On the objectivity, reliability, and validity of deep learning enabled bioimage analyses

Dennis Segebarth[1†], Matthias Griebel[2†], Nikolai Stein[2], Cora R. von Collenberg[1], Corinna Martin[1], Dominik Fiedler[3], Lucas B. Comeras[4], Anupam Sah[5], Victoria Schoeffler[6], Theresa Lüffe[6], Alexander Dürr[2], Rohini Gupta[1], Manju Sasi[1], Christina Lillesaar[6], Maren D. Lange[3], Ramon O. Tasan[4], Nicolas Singewald[5], Hans-Christian Pape[3], Christoph M. Flath[2*‡], Robert Blum[1,7*‡]

***For correspondence:** christoph.flath@uni-wuerzburg.de (CMF); Blum_R@ukw.de (RB)

[†]These authors contributed equally to this work
[‡]These senior authors contributed equally to this work

[1]Institute of Clinical Neurobiology, University Hospital Würzburg, 97078 Würzburg, Germany; [2]Department of Business and Economics, University of Würzburg, 97070 Würzburg, Germany; [3]Institute of Physiology I, Westfälische Wilhelms-University, 48149 Münster, Germany; [4]Department of Pharmacology, Medical University of Innsbruck, 6020 Innsbruck, Austria; [5]Department of Pharmacology and Toxicology, Institute of Pharmacy and Center for Molecular Biosciences Innsbruck, University of Innsbruck, 6020 Innsbruck, Austria; [6]Department of Child and Adolescent Psychiatry, Center of Mental Health, University Hospital of Würzburg, 97080 Würzburg, Germany; [7]Comprehensive Anxiety Center, 97078 Würzburg, Germany

Abstract Fluorescent labeling of biomolecules is widely used for bioimage analyses throughout the life sciences. Recent advances in deep learning (DL) have opened new possibilities to scale the image analysis processes through automation. However, the annotation of fluorescent features with a low signal-to-noise ratio is frequently based on subjective criteria. Training on subjective annotations may ultimately lead to biased DL models yielding irreproducible results. An end-to-end analysis process that integrates data annotation, ground truth estimation, and model training can mitigate this risk. To highlight the importance of this integrated process, we compare different DL-based analysis approaches. Based on data from different laboratories, we show that ground truth estimation from multiple human annotators is indispensable to establish objectivity in fluorescent feature annotations. We demonstrate that ensembles of multiple models trained on the estimated ground truth establish reliability and validity. Our research provides guidelines for reproducible and transparent bioimage analyses using DL methods.

Introduction

Modern microscopy methods enable researchers to capture images that describe cellular and molecular features in biological samples at an unprecedented scale. One of the most frequently used imaging methods is fluorescent labeling of biological macromolecules, both *in vitro* and *in vivo*. In order to test a biological hypothesis, fluorescent features have to be interpreted and analyzed quantitatively, a process known as bioimage analysis (***Meijering et al., 2016***). However, fluorescence does not provide clear signal-to-noise borders, forcing human experts to utilize individual heuristic criteria, such as morphology, size, or signal intensity to classify fluorescent signals as

background, or to, often manually, annotate them as a region of interest (ROI). Constant technological advances in fluorescence microscopy facilitate the automatized acquisition of large image datasets, even at high resolution and with high throughput (***Li et al., 2010***; ***McDole et al., 2018***; ***Osten and Margrie, 2013***). The ever increasing workload associated with image feature annotation therefore calls for computer-aided automated bioimage analysis. However, attempts to replace human experts and to automate the annotation process using traditional image thresholding techniques (e.g., histogram shape-, entropy-, or clustering-based methods (***Sezgin and Sankur, 2004***)) frequently lack flexibility, as they rely on a high signal-to-noise ratio in the images or require computational expertise for user-based adaptation to individual datasets (***von Chamier et al., 2019***).

In recent years, deep learning (DL) and in particular deep convolutional neural networks have shown remarkable capacities in image recognition tasks, opening new possibilities to perform automatized image analysis. DL-based approaches constantly outperform conventional methods (***Caicedo et al., 2019***) and, moreover, are capable of performing complex tasks such as artificial labeling of plain bright-field images (***von Chamier et al., 2019***; ***Christiansen et al., 2018***; ***Ounkomol et al., 2018***). The main difference between conventional and DL algorithms is that conventional algorithms follow predefined rules (hard-coded), while DL algorithms are flexible to learn the respective task on base of a training dataset (***LeCun et al., 2015***). Yet, deployment of DL approaches necessitates both computational expertise and suitable computing resources. These requirements frequently prevent non-AI experts from applying DL to routine image analysis tasks. Initial efforts have already been made to break down these barriers, for instance, by integration into prevalent bioimaging tools such as *ImageJ* (***Falk et al., 2019***) and *CellProfiler* (***McQuin et al., 2018***), or using cloud-based approaches (***Haberl et al., 2018***).

To harness the potentials of these DL-based methods they require integration into the bioimage analysis pipeline. Such pipelines for processing complex data frequently lack standardization and, at present, draw attention as they appear to alter the results of scientific studies (***Lindquist, 2020***). We argue that such an integration into the scientific process ultimately necessitates DL-based approaches to meet the same standards as any method in an empirical study. We can derive these standards from the general quality criteria of qualitative and quantitative research: objectivity, reliability, and validity (***Frambach et al., 2013***).

Objectivity refers to the neutrality of evidence, with the aim to eliminate personal biases, emotions, or simply limitations introduced by the context in which manual feature annotation is performed (***Frambach et al., 2013***). Manual annotation of fluorescent features has long been known to be subjective, especially in the case of weak signal-to-noise thresholds (***Schmitz et al., 1999***; ***Collier et al., 2003***; ***Niedworok et al., 2016***). Notably, there is no objective ground truth reference in the particular case of fluorescent label segmentation, causing a critical problem for training and evaluation of DL algorithms. As multiple studies have pointed out that inconsistent or biased annotations can cause DL algorithms to either fail to train or to reproduce biased annotations on new data (***von Chamier et al., 2019***) this is a crucial obstacle for applying DL to bioimage analysis processes.

Reliability is concerned with the consistency of evidence (***Frambach et al., 2013***). To allow an unambiguous understanding of this concept, we further distinguish between repeatability and reproducibility. Repeatability or test-retest reliability is defined as “closeness of the agreement between the results of successive measurements of the same measure and carried out under the same conditions" (***Taylor and Kuyatt, 1994***, 14), which is guaranteed for any deterministic DL model. Reproducibility, on the other hand, is specified as “closeness of the agreement between the results of measurements of the same measure and carried out under changed conditions" (***Taylor and Kuyatt, 1994***, 14), e.g., different observer, or different apparatus. This is a critical point, since the output of different DL models trained on the same training dataset can vary significantly. This is caused by the stochastic training procedure (e.g., random initialization, random sampling and data augmentation (***Ronneberger et al., 2015***)), the choice of model parameters (e.g., model architecture, weights, activation functions), and the choice of hyperparameters (e.g., learning rate,

mini-batch size, training epochs). Consequently, the reproducibility of DL models merits careful investigation.

Finally, *validity* relates to the truth value of evidence, i.e, whether we in fact measured what we intended to. Moreover, validity implies reliability - but not vice versa (***Frambach et al., 2013***). On a basis of a given ground truth, validity is typically measured using appropriate similarity measures such as F1 score for detection and Intersection over Union (IoU) for segmentation purposes (***Ronneberger et al., 2015***; ***Falk et al., 2019***; ***Caicedo et al., 2019***). In addition, the deep learning community has established widely accepted standards for training models. These comprise, among other things, techniques to avoid overfitting (regularization techniques and cross-validation), tuning hyperparameters, and selecting appropriate metrics for model evaluation. However, these standards do not apply for the training and evaluation of a DL model in the absence of a ground truth, like in the case of fluorescent features.

Taken together and with regard to the discussion about a reproducibility crisis in the fields of biology, medicine and artificial intelligence (***Siebert et al., 2015***; ***Baker, 2016***; ***Ioannidis, 2016***; ***Hutson, 2018***; ***Fanelli, 2018***; ***Chen et al., 2019***), these limitations indicate that DL could aggravate this crisis by adding even more unknowns and uncertainties to bioimage analyses.

However, the present study asks whether DL, if instantiated in an appropriate manner, also holds the potential to instead enhance the objectivity, reproducibility and validity of bioimage analysis. To tackle this conundrum, we investigated different DL-based strategies on five fluorescence image datasets. We show that training of DL models on the pooled input of multiple human experts utilizing ground truth estimation (consensus models) increases objectivity of fluorescent feature segmentation. Furthermore, we demonstrate that ensembles of consensus models are even capable of enhancing the reliability and validity of bioimage analysis of ambiguous image data, such as fluorescence features in histological tissue sections.

Results

In order to evaluate the impact of DL on bioimage analysis results, we instantiated three exemplary DL-based strategies (***Figure 1***; strategies color-coded in gray, blue, and orange) and investigate them in terms of objectivity, reliability and validity of fluorescent feature annotation. The first strategy, *expert models* (gray), reflects mere automation of the annotation process of fluorescent features in microscopy images. Here, manual annotations of a single human expert are used to train an individual (and hence expert-specific) DL model with a U-Net (***Ronneberger et al., 2015***) architecture. U-Net and its variants have emerged as the de facto standard for biomedical image segmentation purposes (***McQuin et al., 2018***; ***Falk et al., 2019***; ***Caicedo et al., 2019***). The second strategy, *consensus models* (blue), addresses the objectivity of signal annotations. Contrary to the first strategy, simultaneous truth and performance level estimation (STAPLE) (***Warfield et al., 2004***) is used to estimate a ground truth and create consensus annotations. The estimated ground truth (est. GT) annotation reflects the pooled input of multiple human experts and is therefore thought to be less affected by a subjective bias of a single expert. We then train a single U-Net model to create a consensus model. The third strategy, *consensus ensembles* (orange), seeks to ensure reliability and eventually validity. Going beyond the second strategy, we train multiple consensus U-Net models to create a consensus ensemble. Such model ensembles are known to be more robust to noise (***Dietterich, 2000***). Hence, we hypothesize that the consensus ensembles mitigate the randomness in the training process. Moreover, deep ensembles are supposed to yield high quality predictive uncertainty estimates (***Lakshminarayanan et al., 2017***).

For each of the three strategies, we complete the bioimage analysis by performing quantification and hypothesis testing on a typical fluorescent microscopy image dataset (***Figure S2***). These images describe changes in fluorescence signal abundance of a protein called cFOS in brain sections of mice. cFOS is an activity-dependent transcription factor and its expression in the brain can be modified experimentally by behavioral testing of the animals (***Gallo et al., 2018***). The low

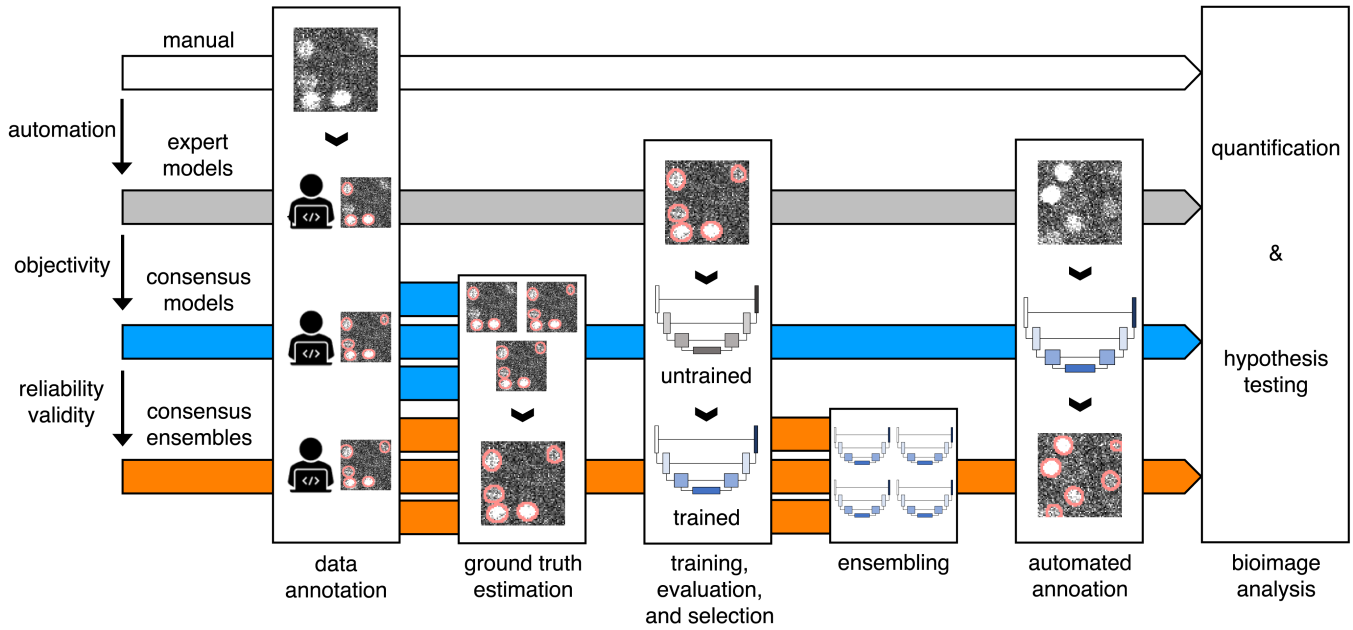


Figure 1. Schematic illustration of bioimage analysis strategies and corresponding hypotheses. Four bioimage analysis strategies are depicted. Manual (white) refers to manual, heuristic fluorescent feature annotation by a human expert. The three DL-based strategies for automatized fluorescent feature annotation are based on expert models (gray), consensus models (blue) and consensus ensembles (orange). For all DL-based strategies, a representative subset of microscopy images is annotated by human experts. Here, we depict labels of cFOS-positive nuclei and the corresponding annotations (pink). These annotations are used in either individual training datasets (gray: expert models) or pooled in a single training dataset by means of ground truth estimation from the expert annotations (blue: consensus models, orange: consensus ensembles). Next, deep learning models are trained on the training dataset and evaluated on a holdout validation dataset. Subsequently, the predictions of individual models (gray and blue) or model ensembles (orange) are used to compute binary segmentation masks for the entire bioimage dataset. Based on these fluorescent feature segmentations, quantification and statistical analyses are performed. The expert model strategy enables the automation of a manual analysis. To mitigate the bias from subjective feature annotations in the expert model strategy we introduce the consensus model strategy. Finally, the consensus ensembles alleviate the random effects in the training procedure and seek to ensure reliability and eventually, validity. Figure 1 - Supplementary figures: ***Figure S1*** and ***Figure S2***.

signal-to-noise ratio of this label, its broad usage in neurobiology and the well-established correlation of its abundance with behavioral paradigms render it an ideal bioimage dataset to test our hypotheses (***Shuvaev et al., 2017***; ***Gallo et al., 2018***).

Consensus ensembles yield the best results for validity and reproducibility metrics

The primary goal in bioimage analysis is to rigorously test a biological hypothesis. To leverage potentials of DL models within this procedure, we need to trust our model – by establishing objectivity, reliability and validity. Pertaining to the case of fluorescent labels, validity (measuring what is intended to be measured) requires objectivity to know what exactly we intend to measure in the absence of a ground truth. Similarly, reliability in terms of repeatability and reproducibility is a prerequisite for a valid and trustworthy model. Starting from the expert model strategy, we seek to establish objectivity (consensus models) and, successively, reliability and validity in the consensus ensemble strategy. In the following analysis, we first turn towards a comprehensive evaluation of the objectivity and its relation to validity before moving on to the concept of reliability.

To assess the three different strategies, a training dataset of 36 images and a test set of nine microscopy images (see ***Figure S2***) showing cFOS immunoreactivity were manually annotated by five independent experts (experts 1-5). In absence of a rigorously objective ground truth, we used STAPLE (***Warfield et al., 2004***) to compute an estimated ground truth (est. GT) based on all expert annotations for each image. First, we trained a set of DL models on the 36 training images and

corresponding annotations, either made by an individual human expert or as reflected in the est. GT (see methods for the data set and detailed training, evaluation and model selection strategy). Then, we used our test set to evaluate the segmentation (Mean IoU) and detection (F1 score) performance of human experts and all trained models by means of similarity analysis on the level of individual images.

For the pairwise comparison of annotations (segmentation masks) we calculated the intersection over union (IoU) for all overlapping pairs of ROIs between two segmentation masks (***Figure 2A***; see Segmentation and detection). Following ***Maška et al.*** (***2014***), we consider two ROIs with an IoU of at least 0.5 as matching and calculated the F1 score $M_{\text{F1 score}}$ as the harmonic mean of precision and recall (***Figure 2B***; see Segmentation and detection). As bioimaging studies predominantly use measures related to counting ROIs in their analyses, we also focused on the feature detection performance ($M_{\text{F1 score}}$). The color coding (gray, blue, orange) introduced in ***Figure 2C*** refers to the different strategies depicted in ***Figure 1*** and applies to all figures, if not indicated otherwise.

