**

(A) Comparison of ROI-wise Jaccard-similarities between the CNN for PV+ somata and Expert 3 across the training data and new images. The PV-model, which was trained in an inter-coding approach, maintains expert-like performance on new images ($p > 0.05$, Mann-Whitney test, $n_{training\ data} = 36$, $n_{new\ data} = 65$).

(B) Immunolabeled PV and cFOS in CA1 under three experimental conditions (HC, C-, and C+). Green: PV, magenta: cFOS, so: stratum oriens, sp: stratum pyramidale, sr: stratum radiatum. Scale bar: 100 µm. The segmentation maps are based on CNN predictions.

(C-E) Proportion of cFOS-positive PV+ interneurons of all PV+ neurons (DG $F_{(2,\ 48)} = 0.180$, $p = 0.836$, $n_{HC} = 18$, $n_{C-} = 18$, $n_{C+} = 15$, one-way ANOVA; CA3 $F_{(2,\ 48)} = 13.128$, $p < 0.0001$, $n_{HC} = 16$, $n_{C-} = 19$, $n_{C+} = 16$, one-way ANOVA with post-hoc pairwise comparisons with Bonferroni's correction, *$P < 0.05$); CA1 $F_{(2,\ 49)} = 23.128$, $p < 0.0001$, $n_{HC} = 18$, $n_{C-} = 19$, $n_{C+} = 15$, one-way ANOVA with post-hoc pairwise comparisons with Bonferroni's correction, *$P < 0.05$).

Image-wise ratio of mean PV fluorescent signal intensity of cFOS-positive PV+ somata compared to the mean signal intensity of cFOS-negative PV+ somata. Both types of cells show similar mean intensity in the PV fluorescent signal, thus indicating the same amount of PV in both cell type somata ($p > 0.05$, one-sample t-test, $n = 152$).

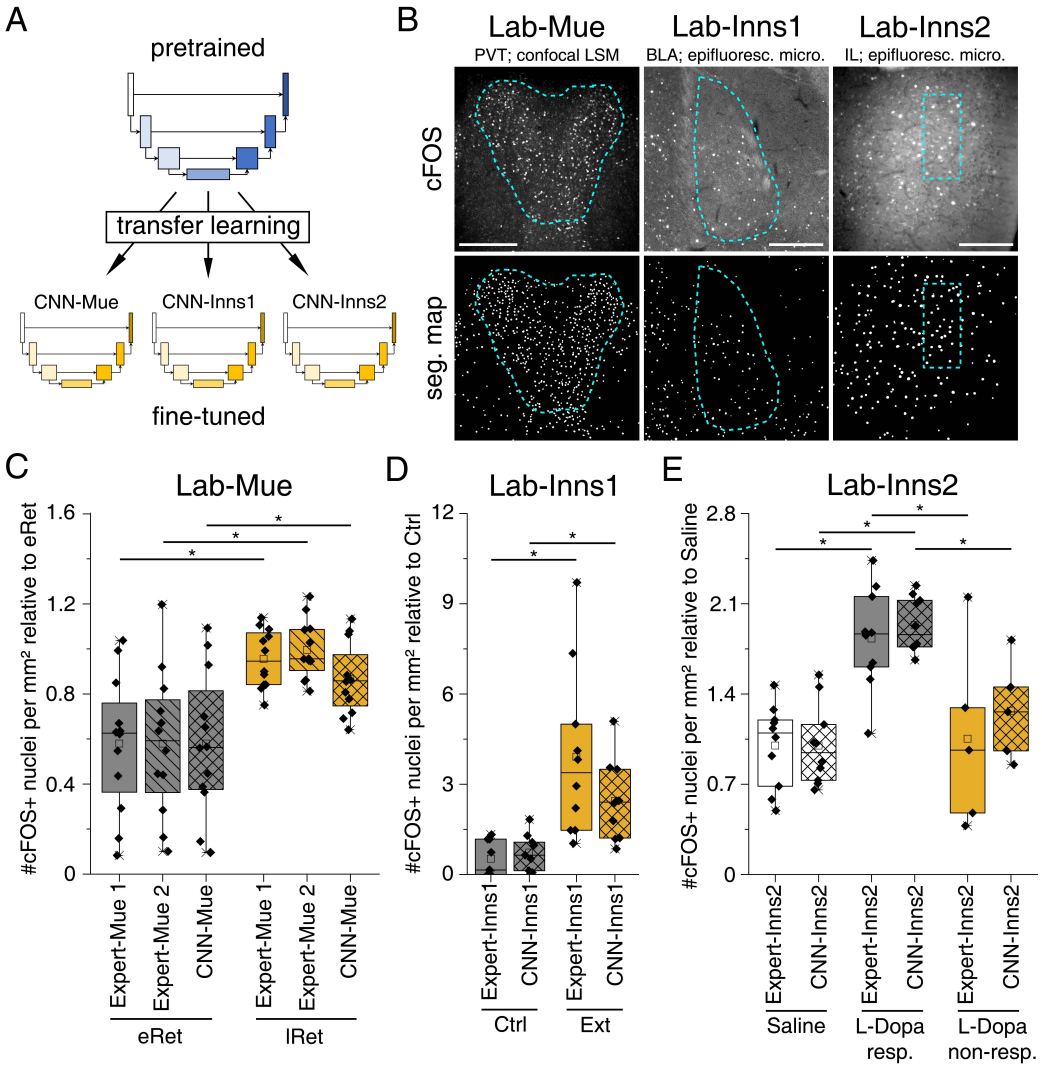

Figure 7: **Transfer learning of pretrained CNN-models to independent datasets with DeepFLaSH**
(A) Transfer learning adapts pretrained-CNN-models to independent datasets. Here, five pairs of microscopy images and manually prepared segmentation maps were sufficient to fine-tune the DeepFLaSH pretrained cFOS-model to the data of three individual laboratories (Lab-Mue, Lab-Inns1, and Lab-Inns2).
(B) Representative images show the lab-specific microscopy data for cFOS labels. The investigated brain region and microscopy technique that was used to acquire the raw image data are indicated. Dashed lines indicate the analyzed regions. The segmentation maps are based on the prediction of the respective CNNs. PVT: para-ventricular nucleus of thalamus, BLA: basolateral amygdala, IL: infralimbic cortex; LSM: laser-scanning microscopy, epifluoresc. micro.: epifluorescence microscopy. Scale bars: Lab-Mue 400 µm, Lab-Inns1 300 µm, Lab-Inns2 150 µm.
(C-E) Comparison of lab-specific experts and the corresponding fine-tuned CNN-models.
(C) Lab-Mue. Two human experts and the transfer-trained CNN-Mue model detect an equal increase of cFOS+ nuclei in the PVT 24 hours after fear conditioning (Expert-Mue1: *$p < 0.05$, Welchs two-sample t-test; Expert-Mue2: *$p < 0.01$, Welchs two-sampe t-test; CNN-Mue: *$p < 0.05$, Welchs two-sample t-test; for all: $N_{eRet} = 4$, $N_{lRet} = 4$, $n_{eRet} = 12$, $n_{lRet} = 12$).
(D) Lab-Inns1. A significant increase of cFOS+ nuclei in the BLA was found by a human expert as well as the fine-tuned CNN-Inns1 model (Expert-Inns1: *$p < 0.001$, Mann-Whitney test; CNN-Inns1: *$p < 0.01$, Welchs two-sample t-test; for both: $N_{Ctrl} = 5$, $N_{Ext} = 5$, $n_{Ctrl} = 9$, $n_{Ext} = 10$).
(E) Lab-Inns2. Only mice that responded to the treatment with L-DOPA showed a significant increase in cFOS+ nuclei in the infralimbic cortex, as detected by a human expert and CNN-Inns2 model (Expert-Inns2: $F_{(2,\ 22)} = 10.045$, $p < 0.001$, one-way ANOVA with post-hoc pairwise comparisons with Bonferroni's correction, *$P < 0.05$; CNN-Inns2: $F_{(2,\ 22)} = 26.849$, $p < 0.0001$, one-way ANOVA with post-hoc pairwise comparisons with Bonferroni's correction, *$P < 0.05$; for both: $N_{Saline} = 6$, $N_{L\text{–}DOPA\ resp.} = 6$, $N_{L\text{–}DOPA\ non\text{–}resp.} = 3$, $n_{Saline} = 10$, $n_{L\text{–}DOPA\ resp.} = 10$, $n_{L\text{–}DOPA\ non\text{–}resp.} = 5$).

Material and methods

Data sets regarding animal behavior, immunofluorescence analysis and image acquisition were performed in four independent laboratories using lab-specific protocols. Experiments were not planned together to ensure the individual character of the datasets. We refer to the lab-specific protocols as follows:

- (*Lab-Wue*) Institute of Clinical Neurobiology, University Hospital, Wrzburg, Germany
- (*Lab-Mue*) Institute of Physiology I, University of Mnster, Germany
- (*Lab-Inns1*) Department of Pharmacology, Medical University of Innsbruck, Austria
- (*Lab-Inns2*) Department of Pharmacology and Toxicology, Institute of Pharmacy and Center for Molecular Biosciences Innsbruck, University of Innsbruck

Contact for reagent and resource sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Robert Blum (`blum_r@UKW.de`). Requests regarding the machine learning model and infrastructure should be directed to Christoph M. Flath (`christoph.flath@uni-wuerzburg.de`).