To better grasp the difficulties in annotating cFOS-positive nuclei as fluorescent features in these images, we first compared manual expert annotations (***Figure 2D***). The analysis revealed substantial differences between the annotations of the different experts and shows varying inter-rater agreement (***Schmitz et al., 1999***; ***Collier et al., 2003***; ***Niedworok et al., 2016***). The level of inter-rater variability was inversely correlated with the relative signal intensities (***Figure S3***) (***Niedworok et al., 2016***).

By comparing the annotations of the expert models (gray) to the annotations of the respective expert (***Figure 2E***) we observed a higher $M_{\text{F1 score}}$ median compared to the inter-rater agreement (***Figure 2D***) in the majority of cases. Furthermore, comparing the similarity analysis results of human experts with those of their respective expert-specific models revealed that they are closely related (***Figure 2F***, ***Figure S5***, and ***Figure S6***). As pointed out by ***von Chamier et al.*** (***2019***), this indicates that our expert models are able to learn and reproduce the subjective bias of the individual experts. This becomes particularly evident in the annotations of the DL models trained on expert 1 (***Figure 2F***, ***Figure S5***, and ***Figure S6***).

Overall, the expert models yield a lower similarity to the est. GT compared to the consensus models (blue) or consensus ensembles (orange). Notably, both consensus models and consensus ensembles perform on par with human experts and even surpass the average human expert. Hereby, the consensus ensembles outperform all other strategies, even at varying IoU thresholds (***Figure 2F*** and ***Figure 2G***).

In order to test for reliability of our analysis, we measured the repeatability and reproducibility of fluorescent feature annotation of our DL strategies. We assumed that the repeatability is assured for all our strategies due to the deterministic nature of our DL models (unchanged conditions imply unchanged model weights). Hence, our evaluation was focused on the reproducibility, meaning the impact of the stochastic training process on the output. Inter-expert and inter-model comparisons within each strategy unveiled a better performance of the consensus ensembles strategy concerning both detection ($M_{\text{F1 score}}$) and segmentation ($\bar{M}_{\text{IoU}}$) of the fluorescent features (***Figure 2H***). Calculating the Fleiss' kappa value (***Fleiss and Cohen, 1973***) revealed that consensus ensemble annotations show a high reliability of agreement (***Figure 2H***). Following the Fleiss' kappa interpretation from ***Landis and Koch*** (***1977***) the results for the consensus ensembles indicate a substantial or almost perfect agreement. In contrast, the Fleiss' kappa values for human experts refer to a fair agreement while the results for the alternative DL strategies lead to a moderate agreement (***Figure 2H***).

In summary, the similarity analysis of the three strategies shows that training of DL models solely on the input of a single human expert imposes a high risk of incorporating an intrinsic bias and therefore resembles, as hypothesized, a mere automation of manual image annotation. Both consensus models and consensus ensembles perform on par with human experts regarding the similarity to the est. GT, but the consensus ensembles yield by far the best results regarding their reproducibility. We conclude that, in terms of similarity metrics, only the consensus ensemble

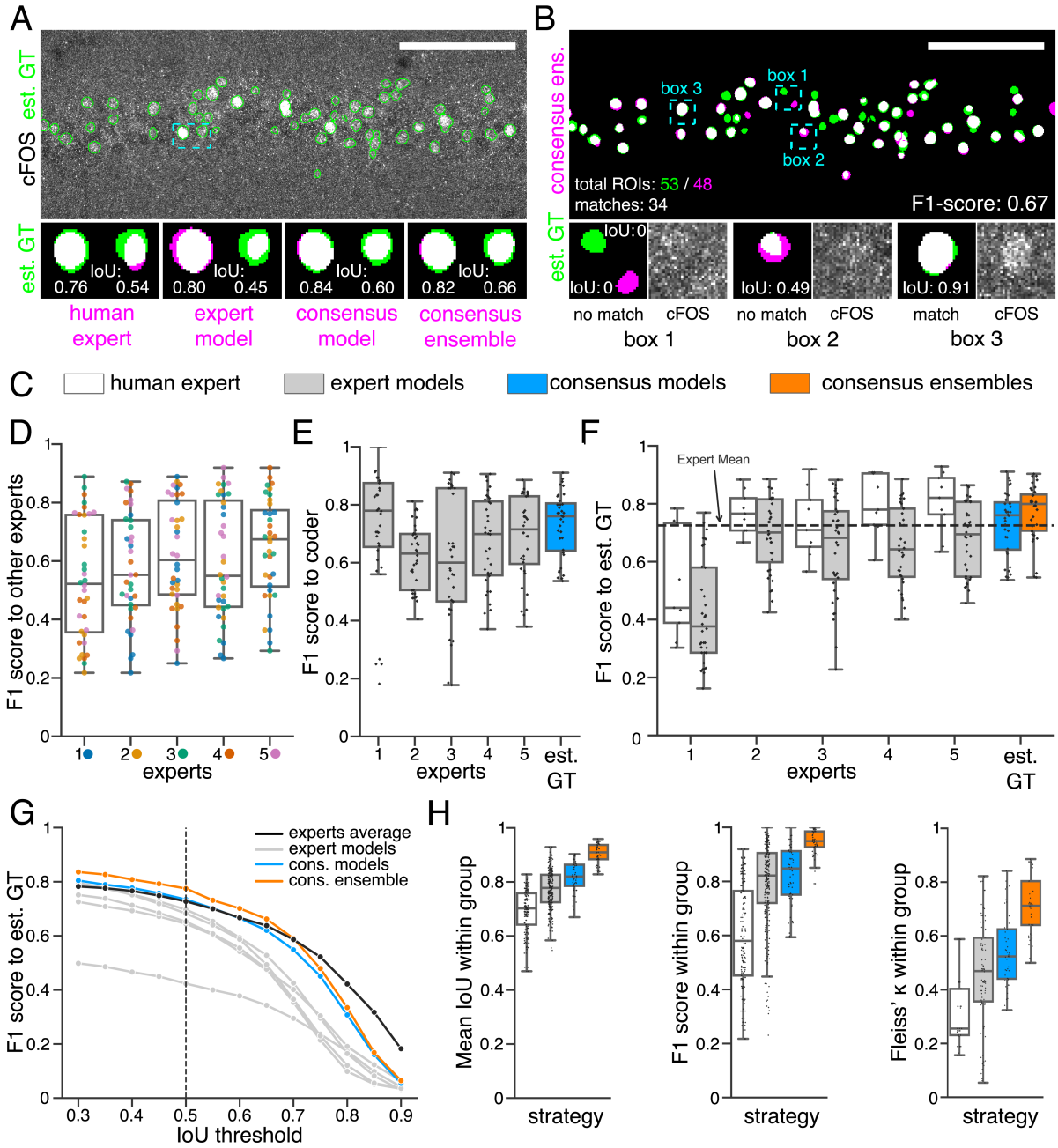


Figure 2. Similarity analysis of fluorescent feature annotations by manual or DL-based strategies. (**A**) Representative example of IoU M_{IoU} calculations on a field of view (FOV) in a bioimage. Image raw data show the labeling of cFOS in a maximum intensity projection image of the CA1 region in the hippocampus (brightness and contrast enhanced). The similarity of estimated ground truth (est. GT) annotations (green), derived from the annotations of five expert neuroscientists, are compared to those of one human expert, an expert model, a consensus model, and a consensus ensemble (magenta, respectively). IoU results of two ROIs are shown in detail for each comparison (magnification of cyan box). Scale bar: 100 µm. (**B**) F1 score $M_{\text{F1 score}}$ calculations on the same FOV as shown in (A). The est. GT annotations (green; 53 ROIs) are compared to those of a consensus ensemble (magenta; 48 ROIs). IoU-based matching of ROIs at an IoU-threshold of $t = 0.5$ is depicted in three magnified subregions of the image (cyan boxes 1-3). Scale bar: 100 µm. (**C-H**) All comparisons are performed exclusively on a separate test set which was withheld from model training and validation. (**C**) Color coding refers to the individual strategies, as introduced in ***Figure 1***: white: manual approach, gray: expert models, blue: consensus models, orange: consensus ensembles. (**D**) $M_{\text{F1 score}}$ between individual manual expert annotations and their overall reliability of agreement given as the mean of Fleiss' κ. (**E**) $M_{\text{F1 score}}$ between annotations predicted by individual models and the annotations of the respective coder, whose annotations were used for training. $N_{\text{models per expert}}$=4. (**F**) $M_{\text{F1 score}}$ between manual expert annotations, the respective expert models, consensus models, and consensus ensembles compared to the est. GT as reference. A horizontal line denotes human expert average. N_{models}=4, $N_{\text{ensembles}}$=4. (**G**) Means of $M_{\text{F1 score}}$ of the individual DL-based strategies and of the human expert average compared to the est. GT plotted for different IoU matching thresholds t. A dashed line indicates the default threshold $t = 0.5$. N_{models}=4, $N_{\text{ensembles}}$=4. (**H**) Annotation reliability of the individual strategies assessed as the similarities between annotations within the respective strategy. We calculated $\bar{M}_{\text{IoU}}$, $M_{\text{F1 score}}$ and Fleiss' κ. N_{experts}=5, N_{models}=4, $N_{\text{ensembles}}$=4. Figure 2 - Supplementary figures: ***Figure S3***, ***Figure S4***, ***Figure S5***, and ***Figure S6***.

strategy meet the bioimaging standards for objectivity, reliability, and validity.

Consensus ensembles yield reliable bioimage analysis results

Similarity analysis is inevitable to assess the quality of a model's output, i.e., the predicted segmentations (***Ronneberger et al., 2015***; ***Caicedo et al., 2019***; ***Falk et al., 2019***). However, the primary goal of bioimage analysis is the unbiased quantification of distinct image features that correlate with experimental conditions. So far, it has remained unclear whether objectivity, reliability and validity for bioimage analysis can be inferred directly from similarity metrics.

In order to systematically address this question, we used our image dataset to quantify the abundance of cFOS in brain sections of mice after Pavlovian contextual fear conditioning. It is well established in the neuroscientific literature that mice show changes in the distribution and abundance of cFOS in a specific brain region, namely the hippocampus, after processing information about places and contexts (***Keiser et al., 2017***; ***Campeau et al., 1997***; ***Huff et al., 2006***; ***Ramamoorthi et al., 2011***; ***Tayler et al., 2013***; ***Murawski et al., 2012***; ***Guzowski et al., 2001***). Consequently, our experimental dataset offered us a second line of evidence, the objective analysis of mouse behavior, in addition to the changes of fluorescent features to validate the bioimage analyses results of our DL-based strategies.

Our dataset comprised three experimental groups (***Figure 3A***). In one group, mice were directly taken from their home cage as naïve learning controls (H). In the second group, mice were re-exposed to a previously explored training context as context controls (C-). Mice in the third group underwent Pavlovian fear conditioning and were also re-exposed to the training context (C+) (***Figure 3A***). These three groups of mice showed different behavioral responses. For instance, fear (threat) conditioned mice (C+) showed increased freezing behavior after fear acquisition and showed strong freezing responses when re-exposed to the training context 24 h later (***Figure S7***). After behavioral testing, brain sections of the different groups of mice were prepared and labeled for the neuronal activity-related protein cFOS by indirect immunofluorescence. Sections were also labeled with the neuronal marker NeuN (Fox3), allowing the anatomical identification of hippocampal subregions of interest. Images were acquired as confocal microscopy image stacks (x,y-z) and maximum intensity projections were used for subsequent bioimage analysis (***Figure S2***). Overall, we quantified the number of cFOS-positive nuclei and their mean signal intensity in five regions of the dorsal hippocampus (DG as a whole, suprapyramidal DG, infrapyramidal DG, CA3, and CA1), and tested for significant differences between the three experimental groups (***Figure 3B-D***). To extend this analysis beyond hypothesis testing at a certain significance level, we calculated the effect size (η^2) for each of these 30 pairwise comparisons.

We illustrate our metrics with the detailed quantification of cFOS-positive nuclei in the stratum pyramidale of CA1 as a representative example and show two analyses for each DL strategy (***Figure 3E***). These two examples represent those two models of each strategy that yielded the lowest and the highest effect sizes, respectively (***Figure 3E***). Despite a general consensus of all models and ensembles on a context-dependent increase in the number of cFOS-positive nuclei, these quantifications already indicate that the variability of effect sizes decreases from expert models to consensus models and is lowest for consensus ensembles (***Figure 3E***).

The analysis in ***Figure 4*** allows us to further explore the impact of the different DL strategies on the bioimage analysis results for each hippocampal subregion. Here, we display a high-level comparison of the effect sizes and corresponding significance levels of 20 independently trained expert models (four per expert), 36 consensus models, and nine consensus ensembles (each derived from four consensus models). In contrast to the detailed illustration of selected models in ***Figure 3E***, ***Figure 4A***, for instance, summarizes the results for all analyses of the stratum pyramidale of CA1. As indicated before, all models and ensembles show a highly significant context-dependent increase in the number of cFOS-positive nuclei, but also a notable variation in effect sizes for both expert and consensus models. Moreover, we identify a significant context-dependent increase in the mean signal intensity of cFOS-positive nuclei for all consensus models and ensembles. The

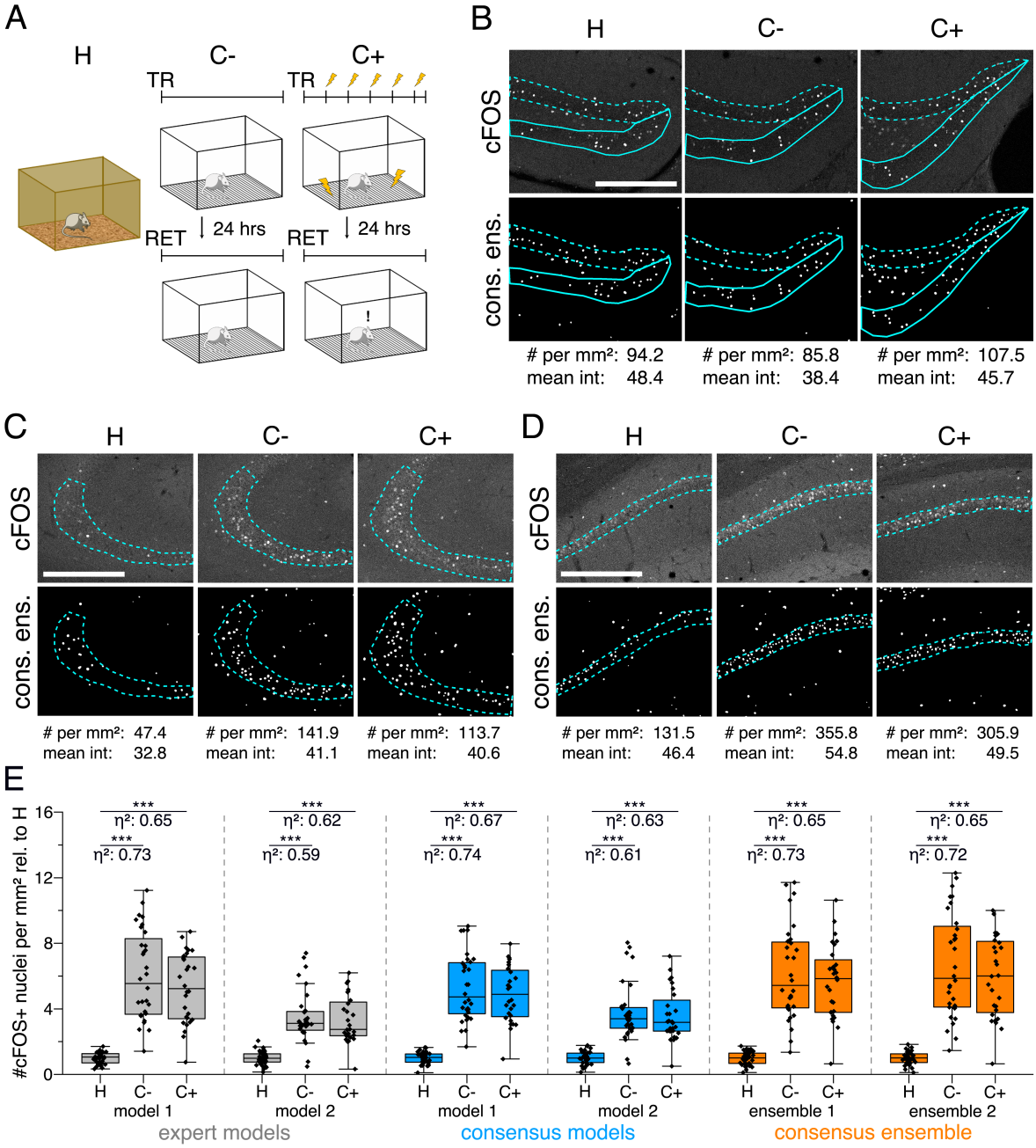


Figure 3. Application of different DL-based strategies for fluorescent feature annotation. The figure introduces how three DL-based strategies are applied for annotation of a representative fluorescent label, here cFOS, in a representative image data set. Raw image data show behavior-related changes in the abundance and distribution of the protein cFOS in the dorsal hippocampus, a brain center for encoding of context-dependent memory. (**A**) Three experimental groups were investigated: Mice kept in their homecage (H), 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). Memory retrieval induces changes in cFOS levels. (B-D) Brightness and contrast enhanced maximum intensity projections showing cFOS fluorescent labels of the three experimental groups (H, C-, C+) with representative annotations of a consensus ensemble, for each hippocampal subregion. The annotations are used to quantify the number of cFOS-positive nuclei for each image (#) per mm^2 and their mean signal intensity (mean int., in bit-values) within the corresponding image region of interest, here the neuronal layers in the hippocampus (outlined in cyan). In B: granule cell layer (supra- and infrapyramidal blade), dotted line: suprapyramidal blade, solid line: infrapyramidal blade. In C: pyramidal cell layer of CA3; in D: pyramidal cell layer in CA1. Scale bars: 200 µm. (E) Analyses of cFOS-positive nuclei per mm^2, representatively shown for stratum pyramidale of CA1. Corresponding effect sizes are given as η^2 for each pairwise comparison. Two quantification results are shown for each strategy and were selected to represent the lowest (model 1 or ensemble 1) and highest (model 2 or ensemble 2) effect sizes (increase in cFOS) reported within each annotation strategy. Total analyses performed: $N_{expert\ models} = 20$, $N_{consensus\ models} = 36$, $N_{consensus\ ensembles} = 9$. Number of analyzed mice (N) and images (n) per experimental condition: $N_H = 7$, $N_{C-} = 7$, $N_{C+} = 6$; $n_H = 36$, $n_{C-} = 32$, $n_{C+} = 28$. ***: $p < 0.001$. Statistical data are available in ***Figure 3 - Source data***. Figure 3 - Supplementary figures: ***Figure S7***.

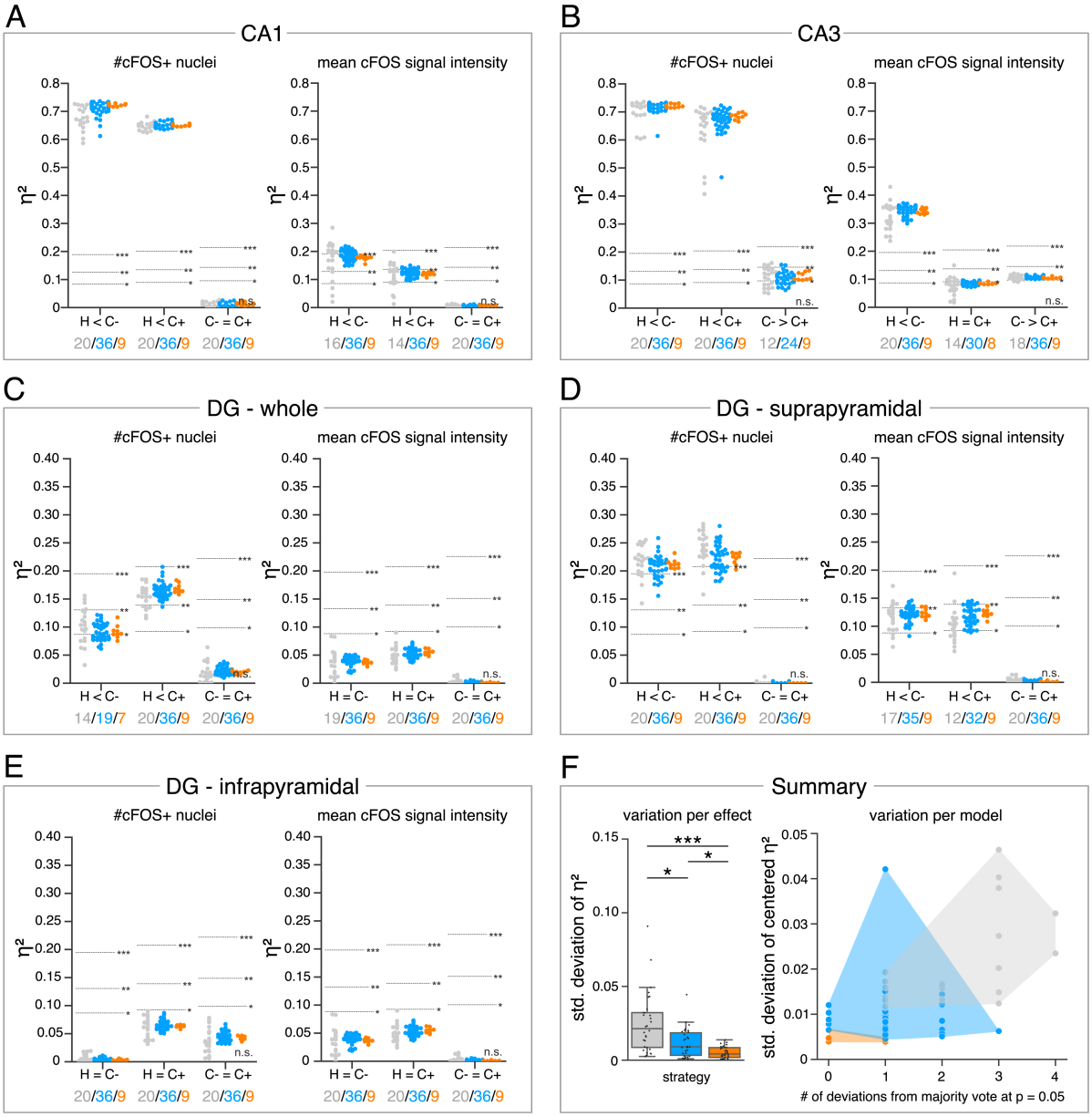


Figure 4. Consensus ensembles significantly increase reliability of bioimage analysis results. (**A-E**) Effect sizes for each pairwise comparison of all individual bioimage analyses for each DL-based strategy (gray: expert models, blue: consensus models, orange: consensus ensembles) in indicated hippocampal subregions. Three horizontal lines separate four significance intervals (n.s.: not significant, *: $0.05 \geq p > 0.01$, **: $0.01 \geq p > 0.001$, ***: $p \leq 0.001$ after Bonferroni correction for multiple comparisons). The quantity of analyses of each strategy that report the respective statistical result of the indicated pairwise comparison (effect, x-axis) at a level of $p \leq 0.05$ are given below each pairwise comparison in the corresponding color coding. In total, we performed all analyses with: $N_{\text{expert models}} = 20$, $N_{\text{consensus models}} = 36$, $N_{\text{consensus ensembles}} = 9$. Number of analyzed mice (N) for all analyzed subregions: $N_H = 7$, $N_{C-} = 7$, $N_{C+} = 6$. Numbers of analyzed images (n) are given for each analyzed subregion. Statistical data are available in ***Figure 4 - Source data***. (**A**) Analyses of cFOS-positive nuclei in stratum pyramidale of CA1. $n_H = 36$, $n_{C-} = 32$, $n_{C+} = 28$. (**B**) Analyses of cFOS-positive nuclei in stratum pyramidale of CA3. $n_H = 35$, $n_{C-} = 31$, $n_{C+} = 28$. (**C**) Analyses of cFOS-positive nuclei in the granule cell layer of the whole DG. $n_H = 35$, $n_{C-} = 31$, $n_{C+} = 27$. (**D**) Analyses of cFOS-positive nuclei in the granule cell layer of the suprapyramidal blade of the DG. $n_H = 35$, $n_{C-} = 31$, $n_{C+} = 27$. (**E**) Analyses of cFOS-positive nuclei in the granule cell layer of the infrapyramidal blade of the DG. $n_H = 35$, $n_{C-} = 31$, $n_{C+} = 27$. (**F**) Reliability of bioimage analysis results are assessed as *variation per effect* (left side) and *variation per model* (right side). For the *variation per effect*, the standard deviation of reported effect sizes was calculated (η^2) within each DL-based strategy for each of the 30 pairwise comparisons. Consensus ensembles show significantly lower standard (std.) deviations of η^2 per pairwise comparison compared to alternative strategies ($X^2(2) = 26.472$, $p < 0.001$, $N_{\text{effects}} = 30$, Kruskal-Wallis ANOVA followed by pairwise Mann-Whitney tests with Bonferroni correction, *: $p < 0.05$, ***: $p < 0.001$). For the *variation per model*, the standard deviation of centered η^2 across all pairwise comparisons was calculated for each individual model and ensemble (y-axis). In addition, the number of deviations from the congruent majority vote (at $p \leq 0.05$ after Bonferroni correction for multiple comparisons) were determined for each individual model and ensemble across all pairwise comparisons (x-axis). 2D plotting of the data reveals that consensus ensembles show the highest reliability of all three DL-based strategies.

expert models, by contrast, yield a high variation in effect sizes at different significance levels. Interestingly, all four expert models trained on the annotations of expert 1 (and two other expert models only in the case of H vs. C+) did not yield a significant increase, indicating that the subjective bias of expert 1 specific models also affects the bioimage analysis results (***Figure 4A***).

The meta analysis discloses a context-dependent increase of cFOS in almost all analyzed hippocampal regions (***Figure 4A-D***), except for the infrapyramidal blade of the dentate gyrus (***Figure 4E***). Notably, the majority votes of all three strategies at a significance level of $p \leq 0.05$ (after Bonferroni correction for multiple comparisons) are identical for each pairwise comparison (***Figure 4A-E***). However, the results can vary between individual models or ensembles (***Figure 4A-E***).

In order to assess the reliability of bioimage analysis results of the individual strategies, we further examined the variation per effect and variation per model in ***Figure 4F***. For the variation per effect, we calculated the standard deviation of reported effect sizes within each strategy for every pairwise comparison (effect). This confirmed the visual impression from ***Figure 4A-E*** as the consensus ensembles yield a significantly lower standard deviation compared to both alternative strategies (***Figure 4F***). To illustrate the variation per model, we show the interaction between the number of biological effects that the corresponding model (or ensemble) reported differently compared to the congruent majority votes versus the standard deviation of its centered effect sizes across all 30 analyzed effects. This analysis shows that no expert model detected all biological effects in the microscopy images that were defined by the majority votes of all models. This is in stark contrast to the consistency of effect interpretation across the consensus ensembles (***Figure 4F***).

Consequently, we conclude that the consensus ensemble strategy is best suited to satisfy the bioimaging standards for objectivity, reliability, and validity.

Applicability of consensus ensemble strategy for the bioimage analysis of external data sets

Bioimage analysis of fluorescent labels comes with a huge variability in terms of investigated model organisms, analyzed fluorescent features and applied image acquisition techniques (***Meijering et al., 2016***). In order to assess our consensus ensemble strategy across these varying parameters, we tested it on four external datasets that were created in a fully independent manner and according to individual protocols (*Lab-Mue, Lab-Inns1, Lab-Inns2* and *Lab-Wue2*; see methods and ***Figure S8***). Due to limited dataset sizes, the lab-specific training datasets consisted of just five microscopy images each and the corresponding est. GT based on the annotations from multiple experts. In the biomedical research field, the limited availability of training data is a common problem when training DL algorithms. For this reason, extensive data augmentation and regularization techniques, as well as transfer learning strategies are widely used to cope with small datasets (***Ronneberger et al., 2015***; ***Christiansen et al., 2018***; ***Falk et al., 2019***). Transfer learning is a technique that enables DL models to reuse the image feature representations learned on another source, such as a task (e.g. image segmentation) or a domain (e.g. the fluorescent feature, here cFOS-positive nuclei). This is particularly advantageous when applied to a task or domain where limited training data is available (***Yosinski et al., 2014***; ***Oquab et al., 2014***). Moreover, transfer learning might be used to reduce observer variability and to increase feature annotation objectivity (***Bayramoglu and Heikkilä, 2016***). There are typically two ways to implement transfer learning for DL models, either by fine-tuning or by freezing features (i.e., model weights) (***Yosinski et al., 2014***). The latter approach, if applied to the same task (e.g., image segmentation), does not require any further model training. These *out-of-the-box* models reduce time and hardware requirements and may further increase objectivity of image analysis, by altogether excluding the need for any additional manual input.

Consequently we hypothesized that transfer learning from pretrained model ensembles would substantially reduce the training efforts (***Falk et al., 2019***) and might even increase objectivity of bioimage analysis. To test this, we followed three different initialization variants of the consensus ensemble strategy (***Figure 5A***). In addition to starting the training of DL models with randomly ini-

tialized weights (***Figure 5A*** - *from scratch*), we reused the consensus ensemble weights from the previous evaluation (*Lab-Wue1*) by means of fine-tuning (***Figure 5A*** - *fine-tuned*) and freezing of all model layers (***Figure 5A*** - *frozen*). Although no training of the *frozen* model is required, we tested and evaluated the performance of *frozen* models to ensure their validity. After performing the similarity analysis, we compared the full bioimage analyses, including quantification and hypothesis testing, of the different initialization variants. Finally, to establish a notion of external validity, we also compared these results with the manually and independently performed bioimage analysis of a lab-specific expert (***Figure 5***, ***Figure S9***, and ***Figure S10***).

Dataset characteristics

The first dataset (*Lab-Mue*) represents very similar image parameters compared to our original *Lab-Wue1* dataset (***Figure 5C*** - *Lab-Mue* and ***Figure S8***). Mice experienced restraint stress and subsequent Pavlovian fear conditioning (cue-conditioning, tone-footshock association) and the number of cFOS-positive cells in the paraventricular thalamus (PVT) was compared between early (eRet) and late (lRET) phases of fear memory retrieval. In the context of transfer learning, this dataset originates from a very similar domain and requires the same task (image segmentation). Another two external datasets are focused on the quantification of cFOS abundance (similar domain), albeit showing less similarity in image parameters to our initial dataset (***Figure S9***, ***Figure S10*** and ***Figure S8***). In *Lab-Inns1*, mice underwent Pavlovian fear conditioning and extinction in the same context. The image dataset of *Lab-Inns2* shows cFOS immunoreactivity in the infralimbic cortex (IL) following fear renewal, meaning return of extinguished fear in a context different from the extinction training context. Since heterogeneity in this behavioral response was observed, mice were classified as responders (Resp) or non-responders (nResp), based on freezing responses (see methods). The image dataset of *Lab-Wue2* shows the least similarity of image parameters to the dataset of *Lab-Wue1*. This dataset represents another commonly used model organism in neurobiology, the zebrafish. Here, cell bodies of specific neurons (GABAergic neurons) instead of nuclei were fluorescently labeled (***Figure 5C*** - *Lab-Wue2* and ***Figure S8***). Hence, this dataset originates from a different domain but was acquired using the same technique.

Similarity analysis

As only limited training data was available we executed the similarity analysis for all external datasets by means of a *k-fold cross-validation*. We observed that the inter-rater variability differed between laboratories and different experts but remained comparable as previously for *Lab-Wue1* (***Figure 5D***, ***Figure S11***, and ***Figure S12***. Both *from scratch* and *fine-tuned* initiation variants resulted in individual consensus models that reached human expert level performance (***Figure 5D***, ***Figure S9***, ***Figure S10***). However, models adapted from pretrained weights yielded a higher validity in terms of similarity to the estimated ground truth. They either exceeded the maximal $M_{\text{F1 score}}$ reached by *from scratch* models (***Figure 5D*** - *Lab-Mue*, ***Figure S9***, ***Figure S10***) or reached them after less training iterations (***Figure 5D*** - *Lab-Wue2*). As expected, the performance of *frozen Lab-Wue1*-specific consensus models was highly dependent on the image similarity between the original and the new dataset. Consequently, the *out-of-the-box* segmentation performance of the *frozen Lab-Wue1* models was very poor on dissimilar images (***Figure 5D*** - *Lab-Wue2*), but we found it to be on par with human experts and adapted models on images that are highly similar to the original dataset (***Figure 5D*** - *Lab-Mue* - very similar domain and the same task).

Bioimage analysis results

To further strengthen the validity of our workflow, we compared all DL-based bioimage analyses to the manual analysis of a human expert from the individual laboratory (***Figure 5E***, ***Figure S9***, ***Figure S10***, and ***Table 1***).