Experimental models

Mice

(*Lab-Wue*) All experiments were in accordance with the Guidelines set by the European Union and approved by our institutional Animal Care, the Utilization Committee and the Regierung von Unterfranken, Wrzburg, Germany (License number: 55.2-2531.01-95/13). C57BL/6J wildtype mice were bred in the animal facility of the Institute of Clinical Neurobiology, University Hospital of Wrzburg, Germany. Mice were housed in groups of 3 to 5 animals under standard laboratory conditions (12h/12h light/dark cycle, food and water ad libitum). All mice were healthy and pathogen-free, with no obvious behavioral phenotypes. Mice were quarterly tested according to the Harlan 51M profile (Harlan Laboratories, Netherlands). Yearly pathogen-screening was performed according to the Harlan 52M profile. All behavioral experiments were performed with male mice at an age of 8-12 weeks during the subjective day-phase of the animals and were randomly allocated to experimental groups.
(*Lab-Mue*) Male C57Bl/6J mice (Charles River, Sulzfeld, Germany) were kept on a 12h-light-dark cycle and had access to food and water ad libitum. No more than five and no less than two mice were kept in a cage. Experimental animals of an age of 9 10 weeks were single housed for 1 week before the experiments started. All animal experiments were carried out in accordance with European regulations on animal experimentation and protocols were approved by the local authorities (Landesamt fr Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen).
(*Lab-Inns1*) Experiments were performed in adult, male C57Bl/6NCrl mice (Charles River, Sulzfeld, Germany) at least 10-12 weeks old, during the light phase of the light/dark cycle. They were bred in the Department of Pharmacology at the Medical University Innsbruck, Austria in Sealsafe IVC cages (1284L Eurostandard Type II L: 365 x 207 x 140 mm, floor area cm2 530, Tecniplast Deutschland GmbH, Hohenpeienberg, Germany). Mice were housed in groups of 3 to 5 animals under standard laboratory conditions (12h/12h light/dark cycle, lights on: 07:00, food and water ad libitum). All procedures involving animals and animal care were conducted in accordance with international laws and policies (Directive 2010/63/EU of the European parliament and of the council of 22 September 2010 on the protection of animals used for scientific purposes; Guide for the Care and Use of Laboratory Animals, U.S. National Research Council, 2011) and were approved by the Austrian Ministry of Science. All effort was taken to minimize the number of animals used and their suffering.
(*Lab-Inns2*) Male 3-month-old 129S1/SvImJ (S1) mice (Charles River, Sulzfeld, Germany) were housed

(four per cage) in a temperature (22±2 °C) and humidity (50-60 %) controlled vivarium with food and water ad libitum under a 12h light/dark cycle. All mice were healthy and pathogen-free, with no obvious behavioral phenotypes. The Austrian Animal Experimentation Ethics Board (Bundesministerium fr Wissenschaft Forschung und Wirtschaft, Kommission fr Tierversuchsangelegenheiten) approved all experimental procedures.

Method details

Mouse behavior

Contextual fear conditioning

(*Lab-Wue*) Male animals, initially kept as siblings in groups, were put to a new cage and kept in single-housing conditions with visual, olfactory and auditory contact in a ventilated cabinet (Scantainer, Scanbur). To habituate the mice to the male experimenter and the experimental rooms, mice were handled twice a day for at least two consecutive days prior to behavioral analysis. Mice were put to three different groups: (1.) the homecage group, (2.) the context control group that experienced the training context, but did not receive any electric foot shock, and (3.) the context-conditioned group, which received electric foot shocks in the training context. Contextual fear (threat) conditioning was performed in a square conditioning arena with a metal grid floor (Multi conditioning setup, 256060 series, TSE, Bad Homburg, Germany). Before each experiment, the arena was cleaned with 70% ethanol. Mice were transported in their homecage to the experimental rooms and were put into the conditioning arena. After an initial habituation phase of 60 s, fear acquisition was induced by five electric foot shocks (unconditioned stimulus, US; 1 s, 0.7 mA) with an inter-stimulus interval of 60 s. After the foot shock presentation, mice remained in the training context for 30 s before being returned to their homecages in their housing cabinet. For fear memory retrieval, 24 hours after the training session, the mice were re-exposed to the conditioning arena for 360 s, without any US presentation. Mice were again put back to their homecage for 90 min, before mice were anaesthetized and prepared for immunohistological analysis. Mouse behavior was videotaped. The videotapes were analyzed with the MCS FCS-SQ MED software (TSE, Bad Homburg, Germany). The software was also used to automatically track mice behavior and to quantify the freezing behavior during all sessions. Freezing was defined as a period of time of at least 2 s showing absence of visible movement except that required for respiration [47]. The percentage time spent freezing was calculated by dividing the amount of time spent in the training chamber.

Restraint stress and Pavlovian fear conditioning

(*Lab-Mue*) Animals were randomly assigned to 4 groups considering the following conditions; stress vs. control and early retrieval vs. late retrieval. Mice experienced restraint stress and a Pavlovian fear-conditioning paradigm as described earlier [48]. In brief, on day one, animals of the stress group underwent restraint stress for 2 h by using a perforated standard 50 ml falcon tube, allowing ventilation, but restricting movement. Animals of the control group remained in their homecages. On day 10, animals were adapted through two presentations of six CS^- (2.5kHz tone, 85dB, stimulus duration 10s, inter-stimulus interval 20s; inter-trial interval 6h). On the next day, fear conditioning was performed in two sessions of three randomly presented CS^+ (10kHz tone, 85dB, stimulus duration 10s, randomized inter-stimulus interval 10-30s; inter-session interval 6h), each of which was co-terminated with an unconditioned stimulus (scrambled foot shock of 0.4mA, duration 1s). Animals of the early retrieval group underwent a retrieval phase on the same day (day 11), 1 h after the last conditioning session, whereas animals of the late retrieval group underwent the retrieval phase on the next day (day 12), 24 h after the conditioning session. For fear memory retrieval, mice were transferred to a new context. After an initial habituation phase of 2 min, mice were exposed to 4 CS^- and 40 s later to 4 CS^+ presentations (stimulus duration 10s, inter-stimulus interval 20s) without receiving foot shocks. Afterwards, mice remained in

this context for another 2 min before being returned to their homecages.

Contextual fear conditioning and extinction

(*Lab-Inns1*) Mice were single housed and stored in the experimental rooms in cages covered by filter tops with food and water *ad libitum* 3 days before behavioral testing. Fear acquisition and fear extinction were performed in a fear conditioning arena consisting of a transparent acrylic rodent conditioning chamber with a metal grid floor (Ugo Basile, Italy). Illumination was 80 lux and the chambers were cleaned with 70% ethanol. On acquisition day, following a habituation period of 120 s, mice were fear conditioned to the context by delivery of 5 foot-shocks (unconditioned stimulus, US, 0.5 mA for 2 s) with a random inter-trial interval of 70-100 s. After the test, mice remained in the test apparatus for an additional 120 s and were then returned to their homecage. On the next day, fear extinction training was performed. For this, mice were placed into the same arena as during acquisition and left undisturbed for 25 min. Freezing behavior was recorded and quantified by a pixel-based analysis software in one min bins (AnyMaze, Stoelting, USA). 90 min after the end of the extinction training, the mice were killed and the brains were processed for immunohistochemistry. Mice for homecage condition were kept in the experimental rooms for the same time period.

(*Lab-Inns2*) Fear conditioning, extinction and extinction retrieval was carried out as previously described [44]. Context dependence of fear extinction memories was assessed using a fear renewal tests in a novel context [49]. Fear conditioning and control of stimulus presentation occurred in a TSE operant system (TSE, Bad Homburg, Germany). Mice were conditioned in a 25 x 25 x 35 cm chamber with transparent walls and a metal-rod floor, cleaned with water, and illuminated to 300 lux (context A). The mice were allowed to acclimatize for 120 s before receiving three pairings of a 30 s, 75 dB 10 kHz sine tone conditioned stimulus (CS) and a 2 s scrambled-foot-shock unconditioned stimulus (US; 0.6 mA), with a 120 s inter-pairing interval. After the final pairing, mice were given a 120 s stimulus-free consolidation period before they were returned to the homecage. Fear extinction training was performed in context B, a 25 x 25 x 35 cm cage with a solid grey floor and black walls, cleaned with a 100% ethanol and illuminated to 10 lux with a red lamp. After a 120 s acclimation period, the mice were subjected to 16x CS-alone trials, separated by 5 s inter-CS intervals. Extinction retrieval was conducted in context B by repeating the conditions used in extinction training procedure but presenting only two CS trials. Fear renewal in a novel context was quantified 11 days following the extinction-retrieval test in a novel context (context C), a round plexiglas cylinder of 20 cm in diameter, and a height of 35 cm. The cylinder was covered on the outside with red diamond-printed white paper with an uneven pale ceramic tiled floor, illuminated to 5 lux with white light. After the mice were acclimated for 120 s, they were given two CS-alone trials, with a 5 s inter-CS interval. A trained observer blind to the animals grouping measured freezing, defined as showing no visible movement except that required for respiration, as an index of fear [50]. The observer manually scored freezing based on video recordings throughout the CS and determined the duration of freezing within the CS per total time of the CS in percent. Freezing during all phases was averaged over two CS presentations and presented in eight trial blocks during extinction training and a one trail block each for extinction retrieval and fear renewal. Freezing during fear conditioning was quantified and presented as single CS.

Immunohistochemistry and microscopy

(*Lab-Wue*) To analyze anti-cFOS and anti-Parvalbumin immunoreactivity after retrieval of a contextual memory, mice were anesthetized 90 minutes after the end of the retrieval session (C+). Mice that spent the same time in the conditioning arena without presentation of the US served as context controls (C-). Single-housed mice that were never exposed to the conditioning arena served as nave learning controls (homecage; HC).
A rodent anesthesia setup (Harvard Apparatus) was used to quickly anesthetize the mice with the volatile narcotic isoflurane (airflow 0.4 L/min, 4% isoflurane, Iso-Vet, Chanelle) for one minute. Then a

mixture of ketamine (120 mg/kg; Ketavet, Pfizer) and xylazine (16 mg/kg; cp-Pharma, Xylavet, Burgdorf, Germany) was injected (12 µml/g bodyweight, intraperitoneal) to provide sedation and analgesia. Then anesthetized mice were transcardially perfused (gravity perfusion) with 0.4% heparin (Ratiopharm) in phosphate-buffered saline (PBS), followed by fixation with 4% paraformaldehyde in PBS. Brains were dissected and post-fixed in 4% paraformaldehyde for two hours at 4°CThe tissue was embedded in 6% agarose and coronal sections (40 µm) were cut using a vibratome (Leica VT1200). A total of 30 sections starting from Bregma -1.22 mm [51] were considered as dorsal hippocampus. Immunofluorescent labeling was performed in 24-well plates with up to three sections per well under constant shaking. Slices were first incubated in 100 mM glycine, buffered at pH 7.4 with 2 M Tris-base for 1 h at room temperature. Slices were transferred in blocking solution consisting of 10% normal horse serum, 0.3% Triton X100, 0.1% Tween 20 in PBS for 1 h at room temperature. Primary antibodies were applied in blocking solution for 48 h at 4°C. The following primary antibodies were used at indicated dilutions: mouse anti-Parvalbumin, SWANT, PV235, 1:5,000; guinea-pig anti-NeuN, SynapticSystems, 266004, 1:400; rabbit anti-cFOS, SynapticSystems, 226003, 1:10,000 (lot# 226003/7). Secondary antibodies were used for 1.5 h at room temperature at a concentration of 0.5 µmg / ml in blocking solution. The following antibodies were used: goat anti-mouse Alexa-488 conjugated (Life sciences, Thermo); donkey anti-rabbit Cy3 conjugated (Jackson ImmunoResearch) and donkey anti-guinea-pig Cy5 conjugated (Jackson ImmunoResearch). Sections were embedded in Aqua-Poly/Mount (Polysciences). Confocal image acquisition was performed with an Olympus IX81 microscope combined with an Olympus FV1000 confocal laser scanning microscope, a FVD10 SPD spectral detector and diode lasers of 473, 559 and 635 nm. Image acquisition was performed using an Olympus UPlan SAPO 20x/0.75 objective. Images with 1024 pixel to monitor 636 μm^2 were taken as 12 bit z-stacks with a step-size of 1.5 µm, covering the whole section. Images of *dentate gyrus* (DG), *Cornu ammonis 1* (CA1) and CA3 were taken in each hemisphere of three sections of the dorsal hippocampus to achieve a maximum of six images (n) per region for each animal (N). During image acquisition, the experimenter was blinded to the treatment condition (C+ versus C- versus HC).

(*Lab-Mue*) Mice were anesthetized via inhalation anesthesia (isoflurane, 5% in O2; CP Pharma, Germany) and perfused with phosphate-buffered saline (PBS) and then 4% paraformaldehyde (PFA; Roti-Histofix 4%, Carl Roth). Brains were isolated and post-fixed overnight in 4% PFA, treated with 30% sucrose/PBS solution for 48 h, and then stored at 4°Cuntil sectioning. Coronal sections (40 µm thick) were prepared on a freezing microtome (Leica, Wetzlar, Germany) and stored in PBS until use. Immunostaining was performed on free-floating sections. Sections were washed 3 x 10 min with PBS and then incubated in blocking solution (10% goat serum, 3% BSA, 0.3% Triton X100 in PBS) for 1 h. After blocking, sections were treated at 4°Covernight with a primary antibody (rabbit anti-cFOS, 1:500, Santa Cruz Biotechnology, California, USA) diluted in blocking solution. On the next day, sections were washed 3 x 10 min with PBS and incubated for 1 h at room temperature with the secondary antibody (goat anti-rabbit Alexa Fluor 488, 1:1000; Invitrogen, Germany) diluted in blocking solution. The incubation was followed by three 5 min washing steps in PBS. Sections were then mounted on SuperFrostPlus slides (Menzel, Braunschweig, Germany) and embedded with Vectashield Mounting Medium (Vector Laboratories, Burlingame, California) + 4,6-diamidino- 2-phenylindole (DAPI). Fluorescence labeling was visualized and photographed using a laser-scanning confocal microscope (Nikon eC1 plus) with a 16x water objective at a step size of 1.5 µm, covering the whole section. Identical exposure settings were used for images that show the same region in the brains. The experimenter was blinded to the treatment conditions.

(*Lab-Inns1*) Ninety minutes after extinction training, mice were injected intraperitoneally with thiopental (150 mg / kg, i.p., Sandoz, Austria) for deep anesthesia. Transaortal perfusion, 3 min with PBS at room temperature followed by 10 min of 4% PFA at 4°C, was performed by a peristaltic pump at a flow rate of 9 ml / min (Ismatec, IPC, Cole-Parmer GmbH, Wertheim, Germany). Subsequently, brains were removed and postfixed in 4% PFA for 90 min at 4°C, cryoprotected for 48 h in 20% sucrose at 4°Cand then snap frozen in isopentane (2-methylbutane, Merck GmbH, Austria) for 3 min at -60°C. Brains were transferred to pre-cooled open tubes and stored at -70°Cuntil further use. For immunohistochemistry, coronal 40 µm sections were cut by a cryostat from rostral to caudal, collected in Tris-buffered saline (TBS) +

0.1% sodium azide. Sections from Bregma -1.22 mm [51] were incubated for 30 min in TBS-triton (0.4%), for 90 min in 10% normal goat/horse serum and overnight with the first primary antibody (diluted in 10% serum containing 0.1% sodium azide). Rabbit anti-cFOS (Millipore, PC-38, 1:20,000) and mouse anti-Parvalbumin (Sigma-Aldrich, P3088, 1:2,500) were used as primary antibodies. After washing with TBS-buffer 3 x 5 min, secondary antibodies (goat anti-rabbit, Vector Laboratories inc., PI-1000, 1:1,000 and biotinylated horse anti-mouse, Vector Laboratories inc., PK-4002, 1:200) were added to the sections for 150 min. Then, sections were incubated in the dark for 8 min in TSA-fluorescein (in-house, 1:100) staining solution (50 mM PBS and 0.02% H2O2). Sections were rinsed 3 x 5 min in TBS buffer and then incubated for 100 min in a solution of streptavidin Dylight 649 (Vector laboratories, SA5649, 1:100) in TBS buffer. Fluorescently stained sections were mounted on slides using gelatin and cover-slipped with glycerol-DABCO anti-fading mounting medium. Photomicrographs were taken on a fluorescent microscope (Zeiss Axio Imager M1) equipped with a halogen light source, respective filter sets and a monochrome camera (Hamamatsu ORCA ER C4742-80-12AG). Images of the basolateral amygdala (BLA) were taken with an EC Plan-Neofluar 10x/0.3 objective. All images were acquired using the same exposure time and software settings and the experimenter was blinded to the treatment conditions (homecage vs extinction).

(*Lab-Inns2*) Mice were killed 2 h after the start of the fear renewal session using an overdose of sodium pentobarbital (200 mg/kg) and transcardially perfused with 40 ml of 0.9% saline followed by 40 ml of 4% paraformaldehyde in 0.1 M phosphate buffer, pH 7.4. Brains were then removed and post fixed at 4°Cfor 2 h in 4% paraformaldehyde in phosphate buffer. Brains were sectioned at the coronal plane with a thickness of 40 µm on a vibratome (VT1000S, Leica). Free-floating sections were incubated for 30 min in blocking solution using 1% BSA in 50 mM Tris buffer (pH 7.4) with 0.1% Triton-X100 and incubated overnight at 4°Cwith a rabbit antibody against cFos (1:1000; sc-52, Szabo-Scandic, Vienna, Austria). The sections were then washed (3 x 15 min in 1% BSA in Tris buffer containing 0.1% Triton-X100) and incubated for 2.5 h with a secondary CY2-conjugated donkey anti rabbit IgG (1:500, #82371, Jackson ImmunoResearch). The sections were then washed (3 x 15 min in 50 mM Tris buffer), mounted on microscope slides and air-dried. Slides were embedded in ProLong Gold anti-fade reagent containing DAPI (P36935, Life Technologies). Immunofluorescence was assessed using a fluorescent microscope (Olympus BX51 microscope, Olympus XM10 video camera, CellSens Dimension 1.5 software, Olympus). Immunolabelled sections were visualized using a 20xoil-objective (UPlanSApo, Olympus) at 488nm excitation.

Manual image processing

(*Lab-Wue*) For image preprocessing, 12-bit grey images were projected (maximum intensity) and converted to 8-bit. The respective regions-of-interest (ROIs) of either the NeuN-positive area, PV-positive somata or cFOS-positive nuclei were segmented manually using ImageJ [7].The following structures were marked as ROI: cFOS-positive nuclei, PV-positive somata, and NeuN immunoreactive layers of the dentate gyrus (granule cell layer), CA1 and CA3. All NeuN-positive areas used for the quantifications of cFOS-positive cells were manually segmented. All human experts were blinded to another and the treatment condition.

(*Lab-Mue*) Images were adjusted in brightness and contrast. The respective regions-of-interest (ROIs) of the cFOS-positive cells in paraventricular thalamus (PVT) were segmented manually using ImageJ. Two independent neuroscientists analyzed these images manually for cFOS-positive cells. Both experts were blinded to the treatment condition.

(*Lab-Inns1*) Number of cFOS-positive neurons was obtained from two basolateral amygdalae (BLA) per animal of five homecage mice and five mice subjected to contextual fear extinction. PV staining was used to identify the localization and extension of the BLA and the borders were manually drawn using the free shape tool of the Improvision Openlab software (PerkinElmer). Boundaries were projected to the

respective cFOS-immunoreactive image and cFOS positive neurons were manually counted inside that area. The experimenters were blinded to the treatment condition.

(*Lab-Inns2*) The anatomical localization of cells within the infralimbic cortex was aided by using illustrations in a stereotaxic atlas [52], published anatomical studies [53] and former studies in S1 mice [54, 55]. All analyses were done in a comparable area under similar optical and light conditions. Images were digitized and viewed on a computer screen using CellSens Dimension 1.5 software (Olympus Corporation, Tokyo, Japan). cFOS positive cells were evaluated within the infralimbic cortex, the brain region of the interest. The experimenter was blinded to the treatment conditions.

Image segmentation via thresholding

(*Lab-Wue*) We used global thresholding of bit-values to divide the pixels of an image into two classes pixels belonging to background and pixels belonging to foreground to create a binary image. For segmentation, the threshold plugin of ImageJ was used. For cFOS-positive nuclei, the threshold isolated the two percent highest bit values (pixel-wise) as foreground. For the segmentation of PV-positive somata, thresholding isolated one percent of the brightest pixels. Subsequently, we used the Particle Analyzer plugin of ImageJ to create segmentation masks. The settings for the particle analyzer were derived from the values we observed in the manual analysis by human experts.

Deep learning approach

Inputs and outputs

Our machine learning model is a deep neural network which takes microscopy images as an input and outputs a segmentation map (mask). For each mask, the network outputs for each pixel a probability distribution of belonging to the positive class (i.e., fluorescent label). The input to the network is a 1024x1024 grayscale image (one channel) and we use a batch size of 4 for training. Thus, the input is a tensor of shape 4x1024x1024x1 of type float32 where the axes represent batch x row x column x channel. The output tensor of the network has a 4x1024x1024x1 shape of type float32 where the axes are batch x row x column x pixel-probability.