For *Lab-Mue*, the bioimage analyses of all DL-based approaches, including the *frozen* consensus models and ensembles pretrained on *Lab-Wue1*, revealed a significantly higher number of cFOS-

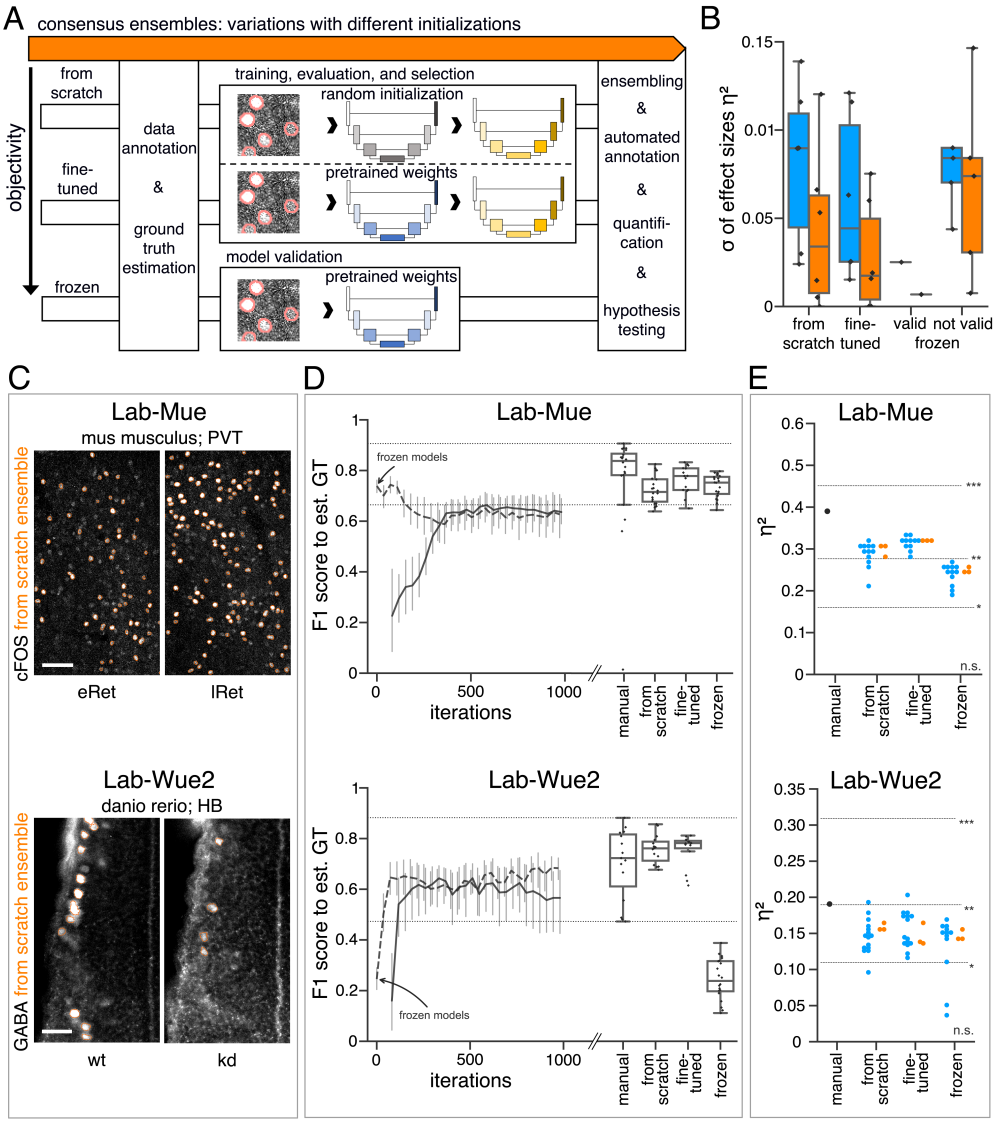


Figure 5. Consensus ensembles for DL-based feature annotation in external bioimage data sets. (**A**) Schematic overview depicting three initialization variants for creating consensus ensembles on new datasets. Data annotation by multiple human experts and subsequent ground truth estimation are required for all three initialization variants. In the *from scratch* variant, a U-Net model with random initialized weights is trained on pairs of microscopy images and estimated ground truth annotations. This variant was used to create consensus ensembles for the initial *Lab-Wue1* dataset. Alternatively, the same training dataset can be used to adapt a U-Net model with pretrained weights by means of transfer-learning (*fine-tuned*). In both variants, models are evaluated and selected on base of a validation set after model training. In a third variant, U-Net models with pretrained weights can be evaluated directly on a validation dataset, without further training (*frozen*). In all three variants, consensus ensembles of the respective models are then used for bioimage analysis. (**B**) Overall reliability of bioimage analysis results of each variant assessed as variation per effect. *Frozen* consensus models and ensembles were further subdivided as *valid* or *not valid* depending on whether they reached human expert performance on the validation dataset or not. In all three strategies, valid consensus ensembles (orange) showed lower standard deviations than consensus models (blue). $N_{\text{pairwise comparisons}} = 6$; $N_{\text{consensus models}} = 15$, and $N_{\text{consensus ensembles}} = 3$ for each variant. (C-E) Detailed comparison of the two external datasets with highest (*Lab-Mue*) and lowest (*Lab-Wue2*) similarity to *Lab-Wue1*. (**C**) Representative microscopy images. Orange: representative annotations of a lab-specific *from scratch* consensus ensemble. PVT: para-ventricular nucleus of thalamus, eRet: early retrieval, lRet: late retrieval, HB: hindbrain, wt: wildtype, kd: *gad1b* knock-down. Scale bars: *Lab-Mue* 100 µm and *Lab-Wue2* 6 µm. (**D**) Mean $M_{\text{F1 score}}$ of *from scratch* (solid line) and *fine-tuned* (dashed line) consensus models on the validation dataset over the course of training (iterations). Mean $M_{\text{F1 score}}$ of *frozen* consensus models are indicated with arrows. Box plots show the $M_{\text{F1 score}}$ among the annotations of human experts as reference and the mean $M_{\text{F1 score}}$ of selected consensus models. Two dotted horizontal lines mark the whisker ends of the $M_{\text{F1 score}}$ among the human expert annotations. (**E**) Effect sizes of all individual bioimage analyses (black: manual experts, blue: consensus models, orange: consensus ensembles). Three horizontal lines separate the significance intervals (n.s.: not significant, *: $0.05 \geq p > 0.01$, **: $0.01 \geq p > 0.001$, ***: $p \leq 0.001$). *Lab-Mue*: $N_{\text{consensus models}} = 12$ and $N_{\text{consensus ensembles}} = 3$ for all initialization variants. $N_{\text{eRet}} = 4$, $N_{\text{lRet}} = 4$; $n_{\text{eRet}} = 12$, $n_{\text{lRet}} = 12$. *Lab-Wue2*: $N_{\text{from scratch consensus models}} = 15$, $N_{\text{fine-tuned consensus models}} = 15$, $N_{\text{frozen consensus models}} = 12$, and $N_{\text{consensus ensembles}} = 3$ for each initialization variant. $N_{\text{wt}} = 5$, $N_{\text{kd}} = 4$, $n_{\text{wt}} = 20$, $n_{\text{kd}} = 15$. Statistical data can be found in ***Figure 5 - Source data***. Figure 5 - Supplementary figures: ***Figure S8***, ***Figure S9***, ***Figure S10***, ***Figure S11***, ***Figure S12***, and ***Figure S13***.

Table 1. Bioimage analyses results of external datasets. Data are based either on manual analysis or on annotations by a consensus ensemble. The results are given for the individual consensus ensemble initialization variants (*from scratch, fine-tuned*, and *frozen*). Results of *frozen* ensembles are only given if they reached human expert level performance (valid). P-values of *Lab-Inns2* are corrected for multiple comparisons using Bonferroni correction. μ_1: mean group 1, μ_2: mean group 2, U: U-statistic, eRet: early retrieval, lRet: late retrieval, Ctrl: control, Ext: extinction, Sal: saline, Res: L-DOPA/MS-275 responder, nRes: L-DOPA/MS-275 non-responder, wt: wildtype, kd: *gad1b* knock-down.

Lab	groups	initialization variant	μ_1	μ_2	U	significance level (p)	η^2
Mue	eRet ~ lRet	manual	1.00	1.65	19.0	** (0.0024)	0.39
		from scratch	1.00	1.70	25.0	** (0.0073)	0.31
		fine-tuned	1.00	1.68	24.0	** (0.0061)	0.32
		frozen	1.00	1.50	30.0	* (0.0166)	0.25
Inns1	Ctrl ~ Ext	manual	1.00	3.92	10.0	** (0.0047)	0.43
		from scratch	1.00	2.26	13.0	* (0.0101)	0.35
		fine-tuned	1.00	1.85	14.0	* (0.0128)	0.33
Inns2	Sal ~ Resp	manual	1.00	1.83	5.0	** (0.0021)	0.59
		from scratch	1.00	1.96	0.0	*** (0.0006)	0.71
		fine-tuned	1.00	2.07	0.0	*** (0.0006)	0.71
	Sal ~ nResp	manual	1.00	1.05	27.0	n.s. (1.0000)	0.00
		from scratch	1.00	1.63	8.0	n.s. (0.1299)	0.29
		fine-tuned	1.00	1.42	12.0	n.s. (0.3774)	0.16
	Res ~ nRes	manual	1.83	1.05	42.0	n.s. (0.1299)	0.29
		from scratch	1.96	1.63	41.0	n.s. (0.1728)	0.26
		fine-tuned	2.07	1.42	42.0	n.s. (0.1299)	0.29
Wue2	wt ~ kd	manual	1.00	0.28	227.5	* (0.0101)	0.19
		from scratch	1.00	0.45	220.0	* (0.0205)	0.16
		fine-tuned	1.00	0.37	216.0	* (0.0290)	0.14

positive cells in the PVT of mice 24h after fear conditioning (lRET), which was confirmed by the manual expert analysis (***Figure 5E*** - *Lab-Mue*, ***Table 1***). Yet again, the formation of model ensembles increased the reproducibility of results by yielding less or almost no variation in the effect sizes (***Figure 5E*** - *Lab-Mue*).

The manual expert analysis of the *Lab-Inns1* dataset revealed a significantly higher number of cFOS-positive nuclei in the basolateral amygdala (BLA) after extinction of a previously learned fear, which was also reliably detected by all consensus ensembles, regardless of initiation variant (***Figure S9***, ***Table 1***). However, this significant difference was only present in the analyses of most individual consensus models, both *from scratch* and *fine-tuned* (***Figure S9***). Again, this could be attributed to a higher variability between the effect sizes of individual models, compared to a higher homogeneity among ensembles (***Figure S9***).

For *Lab-Inns2*, the manual expert analysis as well as all deep-learning based approaches that were adapted to the *Lab-Inns2* dataset show increased numbers of cFOS-positive cells in the infralimbic cortex of L-DOPA/MS-275 responders (Resp) compared to control (Sal) mice (***Figure S10***, ***Table 1***). However, in L-DOPA/MS-275 non-responders (nResp), we did not observe a significant increase of cFOS-positive nuclei (***Figure S10***, ***Table 1***). Furthermore, the high effect sizes of the comparison between L-DOPA/MS-275 responders and non-responders further indicate that the differences observed in the behavioral responses of Resp and nResp mice were also reflected in the abundance of cFOS in the infralimbic cortex (***Figure S10***, ***Table 1***).

Manual expert analysis of the fourth external dataset revealed a significantly lower amount

of GABA-positive somata in *gad1b* knock-down zebrafish, compared to wildtypes (***Figure 5E*** - *Lab-Wue2*, ***Table 1***). Again, this effect was reliably detected by all deep-learning based approaches that included training on the *Lab-Wue2*-specific training dataset and the effect sizes of ensembles showed less variability (***Figure 5E*** - *Lab-Wue2*). Despite its poor segmentation performance and hence, poor validity, this effect was also present in the bioimage analysis of the *frozen* consensus models and ensembles pretrained on *Lab-Wue1* (***Figure 5E*** - *Lab-Wue2*).

As with our initial dataset, we assessed reliability by calculating the variation per effect as the standard deviation of the reported effect sizes within each group and pooled these results across all external datasets. Consistent with the higher reliability of *from scratch* and *fine-tuned* ensemble annotations (***Figure S13***), this analysis shows that the formation of model ensembles reduced the variation per effect in both variants, compared to the respective individual models (***Figure 5B***). For *frozen* models and ensembles, we further distinguished between valid (perform with on par with human experts; *Lab-Mue* only) and not valid (perform worse than human experts; *Lab-Inns1, Lab-Inns2, Lab-Wue2*).

In summary, we assessed the reproducibility of our consensus ensemble strategy by using four external datasets. These datasets were acquired with different image acquisition techniques, investigate two common model organisms, and analyze the two main cellular compartments (nuclei and somata) at varying resolutions (***Figure S8***). In-line with the results obtained on our initial dataset, we observed an increased reproducibility for the consensus ensembles compared to individual consensus models after training on all four external datasets (***Figure 5B***). Moreover, our data also suggests that pretrained consensus models can even be deployed *out-of-the-box*, but only when carefully validated. Thus, sharing pretrained model weights across different laboratories reduces lab-specific biases within the bioimage analysis and may further increase objectivity and validity.

Ultimately, we conclude that our proposed ensemble consensus workflow is reproducible for different data sets and laboratories and increases objectivity, reliability and validity of DL-based bioimage analyses.

Discussion

The present study contributes to bridging the gap between “methods" and “biology" oriented studies in image feature analysis (***Meijering et al., 2016***). We explored the potentials and limitations of DL models utilizing the general quality criteria for quantitative research: objectivity, reliability, and validity. Thereby, we put forward an effective but easily implementable strategy that aims to establish reproducible, DL-based bioimage analysis within the life science community.

The number of DL-based tools for bioimage annotations and their accessibility for non-AI specialists is gradually increasing (***McQuin et al., 2018***; ***Haberl et al., 2018***; ***Falk et al., 2019***). DL models are proven to hold advantages over conventional algorithms (***Caicedo et al., 2019***) and are thus likely to become the new standard for bioimage analysis tasks throughout the life sciences. Usually, the performance of new bioimage analysis tools or methods is assessed by means of similarity measures to a certain ground truth (***Ronneberger et al., 2015***; ***McQuin et al., 2018***; ***Haberl et al., 2018***; ***Falk et al., 2019***; ***Caicedo et al., 2019***). However, this is rarely sufficient to establish trust in the use of DL models for bioimage analysis, as the vast amount of parameters and flexibility to adapt DL models to virtually any task renders them prone to internalize biases (***von Chamier et al., 2019***). This is particularly true in the case of fluorescent feature analysis in bioimage datasets, as an objective ground truth is not available. In conjunction with the stochastic training process, this is a very critical point, because it holds the potential for intended or unintended tampering similar to p-hacking (***Head et al., 2015***), e.g., by training DL models until non-significant results become significant.

To investigate the effects of DL-based strategies on the bioimage analysis of fluorescent features, we acquired a typical bioimage dataset (*Lab-Wue1*) and five experts manually annotated corresponding ROIs (here cFOS-positive nuclei) in a representative subset of images. Then, we

tested three DL-based strategies for automatized feature segmentation. DL models were either trained on the manual annotations of a single expert (expert models) or on the input of multiple experts pooled by ground truth estimation (consensus models). In addition, we formed ensembles of consensus models (consensus ensembles).

Similarity analysis of fluorescent feature annotation

In accordance with previous studies, similarity analyses revealed a substantial level of inter-rater variability in the heuristic annotations of the single experts (***Schmitz et al., 1999***; ***Collier et al., 2003***; ***Niedworok et al., 2016***). Furthermore, we confirmed the concerns already put forward by others (***Falk et al., 2019***; ***von Chamier et al., 2019***) that training of DL models solely on the input of a single human expert imposes a high risk of incorporating an intrinsic bias into the trained models. We therefore conclude that models trained on single expert annotations resemble an automation of manual image annotation, but cannot remove subjective biases from bioimage analyses. Importantly, only consensus ensembles led to a coincident significant increase also in the reliability and validity of fluorescent feature annotations. Our analyses also show that annotations of multiple experts are imperative for two reasons: first, they mitigate or even eliminate the bias of expert-specific annotations and, secondly, are essential for the assessment of the model performance.

Reproducibility and validity of bioimage analyses

Our bioimage dataset from *Lab-Wue1* enabled us to look at the impact of different DL-based strategies on the results of bioimage analyses. This revealed a striking model-to-model variability as the main factor impairing the reproducibility of DL-based bioimage analyses. Convincingly, the majority votes for each effect were identical for all three strategies. However, the variance within the reported effect sizes differed significantly for each strategy. This entailed, for example, that no expert model was in full agreement with the congruent majority votes. On the contrary, consensus ensembles detected all effects with significantly higher reliability. Thus, our data indicates that bioimage analysis performed with a consensus ensemble significantly reduces the risk of obtaining irreproducible results.