Architecture

The design of the deep neural network is inspired by the U-net architecture [22]. U-net like architectures are, at the current state, the most common architecture for biomedical image segmentation. Moreover, they have demonstrated impressive performance in relevant data-science competitions (Kaggle Data Science Bowl: `https://www.kaggle.com/c/data-science-bowl-2018/discussion/54741`). The architecture proposed by Ronneberger et al. [22] is designed to process smaller resolution images (512 x 512 pixel). Therefore, we modified the neural network to enhance the performance on 1024 x 1024 dimensional images. To this end, we increase the depth of the U-net from five to eight modules (Figure S.1). Furthermore, in order to benefit from current research and findings in the quickly emerging field of deep learning, we incorporated two novel components in the modules. Batch Normalization Layers [24] and depthwise separable convolutions [25] instead of standard 2D-convolutional layers on the down part of the network. These changes significantly reduce trainable parameters (i.e., weights) and improved training speed while the level of performance is maintained.
Furthermore, the modular architecture enables a quick adjustment to support applications with different image dimensions.

Data augmentation

Given limited training data availability, we leverage data augmentation by applying elastic deformations

to the available training images. This allows the network to adopt robustness to such deformations. For data augmentation both the original microscopy image and the segmentation maps are transformed using the Keras ImageDataGenerator [25] with the following parameters: degree range for random rotations: 90, width and height shift range: 0.1, shear angle in counter-clockwise direction in degrees: 0.2, randomly flip inputs horizontally and vertically, constant fill mode: 0 (black).

Training loss

The network is trained using a combination of two loss functions. On the one hand, we use the weighted cross entropy loss [22]), which provides a probabilistic measure of the similarity between the prediction and the ground truth and tends toward zero as the neural network gets better at computing the desired output.
On the other hand, we include the weighted dice overlap coefficient loss to address class-imbalance issue [56]. This loss function is based on the Dice coefficient, one of the most common measures of region overlap in medical image analysis [57].

Implementation

We implemented the network in Keras [25], a popular high-level open-source API for deep learning, with TensorFlow [26] in the backend. It was trained using the Adam optimizer [27], a commonly used gradient-based function optimizer included in TensorFlow.

Training procedure

For the training phase, weights are updated using adam optimizer [27] with a learning rate of 0.001. Training is completed after 100 epochs. We use a batch size of 4.

Postprocessing

In order to derive binary segmentation maps from the probabilistic output of the CNN, we considered each pixel with a positive class probability higher than 94% as a potential part of a fluorescent feature (ROIs). Then, we applied the Particle Analyzer of ImageJ to detect areas that correspond to fluorescent features.

Transfer learning

In order to adopt the neural network to different imaging conditions (for instance, different microscopes in different laboratories, see Manual image processing) we used the concept of transfer learning [17].
While the neural network pretrained on the data of *Lab-Wue* yields reasonable results on the datasets of the other labs, the performance improved significantly using only five new sample images and manual segmentation maps of the new imaging conditions. Here, we train the network for another 50 rounds using the same training procedure and data augmentation techniques as before. Due to the layer connection characteristics of the U-net we chose not to freeze any layers for training.

Inter-coding

To train and test the convolutional neural network we used a training dataset of 36 images that contained equal amounts of images of the different regions (DG, CA1, CA3; 12 images each) and of the different experimental conditions (HC, C- and C+; 12 images per condition, 4 images per region). Five independent neuroscientists analyzed these 36 images manually for the NeuN-positive area, PV-positive somata and cFOS-positive nuclei.

Expert consensus and error maps

As a reference, we computed an expert-consensus. The expert consensus represents all pixel information that was annotated by at least three of five human neurobiology experts. To visualize the spatial accuracy of all coders (experts, CNN and threshold), we created error maps using the expert-consensus as reference. We overlaid the expert-consensus with the segmentation map of the respective coder and computed all pixels that were annotated in the reference but were absent in the coders segmentation area and vice versa.

Quantification and statistical analysis

Statistical analysis

All statistical analyses were performed using OriginPro 2018G. Grubb's test was used to test for outliers ($p<0.05$). Normality (Shapiro-Wilk) and equality of variances (Levene's) were tested and parametric or non-parametric tests were used accordingly, as reported in the figure legends (parametric: one-way ANOVA; non-parametric: Kruskal-Wallis-ANOVA, followed by Mann-Whitney tests with Bonferroni correction for multiple comparisons). Throughout all analyses, N represents the number of animals and n the number of analyzed images. In boxplots, the area of the box represents the interquartile range (IQR, 1^{st} to 3^{rd} quartile) and whiskers extend to the maximal or minimal values, but no longer than 1.5 IQR. Normal distribution curves are scaled to 115% of the maximum.

Jaccard similarity coefficient

To evaluate the similarity between two segmentation maps we exploit the Jaccard similarity coefficient (also known as Jaccard Index or Intersection over the Union), which is a widely used similarity measure for biomedical images [58, 59, 60]. We compute the Jaccard similarity between two objects A and B as

$$J(A,B) = \frac{|A n \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|} \tag{1}$$

where $|A n \mathrm{B}|$ represents the intersection and $|\mathrm{A} \cup \mathrm{B}|$ the union of A and B. We apply the Jaccard similarity measure for *ROI matching* and *coder segmentation comparison*.
To address the issue that segmentation maps of the coders or the CNN often differ on a pixel basis even if the same ROIs are labelled, we compute the Jaccard similarity of all possible ROI pairs between two segmentation maps. Similarly to [61] we consider a tuple (A, B) of images with a single ROI as a match if they satisfy the condition:

$$J(A,B) > 0.5 \tag{2}$$

For the coder segmentation comparison, we compute the Jaccard Similarity between two segmentation maps, where $|A n \mathrm{B}|$ denotes the number of matching ROIs and $|\mathrm{A} \cup \mathrm{B}|$ the union of all ROIs in both images. As the CNN is trained to maximize the similarity on pixel level (not ROI level), we assure the consistency of our approach by also computing the Jaccard Similarity between two segmentation maps on pixel level. Here, $|A n \mathrm{B}|$ denotes the number of matching positive, i.e., white, pixels and $|\mathrm{A} \cup \mathrm{B}|$ the union of all positive pixels in both images. As Figure 3B and Figure S2 yield similar results we surmise a strong relationship between both approaches, while the first approach is more suitable for biological interpretable purposes.

Quantification of cFOS-positive cells

In order to compare the number of cFOS-positive cells across images, we normalized in each image the number of cFOS-positive cells to the area of the analyzed region (e.g. NeuN-positive area for *Lab-Wue*).

For one set of experiment, we pooled this data for each condition (e.g. HC, C- and C+ for *Lab-Wue*) and the analyzed brain region (e.g. DG, CA3 and CA1 for *Lab-Wue*). To compare different sets of experiments with each other, we normalized all relative cFOS-positive cell counts to the mean value of the respective control (e.g. HC for *Lab-Wue*).

Quantification of PV-positive interneurons

To test for differences in overall numbers of Parvalbumin-positive (PV-positive) interneurons and their staining intensities, we quantified the number of PV-positive interneurons and their mean staining intensity per image and pooled this data for each condition and the analyzed hippocampal region. To compare the mean staining intensities across experimental sets, we normalized all mean staining intensities to the mean value of the respective homecage control group. In DeepFLaSH analysis, a PV-positive interneuron was defined as cFOS-positive, when the predicted PV-ROI contained a predicted cFOS-positive ROI. The ratio of cFOS-positive PV-positive interneurons (cFOS+ PV+ INs) to the total number of PV-positive interneurons was calculated for each image and pooled according to experimental condition and analyzed hippocampal region.

Data and software availability

DeepFLaSH is a pipeline for label segmentation. The code for running the image segmentation is on GitHub at `https://github.com/matjesg/DeepFlaSH`. It includes links to pretrained network parameters. Users without any programming experience can follow the README to run an iPython Notebook on Google Colaboratory to create segmentation maps from their own microscopy images. Moreover, it allows the adoption of pretrained models for individual demands.

Resource Table

Reagent or resource	Source	Identifier
Antibodies		
(*Lab-Wue*) guinea-pig anti-NeuN	Synaptic systems	Cat# 26604, RRID: AB_2619988
(*Lab-Wue*) mouse anti-Parvalbumin	Swant	Cat# PV235, RRID: AB_10000343
(*Lab-Inns1*) mouse anti-Parvalbumin	Sigma-Aldrich	Cat# P3088, RRID: AB_477329
(*Lab-Wue*) rabbit anti-cFOS	Synaptic systems	Cat# 226003, RRID: AB_2619946
(*Lab-Inns1*) rabbit anti-cFOS	Millipore	Cat# PC38, RRID: AB_2106755
(*Lab-Mue*, *Lab-Inns2*) rabbit anti-cFOS	Santa Cruz	Cat# sc-52, RRID: AB_2106783
Experimental models		
(*Lab-Wue*, *Lab-Mue*) Mouse: C57BL/6J	Jackson Laboratory	Cat# JAX:000664, RRID: IMSR_JAX:000664
(*Lab-Inns1*) Mouse: C57BL/6N	Charles River	Cat# CRL:027, RRID: IMSR_CRL:27
(*Lab-Inns2*) Mouse: 129S1/SvImJ (S1)	Jackson Laboratory	Cat# 002448, RRID: MGI:5658424
(*all labs*) Postnatal brain sections from adult mice as described	This paper	N/A
Software and algorithms		
(*Lab-Wue*, *Lab-Mue*, *Lab-Inns2*) Fiji (ImageJ)	Fiji	`https://fiji.sc/`, RRID: SCR_002285
(*Lab-Wue*) Fluoview FV10-ASW	Olympus	`https://www.photonics.com/Product.aspx?PRID=47380`, RRID: SCR_014215
(*Lab-Inns1*) Improvision Openlab software (5.5.0)	Perkin Elmer	`http://www.perkinelmer.com/pages/020/cellularimaging/products/openlab.xhtml`, RRID: SCR_012158
(*Lab-Inns2*) CellSens Dimension Desktop 1.9 software	Olympus	`https://www.olympus-lifescience.com/en/software/cellsens/`, RRID: SCR_016238
(*Lab-Inns1*) Prism 7.0	Graphpad	`https://www.graphpad.com/scientific-software/prism/`, RRID: SCR_015807
TensorFlow	Abadi et al. [26]	`https://www.tensorflow.org`, RRID: SCR_016345
Keras	Chollet [25]	`https://keras.io`
ImageJ	Rueden et al. [62]	`https://imagej.net/`, RRID: SCR_003070
SciPy	Jones et al. [63]	`https://www.scipy.org`, RRID: SCR_008058
Code and data for this paper	This paper	`https://github.com/matjesg/DeepFlaSH`