Evaluation of consensus ensembles on external datasets

We then tested our consensus ensemble approach and three initialization variants on four external datasets with limited training data and varying similarities in terms of image parameters to our original dataset (*Lab-Wue1*). In line with previous studies on transfer learning, we demonstrate that the adaptation of models from pretrained weights to new, yet similar data requires less training iterations, compared to the training of models *from scratch* (***Falk et al., 2019***). We extend these analyses and show that the reliability of *fine-tuned* ensembles was at least equivalent to *from scratch* ensembles, if not higher. Furthermore, we also provide initial evidence that pretrained ensembles can be used even without any adaptation, if task similarity is sufficiently high. This removes any additional, potentially biased manual input and could further increase the objectivity of bioimage analyses.

Potentials of open source pretrained consensus ensemble libraries

Sharing model weights from validated models in open-source libraries, similarly to *TensorFlow Hub* (https://www.tensorflow.org/hub) or *PyTorch Hub* (https://pytorch.org/hub/), offers a great opportunity to provide annotation experience across labs in an open science community. In this study, for instance, we used the nuclear label of cFOS, an activity-dependent transcription factor, as fluorescent feature of interest. This 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), apotposis markers (Caspase-3), and multiple others. Similarly to the pretrained and shared models of ***Falk et al.*** (***2019***), we surmise that the learned feature representations (i.e., model weights) of our cFOS consensus ensembles may serve as a

good initialization for models that aim at performing nucleosomatic fluorescent label segmentation in brain slices.

In line with the results of the *Kaggle Data Science Bowl 2018* (***Caicedo et al., 2019***), however, our findings indicate that a model adapted to a specific data set usually outperforms a general model trained on different datasets from different domains. To use and share frozen *out-of-the-box* models across the science community, we therefore need to create a well documented library that contains validated model weights for each specific task and domain (e.g., for each organism, marker type, image resolution, etc.). In conjunction with data repositories, this would also allow retrospective data analysis of prior studies.

In summary, open source model libraries may contribute to a better reproducibility of scientific experiments (***Fanelli, 2018***) by reducing the subjective bias in bioimage analysis, by offering openness to analysis criteria, and by sharing pretrained models for (re-)evaluation.

Limitations

The characteristics of our examined strategies are based on the best practices in the field of DL and derived from extant literature (***Meijering et al., 2016***; ***Falk et al., 2019***; ***Caicedo et al., 2019***). Nevertheless, we do not claim universal applicability. Our decisions were driven by our expectation of the likelihood to be deployed in the future or even become the new standard in bioimage analysis. Thus, one important limitation of our study is the focus on the U-Net model architecture (***Ronneberger et al., 2015***). In distinction to conventional studies that introduce a new method and provide a comprehensive performance comparison to the state of the art, we chose U-net as the well-established and widely studied de facto standard for biomedical image segmentation purposes (***McQuin et al., 2018***; ***Falk et al., 2019***; ***Caicedo et al., 2019***). Similarly, we chose to use STAPLE (***Warfield et al., 2004***) as the benchmark procedure for ground truth estimation over its alternatives and variants (***Lampert et al., 2016***). In addition, we tried different ways to incorporate the single expert annotations into one DL model. For instance, we followed the approach of ***Guan et al.*** (***2018***) by modeling individual experts in a multi-head DL model instead of pooling them in the first place. However, we decided to discard the approach as our tests did not yield superior results and scaled up the complexity. With the emergence of improved or more reliable algorithms, the need for some of our proposed concepts, e.g., the formation of model ensembles, may vanish. Still, this paper can serve as a blueprint for any evaluation of DL models along the general quality criteria further on.

Accessibility of our workflow and pretrained consensus ensembles

To enable other researchers to easily access, to interact with, and to reproduce our results and to share our trained models, we provide an open-source *Python* library that is easily accessible for both local installation or cloud-based deployment. With *Jupyter Notebooks* becoming the computational notebook of choice for data scientists (***Perkel, 2018***) we also implemented a training pipeline for non-AI experts in a *Jupyter Notebook* optimized for Google Colab, providing free access to the required computational resources (e.g., GPUs and TPUs). In summary, we recommend to use the annotations of multiple human experts to train and evaluate DL consensus model ensembles. In such a way, DL can be used to increase the objectivity, reliability, and validity of bioimage analyses and pave the way for higher reproducibility in science.

Acknowledgments

We thank all members of our collaborative research center (SFB-TRR58) for all the fruitful discussions and the continuous support over the course of this project. We thank Toni Greif for critically reviewing the mathematical content.

Declaration of interests

All authors declare no conflict of interest.

Author contributions

Dennis Segebarth: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization, Project administration
Matthias Griebel: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization, Project administration
Nikolai Stein: Conceptualization, Methodology, Validation, Writing - original draft, Visualization, Project administration
Cora R. von Collenberg: Investigation, Validation, Data curation, Writing - review & editing
Corinna Martin: Validation, Data curation, Writing - review & editing
Dominik Fiedler: Methodology, Validation, Investigation, Data curation, Writing - review & editing
Lukas B. Comeras: Methodology, Validation, Investigation, Data curation, Writing - review & editing
Anupam Sah: Methodology, Validation, Investigation, Data curation, Writing - review & editing
Victoria Schöffler: Methodology, Validation, Investigation, Data curation, Writing - review & editing
Theresa Lüffe: Methodology, Validation, Investigation, Data curation, Writing - review & editing
Alexander Dürr: Software, Writing - review & editing
Rohini Gupta: Data curation, Writing - review & editing, Funding acquisition
Manju Sasi: Data curation, Writing - review & editing, Funding acquisition
Christina Lillesaar: Methodology, Validation, Resources, Data curation, Writing - review & editing, Project administration, Funding acquisition
Maren D. Lange: Methodology, Validation, Investigation, Resources, Data curation, Writing - review & editing, Project administration, Funding acquisition
Ramon O. Tasan: Methodology, Validation, Resources, Writing - review & editing, Project administration, Funding acquisition
Nicolas Singewald: Methodology, Validation, Resources, Writing - review & editing, Project administration, Funding acquisition
Hans-Christian Pape: Methodology, Resources, Writing - review & editing, Project administration, Funding acquisition
Christoph M. Flath: Conceptualization, Methodology, Validation, Resources, Writing - original draft, Supervision, Project administration, Funding acquisition
Robert Blum: Conceptualization, Methodology, Validation, Resources, Writing - original draft, Supervision, Project administration, Funding acquisition

Funding information

The following researchers were supported by the Deutsche Forschungsgemeinschaft SFB-TRR58 – Fear, Anxiety and Anxiety Disorders: Robert Blum project A10, Hans-Christian Pape project A03 and Maren D. Lange project B08. Rohini Gupta and Manju Sasi were supported by fellowships of the Graduate School of Life Sciences (GSLS) Würzburg. Ramon O. Tasan was supported by the Austrian Science Fund (FWF) P29952 & P25851. Nicolas Singewald was supported by the FWF (Austrian Science Fund, I2433-B26, DK W-1206 and SFB F4410). Christina Lillesaar was supported by IZKF (Interdiziplinäres Zentrum für Klinische Forschung, Universitätsklinikum Würzburg, project number N-320).

Materials and Methods

Data sets regarding animal behavior, immunofluorescence analysis and image acquisition were performed in five 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-Mue*: Institute of Physiology I, University of Münster, 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
- *Lab-Wue1*: Institute of Clinical Neurobiology, University Hospital, Würzburg, Germany
- *Lab-Wue2*: Department of Child and Adolescent Psychiatry, Center of Mental Health, University Hospital of Würzburg, Würzburg, Germany

Contacts 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).

Resource table

Key reagents and resources are listed with their source and identifiers in ***Table 2***.

Experimental models

Mice

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 für 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^{TM} IVC cages (1284L Eurostandard Type II L: 365 x 207 x 140 mm, floor area cm^2 530, Tecniplast Deutschland GmbH, Hohenpeißenberg, 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 $\pm$ 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 für Wissenschaft Forschung und Wirtschaft, Kommission für Tierversuchsangelegenheiten) approved all experimental procedures.

Lab-Wue1: 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, Würzburg, 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 Würzburg, 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

Table 2. Key resources and reagents.

Reagent or resource	Source	Identifier
Antibodies		
Lab-Mue, Lab-Inns2: rabbit anti-cFOS	Santa Cruz	Cat# sc-52, RRID: AB_2106783
Lab-Inns1: mouse-anti-Parvalbumin	Sigma-Aldrich	Cat# P3088, RRID: AB_477329
Lab-Inns1: rabbit anti-cFOS	Millipore	Cat# PC38, RRID: AB_2106755
Lab-Wue1: guinea-pig anti-NeuN	Synaptic Systems	Cat# 266004, RRID: AB_2619988
Lab-Wue1: mouse anti-Parvalbumin	Swant	Cat# PV235, RRID: AB_10000343
Lab-Wue1: rabbit anti-cFOS	Synaptic Systems	Cat# 226003, RRID: AB_2619946
Lab-Wue2: rabbit anti-GABA	Sigma-Aldrich	Cat#A2025, RRID: AB_477652
Experimental models: organism / strains		
Lab-Mue, Lab-Inns1: Mouse: C57BL/6J	Charles River	Cat# CRL:027, RRID:IMSR_CRL:27
Lab-Inns2: Mouse: 129S1/SvImJ (S1)	Charles River	RRID:MGI:5658424
Lab-Wue1: Mouse: C57BL/6J	Jackson Laboratory	Cat# JAX:000664, RRID: IMSR_JAX:000664
Lab-Wue2: Zebrafish: AB/AB		
Software and algorithms		
Lab-Mue, Lab-Inns2, Lab-Wue1, Lab-Wue2: Fiji (ImageJ)	Fiji	https://fiji.sc/, RRID:SCR_002285
Lab-Inns1: Improvision Openlab software (5.5.0)	Perkin Elmer	http://www.perkinelmer.com/pages/020/cellularimaging/products/openlab.xhtml, RRID:SCR_012158
Lab-Inns1: Prism 7.0	Graphpad	https://www.graphpad.com/scientific-software/prism/, RRID: SCR_015807
Lab-Inns2: CellSens Dimension Desktop 1.9 software	Olympus	https://www.olympus-lifescience.com/en/software/cellsens/, RRID: SCR_016238
Lab-Wue1: Fluoview FV10-ASW	Olympus	https://www.photonics.com/Product.aspx?PRID=47380, RRID:SCR_014215
TensorFlow	***Abadi et al.*** (***2016***)	https://www.tensorflow.org, RRID: SCR_016345
Keras	***Chollet*** (***2015***)	https://keras.io
ImageJ	***Rueden et al.*** (***2017***)	https://imagej.net/, RRID: SCR_003070
SciPy	***Jones et al.*** (***2001***)	https://www.scipy.org, RRID: SCR_008058
scikit-learn	***Pedregosa et al.*** (***2011***)	https://scikit-learn.org/
scikit-image	***van der Walt et al.*** (***2014***)	https://scikit-image.org/
Pingouin	***Vallat*** (***2018***)	https://pingouin-stats.org
simpleITK	***Lowekamp et al.*** (***2013***)	https://simpleitk.org/
Code and data for this paper	This paper	https://github.com/matjesg/bioimage_analysis

age of 8-12 weeks during the subjective day-phase of the animals and were randomly allocated to experimental groups.

Zebrafish

Lab-Wue2: Zebrafish (*Danio rerio*) embryos of the AB/AB strain were cultivated at 28°C with a 14/10 h light/dark cycle in Danieau's solution containing 0.2 mM phenylthiocarbamide to prevent pigmentation. The embryos were staged according to ***Kimmel et al.*** (***1995***). To knock-down expression of *gad1b*, fertilized eggs were injected with 500 μM of a *gad1b* splice blocking morpholino targeting the exon 8/intron 8 boundary of *gad1b* (Ensembl, GRCz11). Morpholino sequence: 5'tttgtgatcagtt-

taccaggtgaga3' (Gene Tools). The efficiency of the morpholino was tested by reverse transcription PCR on DNase I treated total RNA collected from 24 hours post fertilization and 5 days post fertilization embryos. Sanger sequencing showed that the morpholino causes a partial inclusion of intron 8, which generates a premature stop codon.

Mouse behavior

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 ***Chauveau et al.*** (***2012***). 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^- (2.5 kHz tone, 85 dB, stimulus duration 10 s, inter-stimulus interval 20 s; inter-trial interval 6 h). On the next day, fear conditioning was performed in two sessions of three randomly presented CS^+ (10 kHz tone, 85 dB, stimulus duration 10 s, randomized inter-stimulus interval 10-30 s; inter-session interval 6 h), each of which was co-terminated with an unconditioned stimulus (scrambled foot shock of 0.4 mA, duration 1 s). 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 10 s, inter-stimulus interval 20 s) without receiving foot shocks. Afterwards, mice remained in this context for another 2 min before being returned to their homecages.

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: Cued fear conditioning, extinction and extinction retrieval was carried out as previously described ***Whittle et al.*** (***2016***). Context dependence of fear extinction memories was assessed using a fear renewal tests in a novel context ***Murphy et al.*** (***2017***). 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 gray floor and black walls, cleaned with a 100% ethanol and illumi-

nated 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 ***Gruene et al.*** (***2015***). 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. In the treatment group, mice which displayed freezing levels 2 times above than the standard deviation of the mean were classified as non-responders.

Contextual fear conditioning

Lab-Wue1: 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 (***Fanselow, 1980***). The percentage time spent freezing was calculated by dividing the amount of time spent in the training chamber.

Drug treatment

MS-275 (Entinostat, Selleck Chemicals, Vienna, Austria; 10 mg kg^{-1} dissolved in saline +25% dimethylsulfoxide vehicle) was administered immediately (<1 min) following an extinction training session and L-DOPA (Sigma-Aldrich, Vienna, Austria; 20 mg kg^{-1} dissolved in saline) was administered 1 h before an extinction training session. All drugs were administered intraperitoneally in a volume of 10 ml kg^{-1} body weight. Control animals received saline. Mice were randomly selected to be administered either vehicle or pharmacological compound (***Whittle et al., 2016***).

Immunohistochemistry and microscopy

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°C until 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. Transcardial 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°C and 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°C until 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 (***Paxinos and Franklin, 2004***) 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°C for 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°C with 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 20x oil-objective (UPlanSApo, Olympus) at 488nm excitation.

Lab-Wue1 To analyze anti-cFOS 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 naïve learning controls (homecage; H). 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 ml/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°C. The 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 (***Paxinos and Franklin*** (***2004***)) 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 mg / 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 H).

Lab-Wue2: For immunohistochemistry of whole-mount specimens, the embryos were fixed at 30 hours post fertilization in 4% PFA at 4°C over night. The specimens were subsequently washed 3x10 min in PBS with 0.1% Tween-20 (PBST) and then once for 5 min in 150 mM Tris-HCl buffer (pH 9.0). The solution was exchanged for fresh Tris-HCl buffer, and the embryos were incubated for 15 min at 70°C, cooled down to room temperature and then washed 2x5 min in PBST. To further increase permeability, the embryos were first rinsed quickly two times in ice-cold dH_2O and then incubated with pre-cooled acetone for 20 min at -20°C. The acetone was quickly washed off with dH_2O, and then with PBS containing 0.8% Triton X100 (PBSTX) for 2x5 min. The specimens were

incubated at room temperature for 1 h in blocking buffer (PBSTX with 10% normal sheep serum and 2% bovine serum albumin) and subsequently with the primary antibody (rabbit-anti-GABA, Sigma-Aldrich A2052, diluted 1:400 in blocking buffer) at 4°C for 3 days with gentle shaking. After extensive washes in PBSTX, the embryos were incubated with the secondary antibody (goat-anti-rabbit AlexaFluor488, Invitrogen, Thermo Fisher Scientific, diluted 1:1,000 in blocking buffer) at 4°C for 2 days with gentle shaking. Finally, the embryos were washed extensively in PBST, transferred and stored in 80% glycerol in PBST at 4°C until imaging. Before microscopy, the yolk was removed and the embryos were mounted in 80% glycerol to be imaged from the dorsal side. Confocal image acquisition was performed using a Nikon eclipse C1 laser scanning microscope with a Plan Apo VC 20x/0.75 DIC N2 objective and a Coherent Saphire 488 nm laser. All specimens were imaged using NIS Elements software (Nikon) with the same acquisition settings. Images with 2048x2048 pixels were taken as 12 bit z-stacks with a step-size of 2.5 µm, covering the whole region, including the dorsal-ventral dimension, of the hindbrain that contains GABA immunoreactive cells.

Image processing and manual analysis

Lab-Mue: Images were adjusted in brightness and contrast using ImageJ. One expert from Lab-Mue manually segmented the paraventricular thalamus (PVT) and quantified cFOS-positive cells in the PVT for bioimage analysis. For the training and validation of DL models, cFOS-positive ROIs in five additional images were manually segmented by four experts from *Lab-Wue1* (expert 1 and experts 3-5) using ImageJ. All experts were blinded to another and 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 by a neuroscientist of Lab-Inns1 using the free shape tool of the Improvision Openlab software (PerkinElmer). Boundaries were projected to the respective cFOS-immunoreactive image and cFOS-positive nuclei were counted manually inside that area by the expert of Lab-Inns1. For the training and validation of DL models, cFOS-positive ROIs in five additional images were manually segmented by four experts from *Lab-Wue1* (expert 1 and experts 3-5) using ImageJ. All experts were blinded to another and the treatment condition.

Lab-Inns2: The anatomical localization of cells within the infralimbic cortex was aided by using illustrations in a stereotaxic atlas ***Franklin and Paxinos*** (***2008***), published anatomical studies ***Van De Werd et al.*** (***2010***) and former studies in S1 mice ***Fitzgerald et al.*** (***2013***); ***Whittle et al.*** (***2010***). All analyses were done in a comparable area under similar optical and light conditions. For manual analysis, an expert of Lab-Inns2 viewed the digitized images on a computer screen using CellSens Dimension 1.5 software (Olympus Corporation, Tokyo, Japan) and evaluated cFOS-positive nuclei within the infralimbic cortex, the brain region of the interest. For the training and validation of DL models, cFOS-positive ROIs in five additional images were manually segmented by four experts from *Lab-Wue1* (expert 1 and experts 3-5) using ImageJ. All experts were blinded to another and the treatment condition.

Lab-Wue1: For image preprocessing, 12-bit gray-scale confocal image z-stacks were projected (maximum intensity) and converted to 8-bit. Five expert neuroscientists from *Lab-Wue1* manually segmented cFOS-positive nuclei and Parvalbumin-positive somata as regions of interest (ROIs) in a total of 45 images (36 training and nine test images). The NeuN immunoreactive granule cell layer of the dentate gyrus and the pyramidal cell layer in CA1 and CA3 were annotated as NeuN-positive ROIs. The NeuN-positive areas used for the quantifications of cFOS-positive cells were identical for all analyses and segmented manually by one human expert. All experts were blinded to another and the treatment condition.

Lab-Wue2: The confocal z-stacks with 12-bit gray-scale images were imported into ImageJ (***Schneider et al., 2012***). An expert of *Lab-Wue2* manually counted GABA-positive somata in every 4th section of each confocal z-stack, covering the entire hindbrain region housing GABAergic neurons ($N_{untreated\ controls}$ = 5, $N_{morphants}$ = 4). For the training and validation of DL models, GABA-positive ROIs

in five additional images were manually segmented by three experts from *Lab-Wue2* (experts 6-8) using ImageJ. All experts were blinded to another and the treatment condition.

Ground truth estimation

In absence of a rigorously objective ground truth, we considered the expert segmentation maps to calculate the expectation-maximization algorithm for simultaneous truth and performance level estimation (STAPLE). Given a collection of segmentation masks, the algorithm computes a probabilistic estimate of the ground truth (est. GT) segmentation as well as a performance measure for each segmentation (***Warfield et al., 2004***). We implemented the algorithm using the simplified interface to the Insight Toolkit (SimpleITK, version 1.2.4) (***Lowekamp et al., 2013***).

Evaluation metrics

All evaluation metrics were calculated using Python (version 3.7.3), SciPy (version 1.4.1), and scikit-image (version 0.16.2). We provide the source code in *Jupyter Notebooks* (see Data and software availability).

Segmentation and detection

Following ***Caicedo et al.*** (***2019***) we based our evaluation on identifying segmentation and detection similarities on object-level (ROI-level). In a segmentation mask, we define an object as a set of pixels that were horizontally, vertically, and diagonally connected (8-connectivity). We only considered ROIs at a biologically justifiable size, depending on the data set characteristics. We approximated the minimum size based on the smallest area that was annotated by a human expert (*Lab-Mue*: 30px, *Lab-Inns1*: 16px, *Lab-Inns2*: 60px, *Lab-Wue1*: 30px, *Lab-Wue2*: 112px).

To compare the segmentation similarity between a source and a target segmentation mask, we first computed the intersection-over-union (IoU) score for all pairs of objects. The IoU, also known as Jaccard similarity, of two sets of pixels $a = \{1, ..., A\}$ and $b = \{1, ..., B\}$ is defined as the size of the intersection divided by the size of the union:

$$M_{\text{IoU}}(a, b) := \frac{|a \cap b|}{|a \cup b|}$$

Second, we used the pairwise IoUs to match the objects of each mask. We solved the assignment problem by maximizing the sum of IoUs by means of the Hungarian Method (***Kuhn, 1955***). This ensures an optimal matching of objects in the case of ambiguity, i.e., overlap of one source object with one or more targets object. We reported the segmentation similarity of two segmentation masks by calculating the arithmetic mean of M_{IoU} over all matching objects:

$$\bar{M}_{\text{IoU}} = \frac{1}{N} \sum_{i=1}^{N} M^i_{\text{IoU}}$$

where $i \in \{0, ..., N\}$ is an assigned match and N denotes number of matching objects. By this definition, the *Mean IoU* only serves as a measure for the segmentation similarity of matching objects and neglects objects that do not overlap at all.

To address this issue, we additionally calculated measures to account for the detection similarity. Therefore, we define a pair of objects with an IoU is above a threshold t as correctly detected (true positive - TP). Objects that match with an IoU at or below t or have no match at all are considered to be false negative (FN) for the source mask and false positive (FP) for the target mask. This allows us to calculate the Precision $M_{\text{Precision}}$, Recall M_{Recall}, and F1 score $M_{\text{F1 score}}$ as the harmonic mean of $M_{\text{Precision}}$ and M_{Recall}:

$$M_{\text{Precision}}(t) := \frac{TP(t)}{TP(t) + FP(t)}$$

$$M_{\text{Recall}}(t) := \frac{TP(t)}{TP(t) + FN(t)}$$

$$M_{\text{F1 score}}(t) := 2 \cdot \frac{\mathrm{M}_{\text{Precision}}(\mathrm{t}) \cdot \mathrm{M}_{\text{Recall}}(\mathrm{t})}{\mathrm{M}_{\text{Precision}}(\mathrm{t}) + \mathrm{M}_{\text{Recall}}(\mathrm{t})}$$

with $t \in [0, 1]$ as a fixed IoU threshold. If not indicated differently, we used $t = 0.5$ in our calculations.

Inter-rater reliability

To quantify the reliability of agreement between different annotators we calculated Fleiss' κ (***Fleiss and Cohen, 1973***). In contrast to the previously introduced metrics, Fleiss' κ accounts for the agreement that would be expected by chance. For a collection of segmentation masks of the same image, each object (ROI) $i \in \{1, ..., N\}$ is assigned to a class $j \in \{0, ..., K\}$. Here, N denotes the total number of unique objects (ROIs) and K the number of categories ($K = 1$ for binary segmentation). Then, n_{ij} represents the number of annotators who assigned object i to class j. We again leveraged the IoU metric to match the ROIs from different segmentation masks above a given threshold $t \in [0, 1]$. Following ***Fleiss and Cohen*** (***1973***) we define the proportion of all assignments for each class:

$$p_j(t) := \frac{1}{Nd} \sum_{i=1}^{N} n_{ij}(t)$$

where d denotes the count of the annotators. We define the extent to which the annotators agree on the i-th object as

$$P_i(t) := \frac{1}{d(d-1)} \sum_{j=1}^{K} n_{ij}(t) \left(n_{ij}(t) - 1\right)$$

Subsequently, we define the mean of the $P_i(t)$ as

$$\bar{P}(t) := \frac{1}{N} \sum_{i=1}^{N} P_i(t)$$

and

$$\bar{P}_e(t) := \sum_{j=1}^{K} p_j(t)^2$$

Finally, Fleiss' κ at a given threshold t is defined as

$$\kappa(t) := \frac{\bar{P}(t) - \bar{P}_e(t)}{1 - \bar{P}_e(t)}$$

where $1 - \bar{P}_e(t)$ denotes the degree of agreement attainable above chance and $\bar{P}(t) - \bar{P}_e(t)$ the actually achieved agreement in excess of chance. To allow a better estimate of the chance we randomly added region proposals of class $j = 0$ (background). If not indicated differently, we use $t = 0.5$ in our calculations.

Deep learning approach

The deep learning pipeline was implemented in Python (version 3.7.3), Tensorflow (version 1.14.0), Keras (version 2.2.4), scikit-image (version 0.16.2), and scikit-learn (version 0.21.2). We provide the source code in *Jupyter Notebooks* (see Data and software availability).

Network Architecture

We instantiated all DL models with a U-Net architecture (***Ronneberger et al., 2015***), a fully convolutional neural network for semantic segmentation. 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 (***Ronneberger et al., 2015***). We adopted the model hyperparameters (e.g., hidden layers, activation functions, weight initialization) from ***Falk et al.*** (***2019***) as these are extensively tested and evaluated on different biomedical data sets. The layers of the U-Net architecture are logically grouped into an encoder and a decoder (***Figure S1***). Following ***Falk et al.*** (***2019***) the VGG-like encoder consists of five convolutional modules. Each

module comprises two convolution layers with no padding, each followed by a leaky ReLU with a leakage factor of 0.1 and a max-pooling operation with a stride of two. The last module, however, does not contain the max-pooling layer and constitutes the origin of the decoder. The decoder consists of four (up-) convolutional modules. Each of these modules comprises of a transposed convolution layer (also called up- or deconvolution), a concatenate layer for the corresponding cropped encoder feature map, and two convolution layers. Again, each layer is followed by a leaky ReLU with a leakage factor of 0.1. The final layer consists of a 1 x 1 convolution with a softmax activation function. The resulting (pseudo-) probabilities allow a comparison to the target segmentation mask using cross-entropy on pixel level. Unless indicated differently, we used a kernel size of 3 x 3. To allow faster convergence during training we included batch normalization layers (***Ioffe and Szegedy, 2015***) after all (up-) convolutions below the first level. By this, an unnormalized path from the input features to the output is remaining to account for absolute input values, e.g., the brightness of fluorescent labels.

Weighted soft-max cross-entropy loss

Fluorescent microscopy images typically exhibit more background than fluorescent features of interest. To control the impacts of the resulting class imbalance we implemented a pixel-weighted softmax cross-entropy loss. Thus, we compute the loss from the raw score (logits) of the last 1x1 convolution without applying the softmax. As proposed by ***Falk et al.*** (***2019***) we define the weighted cross entropy loss for an input image I as

$$L_{\text{wce}}(I) := -\sum_{x\in\Omega} w(x) \log \frac{\exp\left(\hat{y}_{y(x)}(x)\right)}{\sum_{k=0}^{K} \exp\left(\hat{y}_k(x)\right)}$$

where x is a pixel in image domain Ω, $w : \Omega \to \mathbb{R}_{\geq 0}$ the pixel-wise weight map, $y : \Omega \to \{0, ..., K\}$ the target segmentation mask, $\hat{y}_k : \Omega \to \mathbb{R}$ the predicted score for class $k \in \{0, ..., K\}$, and K the number of classes ($K = 1$ for binary classification). Consequently, $\hat{y}_{y(x)}(x)$ is the predicted score for the target class $y(x)$ at position x.

Similar to ***Falk et al.*** (***2019***) we compose the weight map w from two different weight maps w_{bal} and w_{sep}. The former allows to mitigate the class imbalances by decreasing the weight of background pixels by the factor $v_{\text{bal}} \in [0, 1]$. We add a smoothly decreasing Gaussian function at the edges of the foreground objects accordingly and define

$$w_{bal}(x) := \begin{cases} 1 & y(x) > 0 \\ v_{\text{bal}} + \left(1 - v_{bal}\right) \exp\left(-\frac{d_1^2(x)}{2\sigma_{\text{bal}}^2}\right) & y(x) = 0 \end{cases}$$

where $d_1(x)$ denotes the distance to the closest foreground object and σ_{bal} the standard deviation of the Gaussian function.

By definition, semantic segmentation performs a pixel-wise classification and is unaware of different object instances (ROIs). Following ***Falk et al.*** (***2019***) we the force learning of the different instances by increasing the weight of the separating ridges. We estimate the width of a ridge by adding d_1 (distance to nearest ROI) and d_2 (distance to second nearest ROI) at each pixel. We define

$$w_{\text{sep}}(x) := \exp\left(-\frac{\left(d_1(x) + d_2(x)\right)^2}{2\sigma_{\text{sep}}^2}\right)$$

where σ_{sep} defines the standard deviation of the decreasing Gaussian function. The final weight map is given by

$$w := w_{\text{bal}} + \lambda w_{\text{sep}}$$

where $\lambda \in \mathbb{R}_{\geq 0}$ allows to control the focus on instance separation. We used the following parameter set in our experiments: $\lambda = 50$, $v_{bal} = 0.1$, $\sigma_{\text{bal}} = 10$ and $\sigma_{\text{sep}} = 6$.

Tile sampling and augmentation.

Given limited training data availability, we leveraged effective data augmentation techniques for biomedical images as proposed by (***Falk et al., 2019***). These comprise transformations and elastic deformations by means of a random deformation field. To become invariant to the input sizes (image shapes) we leveraged the overlap tile strategy introduced by (***Ronneberger et al., 2015***). Thus, images of any size can be processed. Both data augmentation and overlap tile strategy were adopted from a Tensorflow implementation of (***Falk et al., 2019***). We used an input tile shape of 540x540x1 (height x width x channels) and a corresponding output tile shape of 356x356x1 for all our experiments.

Training, evaluation and model selection

We trained, evaluated and selected all deep learning models for our different strategies – *expert models*, *consensus models*, *consenus ensembles* – following the same steps:

1. Splitting the data into train and validation set (random stratified sampling)
2. Determining the learning rate using the *learning rate finder* (***Smith, 2018***)
3. Training the model with cyclical learning rates according to the *fit-one-cycle* policy of ***Smith (2018)***
4. Selecting the model with the highest $M_{\text{F1 score}}$ on the validation set (post-hoc evaluation)

We used the annotations from individual experts to train the *expert models* and the consensus annotations (est. GT) for the *consensus models* and *consensus ensembles*. To ensure the validity of the consensus approaches, we only considered valid models for further use and evaluation. We defined all models as valid that yielded a higher $M_{\text{F1 score}}$ to the estimated ground truth than the "worst" human expert on each validation image. For the *consensus ensemble* approach, we repeated the steps above according to the principle of a *k-fold cross-validation*. We ensembled the resulting k models (invalid models excluded) by averaging the softmax predictions. The post-hoc evaluation on the validation set was performed using the saved model weights (checkpoints) from each epoch. For the similarity analysis we converted the model output (pixel-wise softmax score) to a segmentation mask by assigning each pixel to the class with the highest softmax score.

We trained all models on a NVIDIA GeForce GTX 1080 TI with 11 GB GDDR5X RAM using the Adam optimizer (***Kingma and Ba, 2014***) and a mini-batch size of four. If not indicated differently, the initial weights were drawn from a truncated normal distribution (***He et al., 2015***). We chose the appropriate maximum learning rates according to the learning rate finder (step two). For *Lab-Wue1* we selected a maximum learning rate of 4e-4 and a minimum learning rate of 4e-5 over a training cycle length of 972 iterations within $k = 4$ validation splits. For *Lab-Mue, Lab-Inns1, Lab-Inns2* and *Lab-Wue2* we chose a maximum learning rate of 1e-4 and a minimum learning rate of 1e-5 over a training cycle length of 972 iterations within $k = 5$ validation splits.

Transfer learning

To implement transfer learning we adapted the training procedure from above. For the *fine-tuning* approach we initialized the weights from the *consensus models* of *Lab-Wue1* and performed all steps for model training, evaluation and selection. For the *frozen* approach we also initialized the weights from the *consensus models* of *Lab-Wue1* but skipped steps two (finding a learning rate) and three (model training). Hence, we did not adjust the model weights to the new training data. Hardware and training hyperparameters remained unchanged.

Quantification of fluorescent features

Fluorescent features were analyzed on base of the binary segmentation masks derived from the output of DL models or model ensembles, or counted manually by lab-specific experts. In order to compare the number of fluorescent features across images, we normalized in each image the number of annotated fluorescent features to the area of the analyzed region (e.g. the number

of cFOS-positive features per NeuN-positive area for *Lab-Wue1*). For one set of experiment, we pooled this data for each condition (e.g. H, C- and C+ for *Lab-Wue1*) and the analyzed brain region (e.g. whole DG, infrapyramidal DG, suprapyramidal DG, CA3, or CA1 for *Lab-Wue1*). To compare different sets of experiments with each other, we normalized all relative fluorescent feature counts to the mean value of the respective control group (e.g. H for *Lab-Wue1*).

The mean signal intensity for each image was calculated as the mean signal intensity of all ROIs annotated within the analyzed NeuN-positive region (only performed for *Lab-Wue1*). Subsequent pooling steps were identical as described above for the count of fluorescent features.