Supplementary figures:

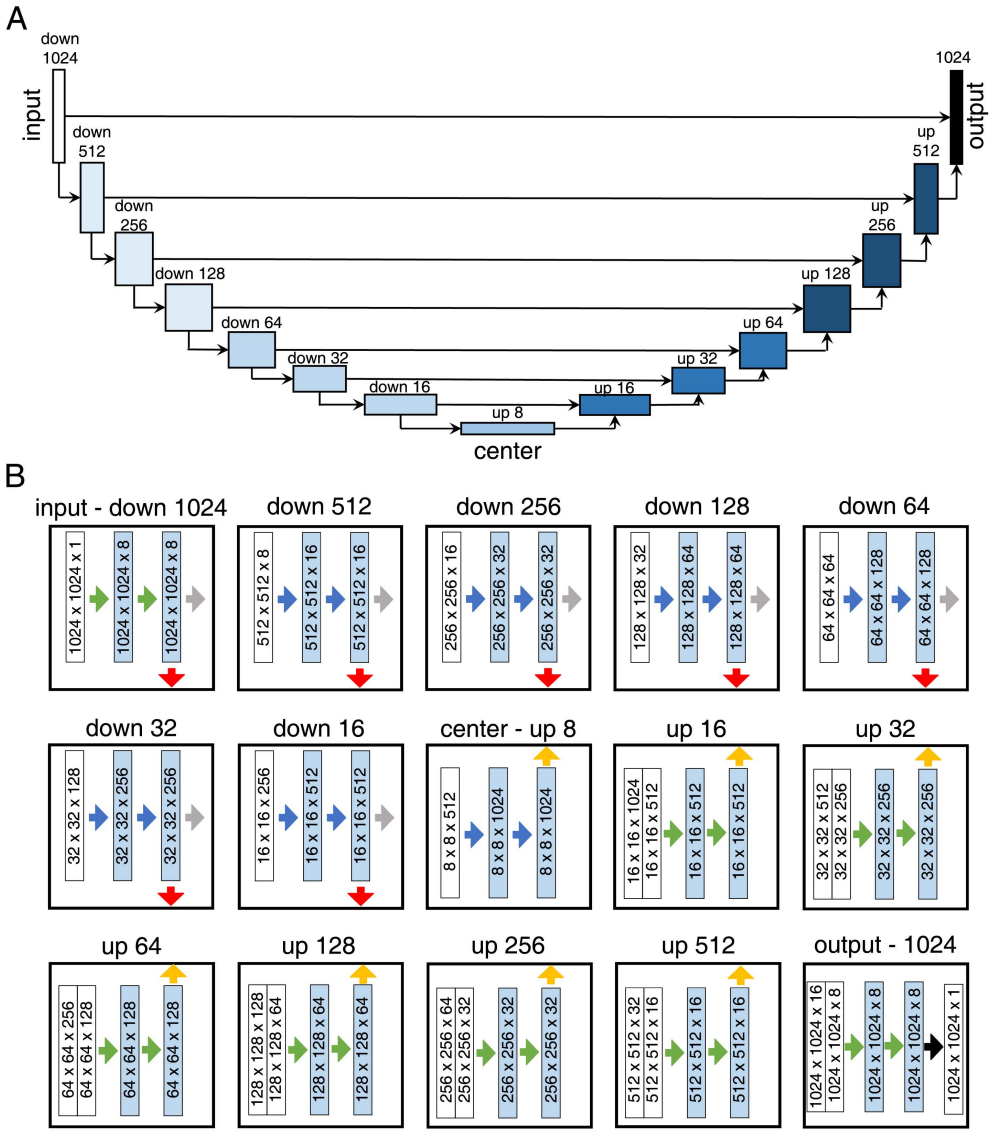

Figure S.1: **Schematic overview of the algorithm architecture**
(A) A convolutional neural network with a U-net architecture [22] comprising 15 modules. It takes a 1024x1024 pixel greyscale image as an input and outputs a probability segmentation map of the same size. Each box corresponds to a module. The name of the module indicates whether the feature maps are reduced (Down) or upsampled (Up). The number denotes the row and column dimensions of the tensors.
(B) A detailed illustration of the U-net modules. The white boxes correspond to the input multi-channel feature map of the module, the blue boxes to the multi-channel feature maps after applying a certain operation defined by the arrows. Arrows that point at the modules edges connect to the other modules as depicted in (A). The three dimensions of the feature map are row x column x feature-channel. In addition to Ronneberger et al. [22], we attach Batch Normalization Layers [24] and replace the standard 2D-Convolutional Layers with depthwise separable convolutions [25] on the down part of the network. These changes significantly reduce trainable parameters (i.e., weights) and improved training speed while the level of performance is maintained. Arrows indicate the following computations: green: 2D convolution 3x3, batch normalization, ReLu activation; grey: concatenate; red: 2D max pooling 2x2; blue: 2D separable convolution 3x3, batch normalization, ReLu activation; yellow: 2D upsampling; black: 2D convolution 1x1, sigmoid activation.

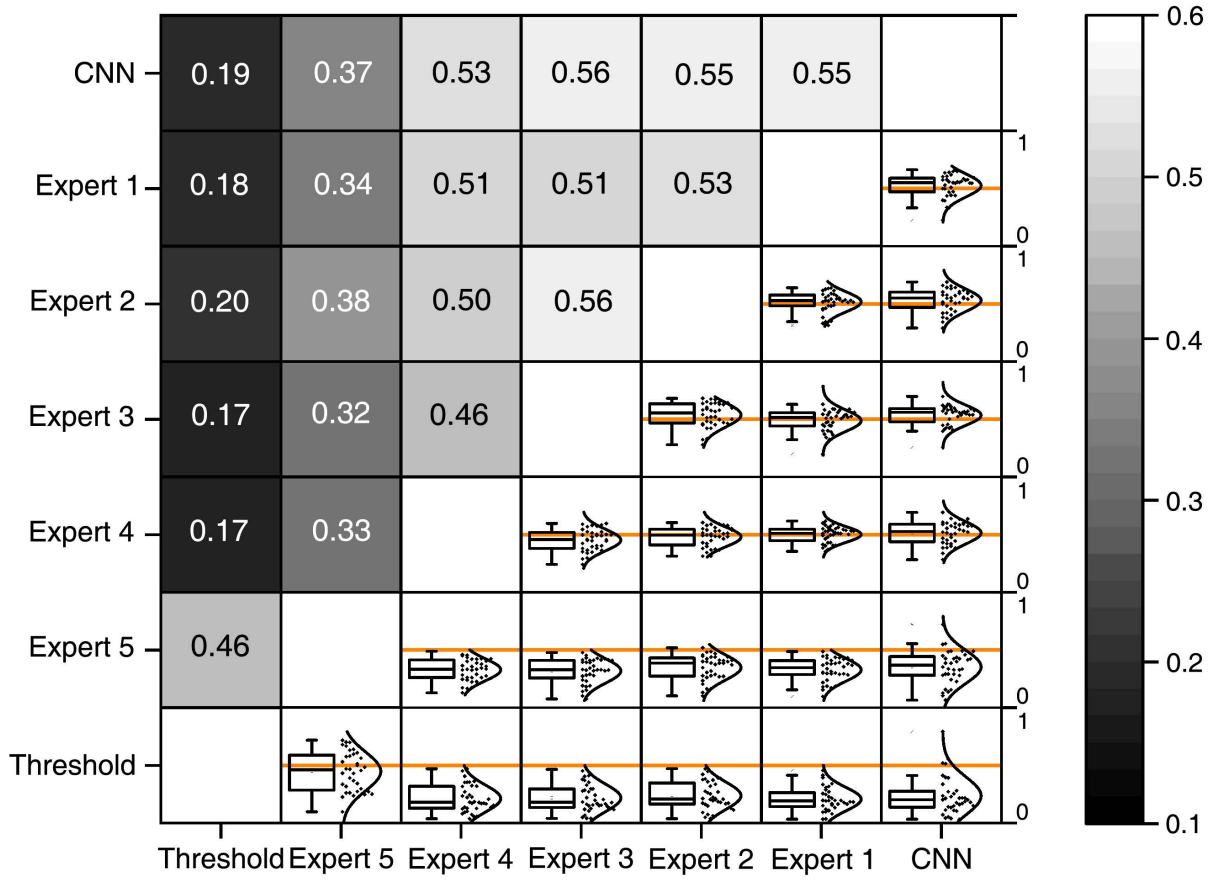

Figure S.2: **Automatic segmentation of cFOS immunolabels by a trained CNN-model shows expert-like performance also on pixel-level.**
Heat map showing the Jaccard similarity (pixel-level) between corresponding segmentation maps. Compared are the segmentation maps of the five human experts, a semi-automated signal thresholding approach and the CNN. The Jaccard similarities are shown as median value (color-coded, upper-left half) and as boxplot, together with the individual data points in the normal distribution curve (lower-right half). The orange lines mark the value of 0.5 (n=36 for each comparison between two coders). Also on pixel-level, our trained cFOS-model is as similar to human experts (0.37-0.56), as they are among themselves (0.32-0.56).