Statistical analysis

All statistical analyses were performed using Python (version 3.7.3), SciPy (version 1.4.1), and Pingouin (version 0.3.4). We provide all source datasets and source codes in *Jupyter Notebooks* (see Data and software availability). In box plots, the area of the box represents the interquartile range (IQR, 1st to 3rd quartile) and whiskers extend to the maximal or minimal values, but no longer than $1.5 \times$ IQR.

Statistical analysis of fluorescent feature quantifications

All DL-based quantifications of fluorescent features were tested for significant outliers (Grubb's test). Individual images were excluded from the analysis if they were detected as significant outlier in the majority of DL-based quantification results. Throughout all bioimage analyses, N represents the number of investigated animals and n the number of analyzed images. Normality (Shapiro-Wilk) and homogeneity of variance (Levene´s) were tested for all DL-based quantification results. For comparison of multiple quantifications of the same image dataset, non-parametric statistical tests were applied to all bioimage analyses. This ensured comparability of the results. To compare two experimental conditions (*Lab-Mue, Lab-Inns1,* and *Lab-Wue2*), Mann-Whitney-U tests were used. In case of three experimental conditions (*Lab-Wue1* and *Lab-Inns2*), Kruskal-Wallis-ANOVA followed by Mann-Whitney-U tests with Bonferroni correction for multiple comparisons was applied.

Effect size calculation

Effect sizes (η^2) were calculated for each pairwise comparison. First, the Z-statistic (Z) was calculated from the U-statistic (U) of the Mann-Whitney-U test as:

$$Z = \frac{U - \frac{n_1 \cdot n_2}{2}}{\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}}$$

where n_1 and n_2 are the numbers of analyzed images of the two compared groups, group 1 and group 2, respectively. Following ***Rosenthal and DiMatteo*** (***2002***), η^2 was calculated as:

$$\eta^2 = \frac{Z^2}{n_1 + n_2}$$

Furthermore, the three critical values of η^2 that mark the borders between the four significance levels (e.g. for p = 0.05, p = 0.01, and p = 0.001 for a pairwise comparison without Bonferroni correction for multiple comparisons) were calculated from the chi-square distribution.

All other statistical analyses

Data was tested for normal distribution (Shapiro-Wilk) and homoscedasticity (Levene´s) and parametric or non-parametric tests were used accordingly, as reported in the figure legends (parametric: one-way ANOVA, followed by T-tests (or Welchs T-test for unequal sample sized) with Bonferroni correction for multiple comparisons; non-parametric: Kruskal-Wallis ANOVA, followed by Mann-Whitney tests with Bonferroni correction for multiple comparisons).

Table 3. List of supplementary figures.

Main Figure	Supplementary Figure	Caption
Figure 1	***Figure S1***	U-Net Architecture
	Figure S2	Illustration of bioimage dataset *Lab-Wue1*
Figure 2	***Figure S3***	Extended subjectivity analysis
	Figure S4	Ensemble size and reliability
	Figure S5	Extended similarity analysis: F1 score
	Figure S6	Extended similarity analysis: mean IoU
Figure 3	***Figure S7***	Behavioral analysis *Lab-Wue1*
Figure 5	***Figure S8***	Characteristics of all five bioimage datasets
	Figure S9	Performance of consensus ensembles on feature annotation in image dataset *Lab-Inns01*
	Figure S10	Performance of consensus ensembles on feature annotation in image dataset *Lab-Inns02*
	Figure S11	Expert similarity across all datasets: F1 score
	Figure S12	Expert similarity across all datasets: mean IoU
	Figure S13	Reliability of the consensus approaches across different datasets

Data and software availability

Data

We provide the full bioimage datasets of *Lab-Wue1* and *Lab-Mue*, including microscopy images, segmentation masks of all DL models and ensembles, and annotations of analyzed regions of interest. For all five bioimage datasets, we provide the quantification results of the bioimage analyses for all models and ensembles as source data. Likewise, the results of the behavioral analysis of *Lab-Wue1* are available as source data. As indicated in the respective figure legends, we also provide all statistical data in full detail as source data files. Furthermore, we provide all training and test datasets that were created in the course of this study. These include all microscopy images with the corresponding manual expert annotations and estimated ground-truth annotations. As part of our proposed pipeline we share the trained and validated consensus ensembles within our open-source model library. All data can be accessed at our GitHub repository (https://github.com/matjesg/bioimage_analysis).

Software

We provide all source code as Jupyter Notebooks in our GitHub repository (https://github.com/matjesg/bioimage_analysis). This includes the code for the bioimage analyses, all statistical analyses, and our proposed pipeline to create, use, and share consensus ensembles for fluorescent feature annotations.

List of supplementary figures

A list of all supplementary figures is provided in ***Table 3***.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:160304467. 2016; .

Baker M. Reproducibility crisis? Nature. 2016; 533(26):353–66.

Bayramoglu N, Heikkilä J. Transfer learning for cell nuclei classification in histopathology images. In: *European Conference on Computer Vision* Springer; 2016. p. 532–539.

Caicedo JC, Goodman A, Karhohs KW, Cimini BA, Ackerman J, Haghighi M, Heng CK, Becker T, Doan M, McQuin C, Rohban M, Singh S, Carpenter AE. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. Nature Methods. 2019; 16(12):1247–1253. http://dx.doi.org/10.1038/s41592-019-0612-7, doi: 10.1038/s41592-019-0612-7.

Campeau S, Falls W, Cullinan W, Helmreich D, Davis M, Watson S. Elicitation and reduction of fear: behavioural and neuroendocrine indices and brain induction of the immediate-early gene c-fos. Neuroscience. 1997; 78(4):1087–1104.

von Chamier L, Laine RF, Henriques R. Artificial intelligence for microscopy: what you should know. Biochemical Society Transactions. 2019; 47(4):1029–1040.

Chauveau F, Lange MD, Jungling K, Lesting J, Seidenbecher T, Pape HC. Prevention of Stress-Impaired Fear Extinction Through Neuropeptide S Action in the Lateral Amygdala. Neuropsychopharmacology. 2012; 37:1588–1599. doi: http://www.nature.com/npp/journal/v37/n7/suppinfo/npp20123s1.html.

Chen X, Dallmeier-Tiessen S, Dasler R, Feger S, Fokianos P, Gonzalez JB, Hirvonsalo H, Kousidis D, Lavasa A, Mele S, et al. Open is not enough. Nature Physics. 2019; 15(2):113–119.

Chollet F. Keras. https://kerasio. 2015; .

Christiansen EM, Yang SJ, Ando DM, Javaherian A, Skibinski G, Lipnick S, Mount E, O'Neil A, Shah K, Lee AK, et al. In silico labeling: predicting fluorescent labels in unlabeled images. Cell. 2018; 173(3):792–803.

Collier DC, Burnett SSC, Amin M, Bilton S, Brooks C, Ryan A, Roniger D, Tran D, Starkschall G. Assessment of consistency in contouring of normal-tissue anatomic structures. Journal of applied clinical medical physics / American College of Medical Physics. 2003; 4(1):17–24. doi: 10.1120/jacmp.v4i1.2538.

Dietterich TG. Ensemble methods in machine learning. In: *International workshop on multiple classifier systems* Springer; 2000. p. 1–15.

Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, Böhm A, Deubner J, Jäckel Z, Seiwald K, Dovzhenko A, Tietz O, Dal Bosco C, Walsh S, Saltukoglu D, Tay TL, Prinz M, Palme K, Simons M, Diester I, et al. U-Net: deep learning for cell counting, detection, and morphometry. Nature Methods. 2019; 16(1):67–70. doi: 10.1038/s41592-018-0261-2.

Fanelli D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? Proceedings of the National Academy of Sciences. 2018; 115(11):2628–2631.

Fanselow MS. Conditioned and unconditional components of post-shock freezing. Pavlov J Biol Sci. 1980; 15(4):177–82. http://www.ncbi.nlm.nih.gov/pubmed/7208128.

Fitzgerald PJ, Whittle N, Flynn SM, Graybeal C, Pinard CR, Gunduz-Cinar O, Kravitz AV, Singewald N, Holmes A. Prefrontal single-unit firing associated with deficient extinction in mice. Neurobiol Learn Mem. 2013; http://www.ncbi.nlm.nih.gov/pubmed/24231425, doi: 10.1016/j.nlm.2013.11.002.

Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement. 1973; 33(3):613–619.

Frambach JM, van der Vleuten CP, Durning SJ. AM last page: Quality criteria in qualitative and quantitative research. Academic Medicine. 2013; 88(4):552.

Franklin KBJ, Paxinos G. The mouse brain in stereotaxic coordinates. compact 3. ed. Amsterdam [u.a.]: Elsevier; 2008.

Gallo FT, Katche C, Morici JF, Medina JH, Weisstaub NV. Immediate early genes, memory and psychiatric disorders: focus on c-Fos, Egr1 and Arc. Frontiers in behavioral neuroscience. 2018; 12:79.

Gruene TM, Flick K, Stefano A, Shea SD, Shansky RM. Sexually divergent expression of active and passive conditioned fear responses in rats. Elife. 2015; 4. https://www.ncbi.nlm.nih.gov/pubmed/26568307, doi: 10.7554/eLife.11352.

Guan MY, Gulshan V, Dai AM, Hinton GE. Who said what: Modeling individual labelers improves classification. In: *Thirty-Second AAAI Conference on Artificial Intelligence*; 2018. p. 3109–3118.

Guzowski JF, Setlow B, Wagner EK, McGaugh JL. Experience-dependent gene expression in the rat hippocampus after spatial learning: a comparison of the immediate-early genesArc, c-fos, and zif268. Journal of Neuroscience. 2001; 21(14):5089–5098.

Haberl MG, Churas C, Tindall L, Boassa D, Phan S, Bushong EA, Madany M, Akay R, Deerinck TJ, Peltier ST, Ellisman MH. CDeep3M—Plug-and-Play cloud-based deep learning for image segmentation. Nature Methods. 2018; 15(9):677–680. http://dx.doi.org/10.1038/s41592-018-0106-z, doi: 10.1038/s41592-018-0106-z.

He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE international conference on computer vision*; 2015. p. 1026–1034.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS biology. 2015; 13(3).

Huff NC, Frank M, Wright-Hardesty K, Sprunger D, Matus-Amat P, Higgins E, Rudy JW. Amygdala regulation of immediate-early gene expression in the hippocampus induced by contextual fear conditioning. Journal of Neuroscience. 2006; 26(5):1616–1623.

Hutson M, Artificial intelligence faces reproducibility crisis. American Association for the Advancement of Science; 2018.

Ioannidis JP. Why most clinical research is not useful. PLoS medicine. 2016; 13(6):e1002049.

Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167. 2015; .

Jones E, Oliphant T, Peterson P, et al. SciPy: Open source scientific tools for Python. http://wwwscipyorg/. 2001; .

Keiser AA, Turnbull LM, Darian MA, Feldman DE, Song I, Tronson NC. Sex Differences in Context Fear Generalization and Recruitment of Hippocampus and Amygdala during Retrieval. Neuropsychopharmacology. 2017; 42:397–407.

Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. Stages of embryonic development of the zebrafish. Developmental dynamics. 1995; 203(3):253–310.

Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014; .

Kuhn HW. The Hungarian method for the assignment problem. Naval research logistics quarterly. 1955; 2(1-2):83–97.

Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in neural information processing systems*; 2017. p. 6402–6413.

Lampert TA, Stumpf A, Gançarski P. An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. IEEE Transactions on Image Processing. 2016; 25(6):2557–2572.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977; p. 159–174.

LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015; 521(7553):436–444.

Li A, Gong H, Zhang B, Wang Q, Yan C, Wu J, Qian L, Zeng S, Luo Q. Micro-Optical Sectioning Tomography to Obtain a HIgh-Resolution Atlas of the Mouse Brain. Science. 2010; 1404(2010):1404–1408. http://www.sciencemag.org/content/330/6009/1404.short, doi: 10.1126/science.1191776.

Lindquist M. Neuroimaging results altered by varying analysis pipelines. Nature. 2020 May; https://doi.org/10.1038/d41586-020-01282-z, doi: 10.1038/d41586-020-01282-z.

Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The design of SimpleITK. Frontiers in neuroinformatics. 2013; 7:45.

Maška M, Ulman V, Svoboda D, Matula P, Matula P, Ederra C, Urbiola A, España T, Venkatesan S, Balak DM, et al. A benchmark for comparison of cell tracking algorithms. Bioinformatics. 2014; 30(11):1609–1617.

McDole K, Guignard L, Amat F, Berger A, Malandain G, Royer LA, Turaga SC, Branson K, Keller PJ. In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level. Cell. 2018; p. 1–18. https://linkinghub.elsevier.com/retrieve/pii/S0092867418312431, doi: 10.1016/j.cell.2018.09.031.

McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, Doan M, Ding L, Rafelski SM, Thirstrup D, Wiegraebe W, Singh S, Becker T, Caicedo JC, Carpenter AE. CellProfiler 3.0: Next-generation image processing for biology. PLOS Biology. 2018 jul; 16(7):e2005970. https://dx.plos.org/10.1371/journal.pbio.2005970, doi: 10.1371/journal.pbio.2005970.

Meijering E, Carpenter AE, Peng H, Hamprecht FA, Olivo-Marin JC. Imagining the future of bioimage analysis. Nature Biotechnology. 2016 dec; 34(12):1250–1255. http://www.nature.com/articles/nbt.3722, doi: 10.1038/nbt.3722.

Murawski NJ, Klintsova AY, Stanton ME. Neonatal alcohol exposure and the hippocampus in developing male rats: effects on behaviorally induced CA1 c-Fos expression, CA1 pyramidal cell number, and contextual fear conditioning. Neuroscience. 2012; 206:89–99.

Murphy CP, Li X, Maurer V, Oberhauser M, Gstir R, Wearick-Silva LE, Viola TW, Schafferer S, Grassi-Oliveira R, Whittle N, Huttenhofer A, Bredy TW, Singewald N. MicroRNA-Mediated Rescue of Fear Extinction Memory by miR-144-3p in Extinction-Impaired Mice. Biol Psychiatry. 2017; 81(12):979–989. https://www.ncbi.nlm.nih.gov/pubmed/28104225, doi: 10.1016/j.biopsych.2016.12.021.

Niedworok CJ, Brown APY, Jorge Cardoso M, Osten P, Ourselin S, Modat M, Margrie TW. AMAP is a validated pipeline for registration and segmentation of high-resolution mouse brain data. Nature Communications. 2016; 7(May):1–9. http://dx.doi.org/10.1038/ncomms11879, doi: 10.1038/ncomms11879.

Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2014. p. 1717–1724.

Osten P, Margrie TW. Mapping brain circuitry with a light microscope. Nature Methods. 2013; 10(6):515–523. doi: 10.1038/nmeth.2477.

Ounkomol C, Seshamani S, Maleckar MM, Collman F, Johnson GR. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. Nature methods. 2018; 15(11):917–920.

Paxinos G, Franklin KBJ. The mouse brain in stereotaxic coordinates. Compact 2nd ed. Amsterdam ; Boston: Elsevier Academic Press; 2004. http://www.loc.gov/catdir/description/els041/2003113435.htmlhttp://www.loc.gov/catdir/toc/els041/2003113435.html.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

Perkel JM. Why Jupyter is data scientists' computational notebook of choice. Nature. 2018; 563(7732):145–147.

Ramamoorthi K, Fropf R, Belfort GM, Fitzmaurice HL, McKinney RM, Neve RL, Otto T, Lin Y. Npas4 regulates a transcriptional program in CA3 required for contextual memory formation. Science. 2011; 334(6063):1669–1675.

Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention* Springer; 2015. p. 234–241.

Rosenthal R, DiMatteo MR. Meta-analysis. Stevens' handbook of experimental psychology. 2002; .

Rueden CT, Schindelin J, Hiner MC, DeZonia BE, Walter AE, Arena ET, Eliceiri KW. ImageJ2: ImageJ for the next generation of scientific image data. BMC Bioinformatics. 2017; 18(1):1–26. doi: 10.1186/s12859-017-1934-z.

Schmitz C, Korr H, Heinsen H. Design-based counting techniques: the real problems. Trends in neurosciences. 1999; 22(8):345.

Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012; 9(7):671–5. http://www.ncbi.nlm.nih.gov/pubmed/22930834.

Sezgin M, Sankur B. Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic imaging. 2004; 13(1):146–166.

Shuvaev SA, Lazutkin AA, Kedrov AV, Anokhin KV, Enikolopov GN, Koulakov AA. Dalmatian: An algorithm for automatic cell detection and counting in 3d. Frontiers in neuroanatomy. 2017; 11:117.

Siebert S, Machesky LM, Insall RH. Point of View: Overflow in science and its implications for trust. eLife. 2015 sep; 4:e10825. https://doi.org/10.7554/eLife.10825, doi: 10.7554/eLife.10825.