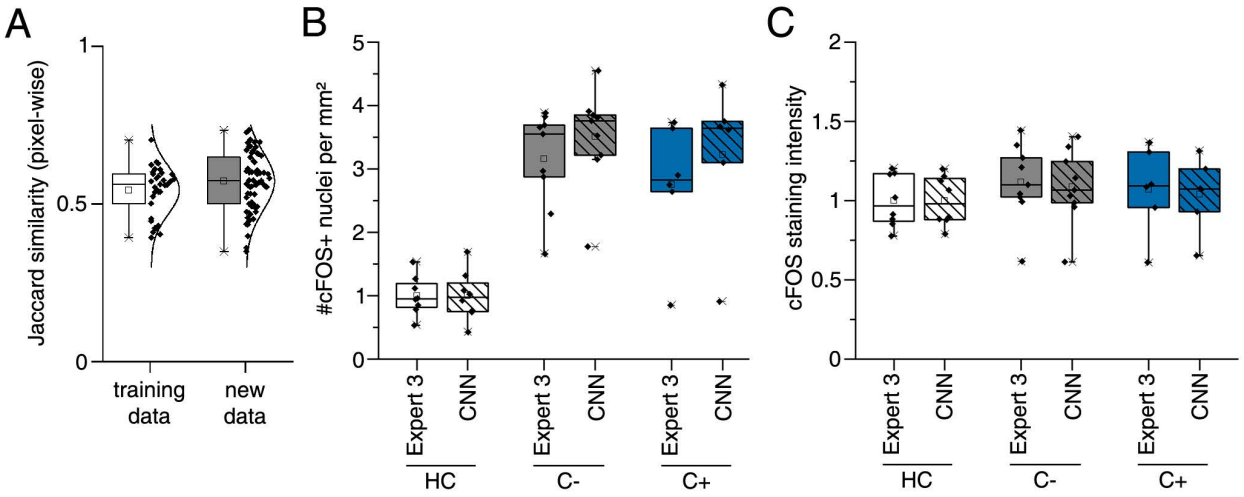

Figure S.3: **Model performance is maintained on images outside of the training dataset.**
To exclude that the trained cFOS-model is over-fitted on the initial 36 images (training data), we tested the model for generalizability on 65 new images (new data).
(A) The calculation of Jaccard similarities on pixel-level between the segmentation maps created by the cFOS-model or Expert 3 show no differences between training and new data ($p < 0.05$, Mann-Whitney test, $n_{training\ data} = 36$, new data: $n_{new\ data} = 65$)
(B and C) CNN-based analysis compared to manual analysis by a human expert. We calculated the number (B) and cFOS signal intensity (C) of cFOS-positive ROIs annotated by either Expert 3 or the cFOS-model on images outside of the training dataset. Both analyses indicate a context-dependent increase in the number of cFOS+ nuclei of 3.5 (C- and C+), compared to a nave learning control (HC, B). The mean cFOS signal intensity remained unchanged in both analyses (C). All values are normalized to the mean of the respective homecage control.

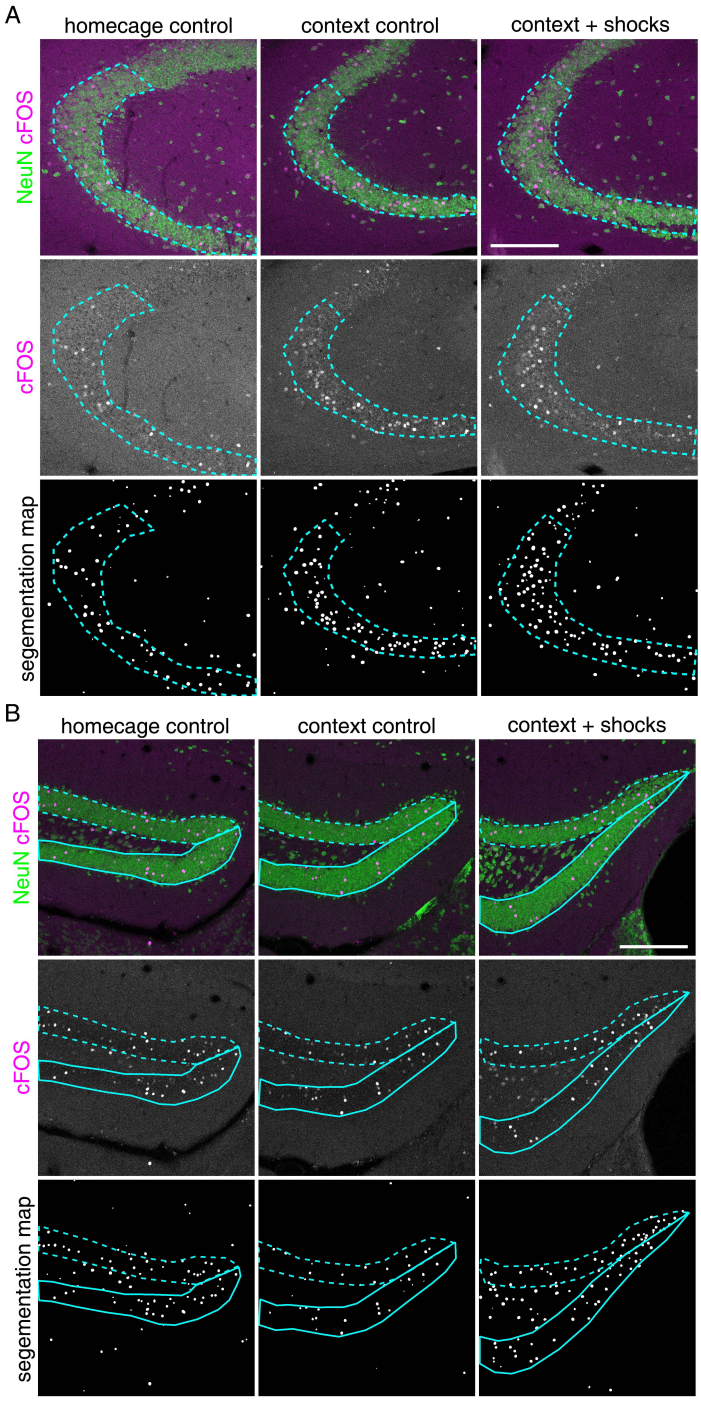

Figure S.4: **Representative images of cFOS labels in the hippocampus after behavioral training.**
(A) Representative images showing cFOS labels in CA3 of the three experimental groups (HC, C-, C+). A dashed line marks the analyzed area in CA3. Green: NeuN, magenta: cFOS. Scale bar: 200 µm.
(B) Representative images showing cFOS labels in the dentate gyrus of the three experimental groups (HC, C-, C+). A dashed line marks the analyzed area of the suprapyramidal blade. A solid line marks the investigated area of the infrapyramidal blade. The segmentation maps are based on CNN predictions. Green: NeuN, magenta: cFOS. Scale bar: 200 µm.

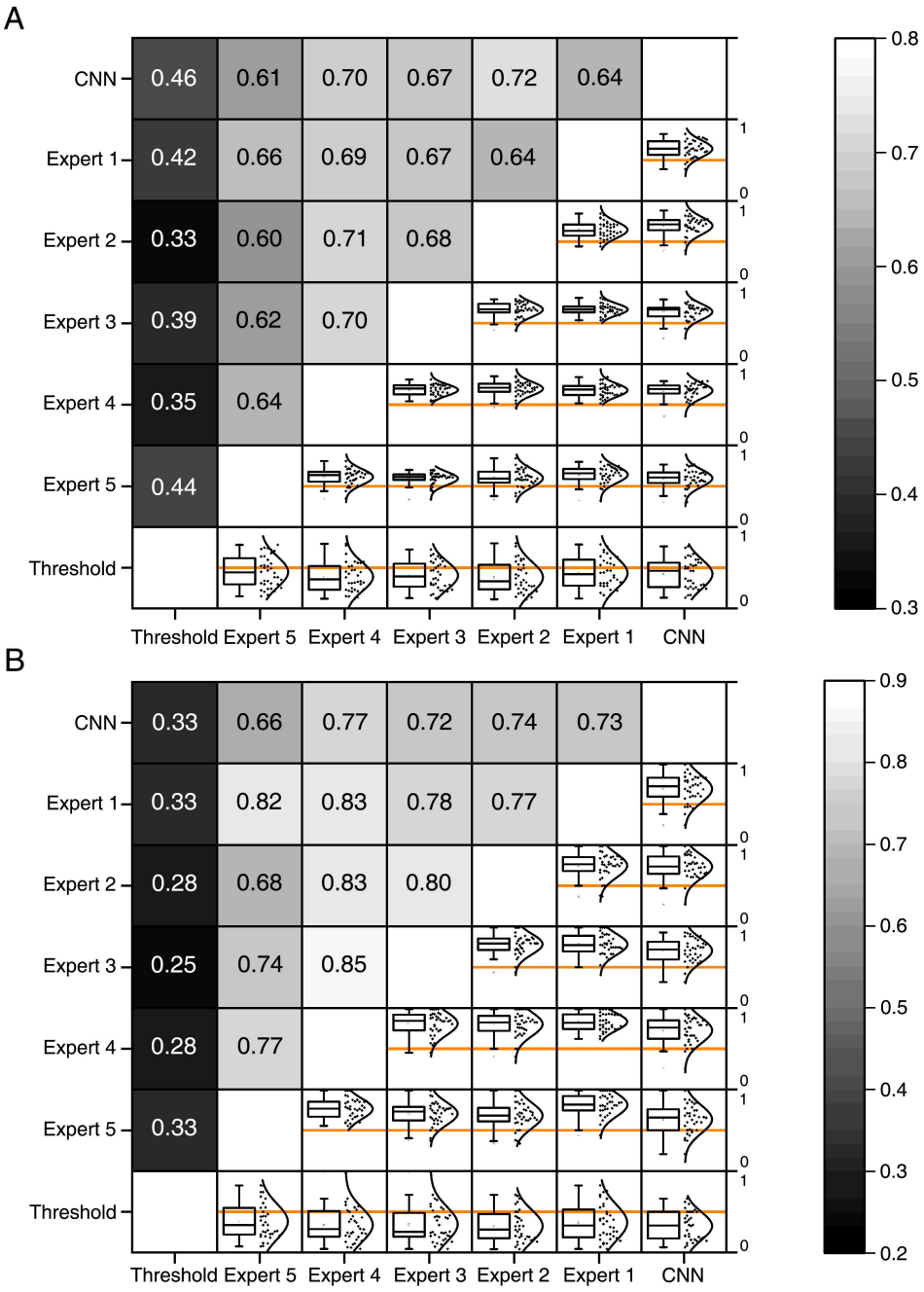

Figure S.5: **Automatic segmentation of Parvalbumin immunolabels by a trained CNN-model shows expert-like performance.** To demonstrate the flexibility of DeepFLaSH, we trained a second model again in an inter-coding approach to segment the somata of Parvalbumin-positive interneurons. In absence of a ground-truth, we assessed its performance on base of similarity analysis. (A and B) Heat map showing the Jaccard similarity on pixel- (A) and on ROI-level (B) between corresponding segmentation maps. Compared are the segmentation maps of the five human experts, a semi-automated signal thresholding approach and the CNN. The Jaccard similarities are shown as median value (color-coded, upper-left half) and as boxplot, together with the individual data points in the normal distribution curve (lower-right half). The orange lines mark the value of 0.5 (n=36 for each comparison between two coders). CNN-based segmentation maps are on both, pixel- and ROI-level as similar to those of human experts, as they are among themselves (inter-coder variability).

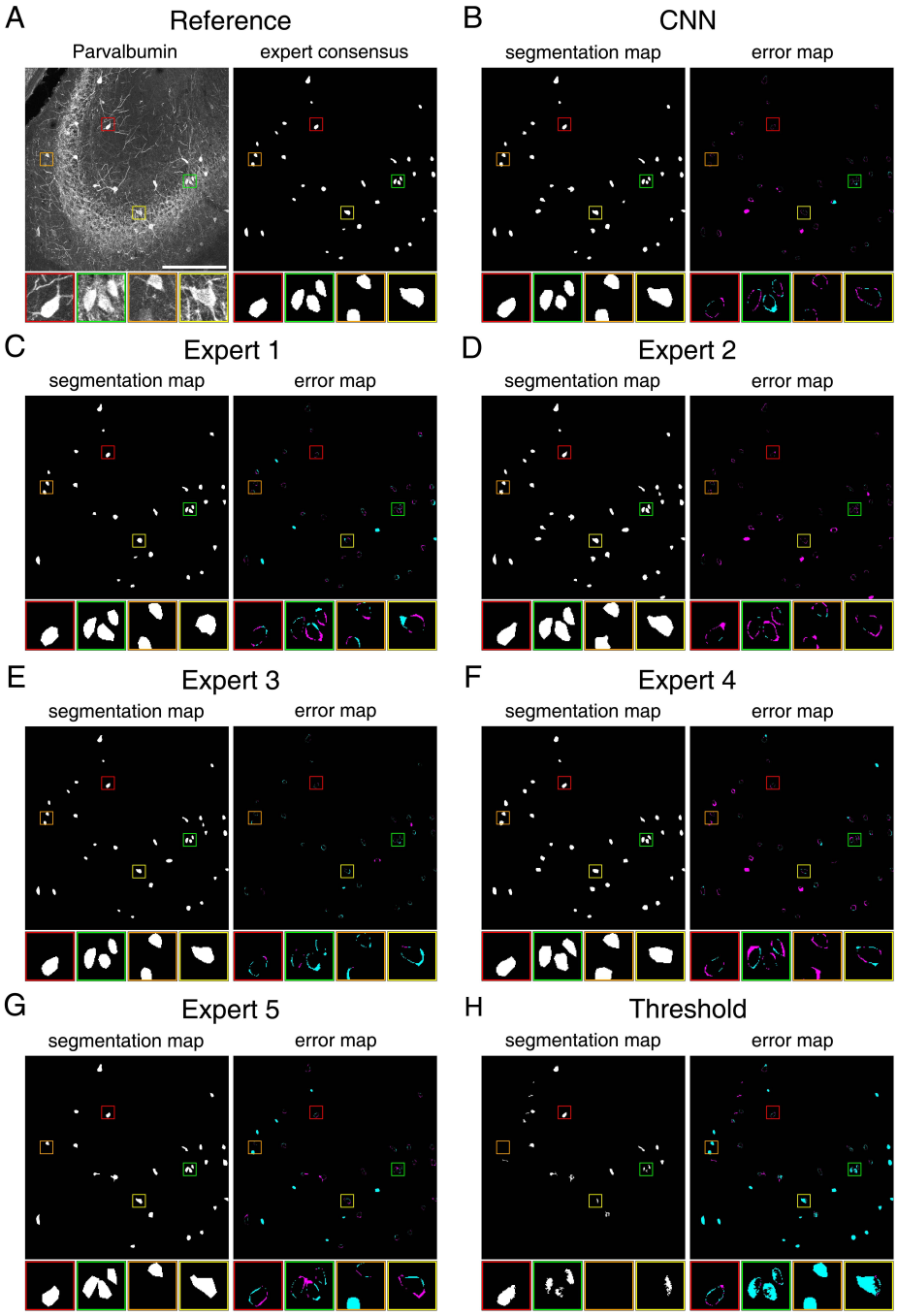

Figure S.6: **Accuracy of segmentation of Parvalbumin-positive somata at pixel-wise resolution.**
Based on a computed expert-consensus, we created error-maps that show the pixel-wise deviation of the individual coder segmentation of Parvalbumin-positive somata from the expert-consensus.
(A) Comparison of Parvalbumin microscopy image with expert consensus segmentation map. Four inlets are selected to highlight the variability of typical anti-Parvalbumin fluorescent labels.
(B-H) Segmentation maps and computed error maps are shown for the indicated coder. The error maps are pixel-wise comparisons of the corresponding coder segmentation to the expert consensus map. In cyan: pixels exclusively present in the expert consensus; in magenta: pixels that were exclusively labeled by the indicated coder. The trained PV-model was able to detect also ROIs with a close-to-noise signal intensity (orange inlet).

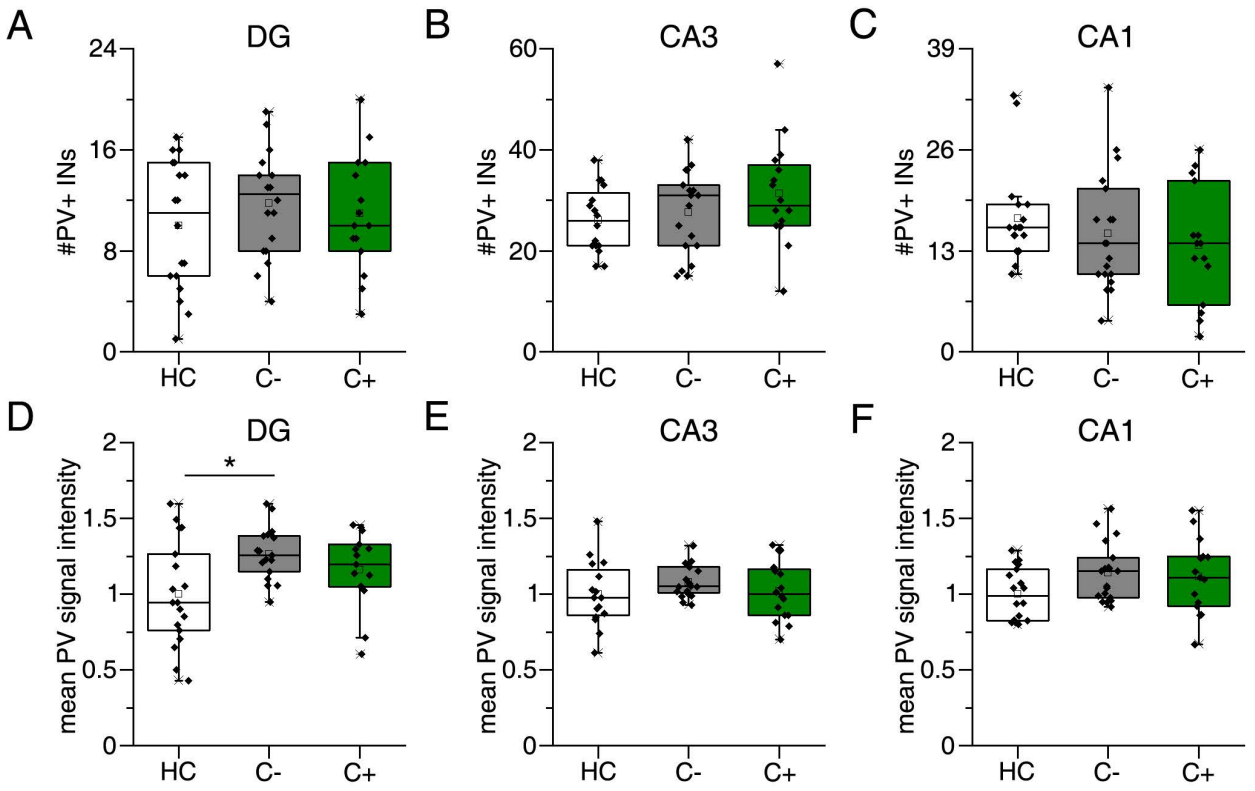

Figure S.7: **Image analysis with a DeepFLaSH-model for Parvalbumin fluorescent label segmentation.**
We used the PV-model to analyze behavior-related changes in Parvalbumin-positive interneurons. The deep segmentation maps (118 images) and manual segmentation data (36 images) were pooled for the statistical analysis to ensure the use of all data.
(A-C) Analysis of the number of Parvalbumin-positive somata per image in the indicated regions show no differences between the three experimental groups (DG: $F_{(2,\ 48)} = 0.653$, $p = 0.525$, $n_{HC} = 18$, $n_{C-} = 18$, $n_{C+} = 15$, one-way ANOVA; CA3: $F_{(2,\ 48)} = 1.623$, $p = 0.208$, $n_{HC} = 16$, $n_{C-} = 19$, $n_{C+} = 16$, one-way ANOVA; CA1: $X^2(2) = 2.452$, $p = 0.293$, $n_{HC} = 18$, $n_{C-} = 19$, $n_{C+} = 15$, Kruskal-Wallis ANOVA; for all: $N_{HC} = 3$, $N_{C-} = 4$, $N_{C+} = 3$).
(D-F) Analysis of mean PV signal intensity of all PV+ somata per image, normalized to the mean of the respective homecage control, in indicated regions. No significant differences were observed for the regions CA3 (E) and CA1 (F). A significantly higher mean PV signal intensity was detected in the DG of C- mice compared to HC, but not to C+ or between HC and C+ (DG: $X^2(2) = 6.415$, $p<0.05$, $n_{HC} = 18$, $n_{C-} = 17$, $n_{C+} = 15$, Kruskal-Wallis ANOVA followed by Mann-Whitney tests with Bonferroni correction, $^*P<0.05$; CA3: $F_{(2,\ 47)} = 1.022$, $p = 0.368$, $n_{HC} = 16$, $n_{C-} = 18$, $n_{C+} = 16$, one-way ANOVA; CA1: $X^2(2) = 5.013$, $p = 0.082$, $n_{HC} = 18$, $n_{C-} = 19$, $n_{C+} = 15$, Kruskal-Wallis ANOVA; for all: $N_{HC} = 3$, $N_{C-} = 4$, $N_{C+} = 3$).

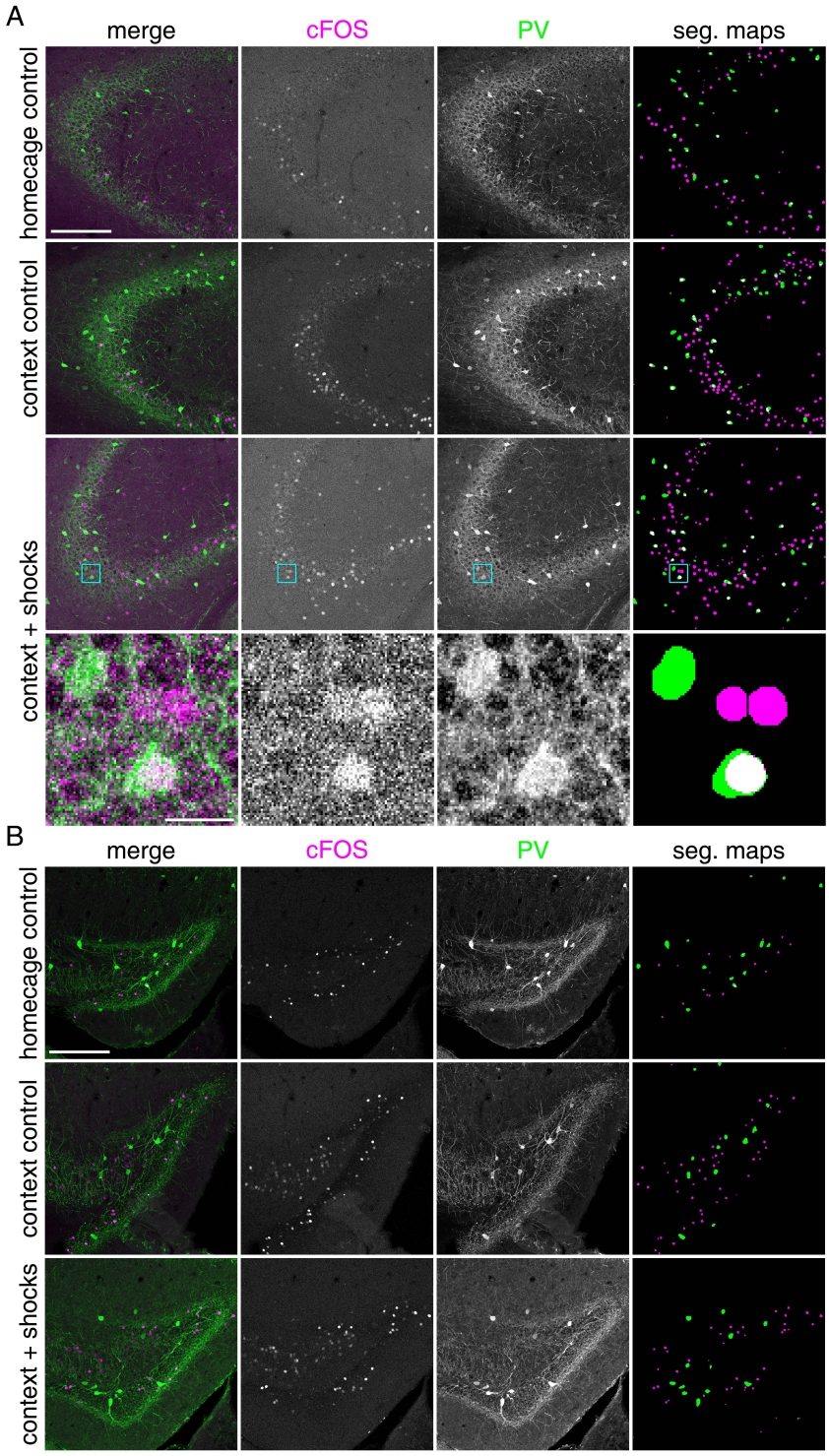

Figure S.8: **Representative images of cFOS and Parvalbumin labels in the hippocampus after behavioral training.**
(A and B) Representative images showing cFOS and Parvalbumin fluorescent labels in CA3 (A) or in the dentate gyrus (B) of the three experimental groups (HC, C-, C+). The segmentation maps are based on CNN-predictions. Scale bars: 200 μm.
(in A) The cyan inset shows a cFOS+ PV+ soma and a cFOS- PV+ soma, as well as two additional cFOS+ nuclei. Scale bar: 20 μm.

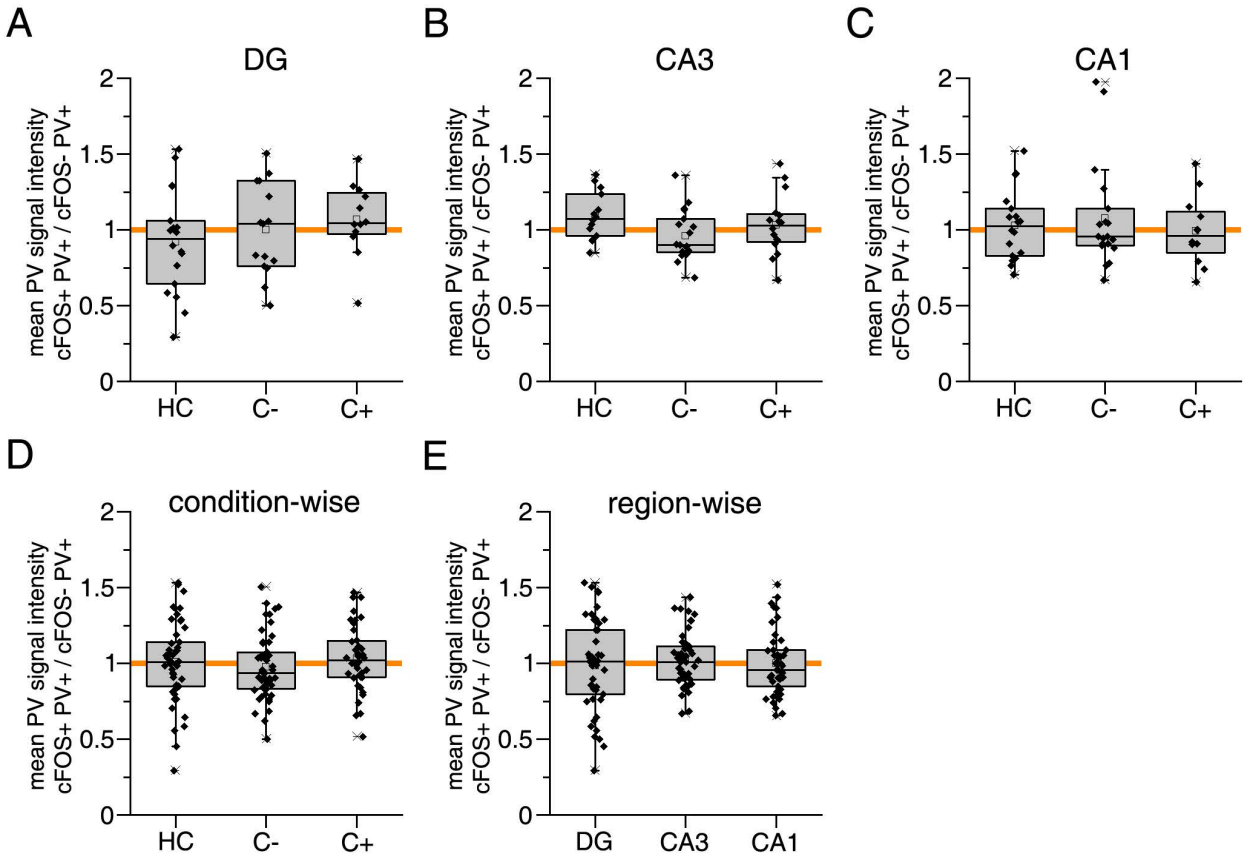

Figure S.9: **Extended analysis of hippocampal Parvalbumin-positive interneurons after behavioral training using two DeepFLaSH-models in combination.**
To analyze behavior-related changes in Parvalbumin-positive interneurons, we used a combination of the cFOS- and the PV-model. The deep segmentation maps (118 images) and manual segmentation data (36 images) were pooled for the statistical analysis to ensure the use of all data.
A-E) Image-wise ratio of mean PV fluorescent signal intensity of cFOS-positive PV+ somata compared to the mean signal intensity of cFOS-negative PV+ somata. Both types of cells show similar mean intensity in the PV fluorescent signal, thus indicating the same amount of PV in both cell type somata (DG: $F_{(2,\ 42)} = 0.885$, $p = 0.420$, $n_{HC} = 18$, $n_{C-} = 15$, $n_{C+} = 12$, $N_{HC} = 3$, $N_{C-} = 3$, $N_{C+} = 3$, one-way ANOVA; CA3: $F_{(2,\ 46)} = 2.422$, $p = 0.100$, $n_{HC} = 14$, $n_{C-} = 19$, $n_{C+} = 16$, $N_{HC} = 3$, $N_{C-} = 4$, $N_{C+} = 3$, one-way ANOVA; CA1: $X^2(2) = 0.211$, $p = 0.900$, $n_{HC} = 18$, $n_{C-} = 19$, $n_{C+} = 12$, $N_{HC} = 3$, $N_{C-} = 4$, $N_{C+} = 3$, Kruskal-Wallis ANOVA; condition-wise: $p>0.05$ for HC, C- and C+, one-sample t-tests, $n_{HC} = 50$, $n_{C-} = 51$, $n_{C+} = 40$, $N_{HC} = 3$, $N_{C-} = 4$, $N_{C+} = 3$; region-wise: $p>0.05$ for DG, CA3 and CA1, one-sample t-tests, $n_{DG} = 45$, $n_{CA3} = 49$, $n_{CA1} = 47$, $N_{DG} = 9$, $N_{CA3} = 10$, $N_{CA1} = 10$).