Smith LN. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:180309820. 2018; .

Tayler KK, Tanaka KZ, Reijmers LG, Wiltgen BJ. Reactivation of neural ensembles during the retrieval of recent and remote memory. Current Biology. 2013; 23(2):99–106.

Taylor BN, Kuyatt CE. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. . 1994; .

Vallat R. Pingouin: statistics in Python. The Journal of Open Source Software. 2018 Nov; 3(31):1026.

Van De Werd HJ, Rajkowska G, Evers P, Uylings HB. Cytoarchitectonic and chemoarchitectonic characterization of the prefrontal cortical areas in the mouse. Brain Struct Funct. 2010; 214(4):339–53. https://www.ncbi.nlm.nih.gov/pubmed/20221886, doi: 10.1007/s00429-010-0247-z.

van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, the scikit-image contributors. scikit-image: image processing in Python. PeerJ. 2014 6; 2:e453. https://doi.org/10.7717/peerj.453, doi: 10.7717/peerj.453.

Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE transactions on medical imaging. 2004; 23(7):903–921.

Whittle N, Hauschild M, Lubec G, Holmes A, Singewald N. Rescue of impaired fear extinction and normalization of cortico-amygdala circuit dysfunction in a genetic mouse model by dietary zinc restriction. J Neurosci. 2010; 30(41):13586–96. https://www.ncbi.nlm.nih.gov/pubmed/20943900, doi: 10.1523/JNEUROSCI.0849-10.2010.

Whittle N, Maurer V, Murphy C, Rainer J, Bindreither D, Hauschild M, Scharinger A, Oberhauser M, Keil T, Brehm C, et al. Enhancing dopaminergic signaling and histone acetylation promotes long-term rescue of deficient fear extinction. Translational psychiatry. 2016; 6(12):e974–e974.

Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: *Advances in neural information processing systems*; 2014. p. 3320–3328.

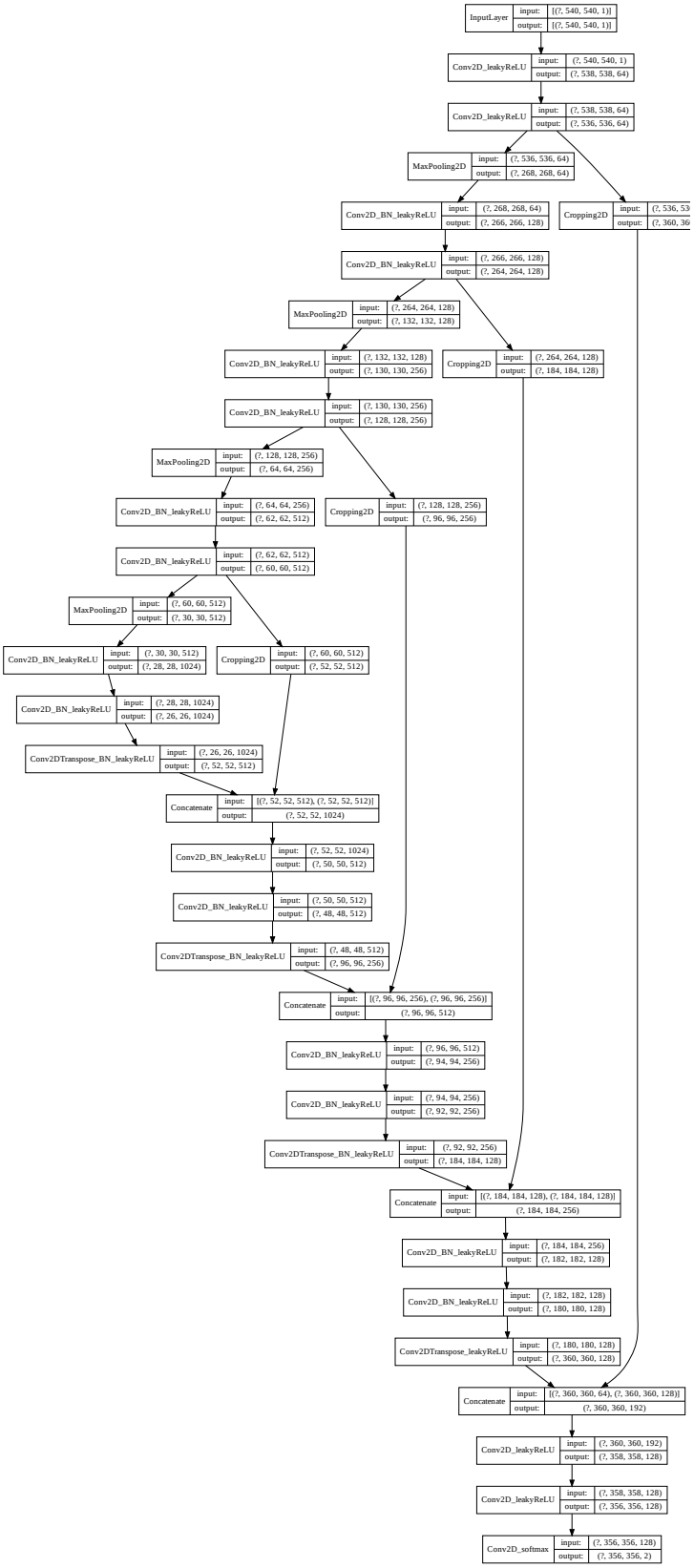
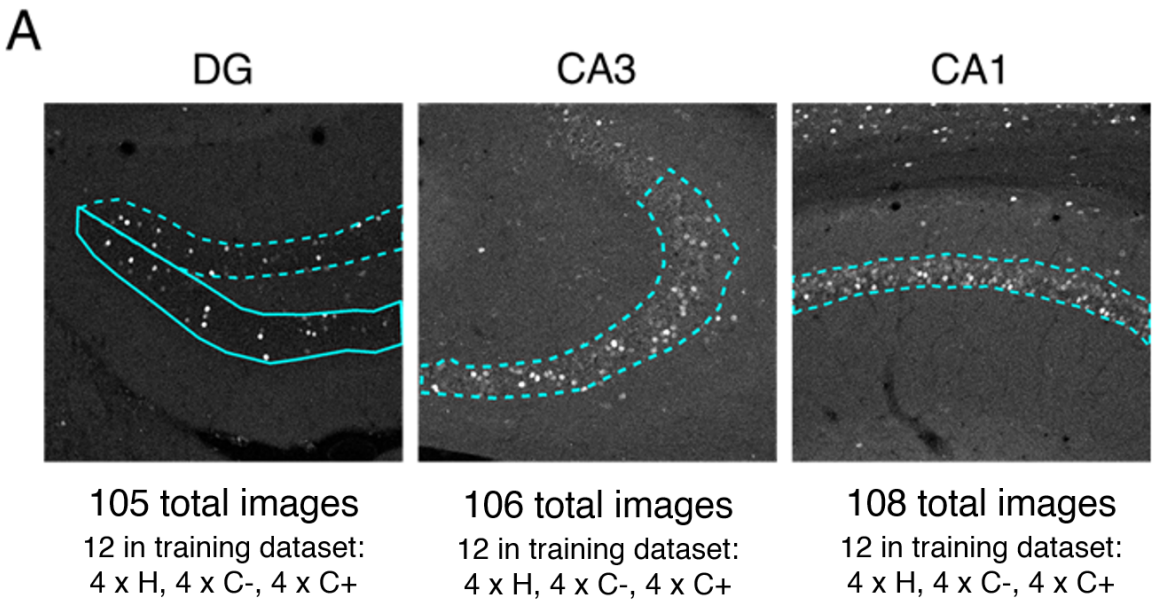
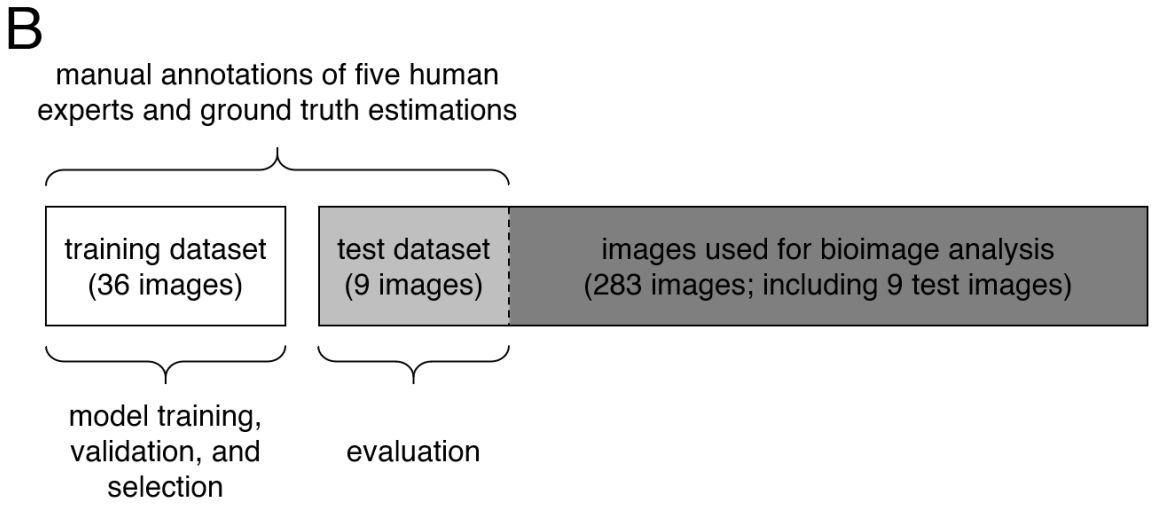


Figure S1. U-Net Architecture. The U-Net architecture was adapted from ***Falk et al.*** (***2019***). The input and output shapes denote (minibatch size x height x width x channels), where the mini-batch size is not specified. If applicable, the names of the layers describe the containing operations (e.g.: *Conv2D_BN_leakyReLU* represents a 2D convolution followed by a batch normalization layer and a leakyReLu activation function). All convolutional layers were instantiated with a kernel size of 3x3, a stride of one and, no padding, except for the last convolution (1x1 kernel). The leaky ReLU has a leakage factor of 0.1 and the max-pooling operation a stride of two. The up-convolution (Conv2DTranspose) has a kernel size of 2x2 and strides of two.





C

	all 283 images	mean per image
features annotated by cons. ensembles (within NeuN+ area)	~ 10,000	~ 35

Figure S2. Illustration of bioimage dataset *Lab-Wue1*. (**A**) A total of 319 images showing cFOS immunoreactivity in the dorsal hippocampus of mice was split up in 105 images of the Dentate gyrus, 106 images of CA3 and 108 images of CA1. To create a balanced training dataset, four images of each experimental condition were randomly selected (H, C-, C+) from each hippocampal subregion (DG, CA3, CA1; 4 x 3 x 3 = 36 images). (**B**) Five expert neuroscientists (experts 1-5) manually annotated cFOS-positive nuclei in the selected 36 images of the training dataset and in nine additional images (test dataset). The test images represented one image per region and condition (3 x 3). Annotation was performed independently and on different computers and screens. The training dataset was used to train either expert specific models (only annotations of a single expert were used) or consensus models (est. GT annotations computed from the annotations of all five experts were used). Using *k-fold cross-validation* during the training, we were able to test the model performance and to ultimately select only those models that reached human level performance. The final evaluation of all models was then performed on the additional nine images of the test dataset. For bioimage analyses, we used the remaining 274 images and the nine test images. (**C**) On average, each consensus ensemble annotated ~ 10,000 cFOS-positive feature within the NeuN-positive areas in all 283 images used for bioimage analysis, which is equivalent to ~ 35 features per image.

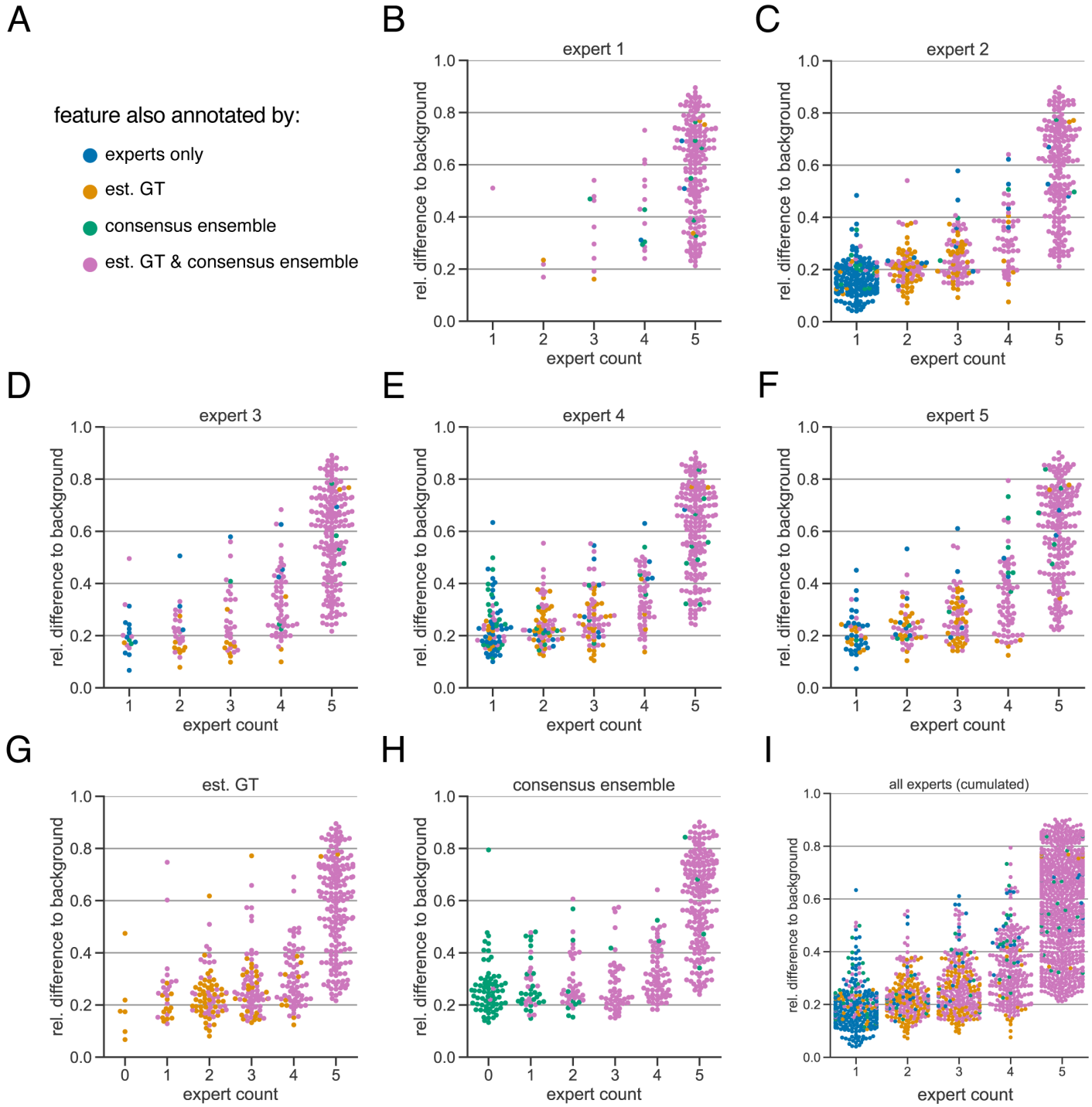


Figure S3. Extended subjectivity analysis. The subjectivity analysis depicts the relationship between the relative intensity difference of a florescent feature (ROI) to the background and the annotation count of human experts. A visual interpretation indicates that the annotation probability of a ROI is positively correlated with its relative relative intensity. The relative intensity difference is calculated as $\frac{\mu_{inner}-\mu_{outer}}{\mu_{inner}}$, where μ_{inner} is the mean signal intensity of the ROI and μ_{outer} the mean signal intensity of its nearby outer area. We considered matching ROIs at an IoU threshold of $t = 0.5$. The coder in the title of the respective plot was used to create the region proposals of the ROIs, i.e., the annotations served as origin for the other pairwise comparisons. (**A**) Legend of color codes: blue depicts that a ROI was only annotated by one or more human experts; yellow depicts the ROIs that were present in the estimated ground truth; green shows the ROIs that are only present in an exemplary consensus ensemble; pink depicts ROIs that are present in both estimated ground truth and consensus ensemble (**B-I**) All calculations are performed on the test set (n=9 images) which was withheld from model training and validation. (**B-F**) The individual expert analysis shows the effects of different heuristic evaluation criteria. (**G**) The analysis of the est. GT annotations reveals the limitations of the ground truth estimation algorithm, which is based on the human annotations. An expert count of zero can result from merging different ROIs. (**H**) The analysis of a representative consensus ensemble shows that human annotators may have missed several ROIs (green) even with a large relative difference to the background. (**I**) Cumulative summary of **B-F**.

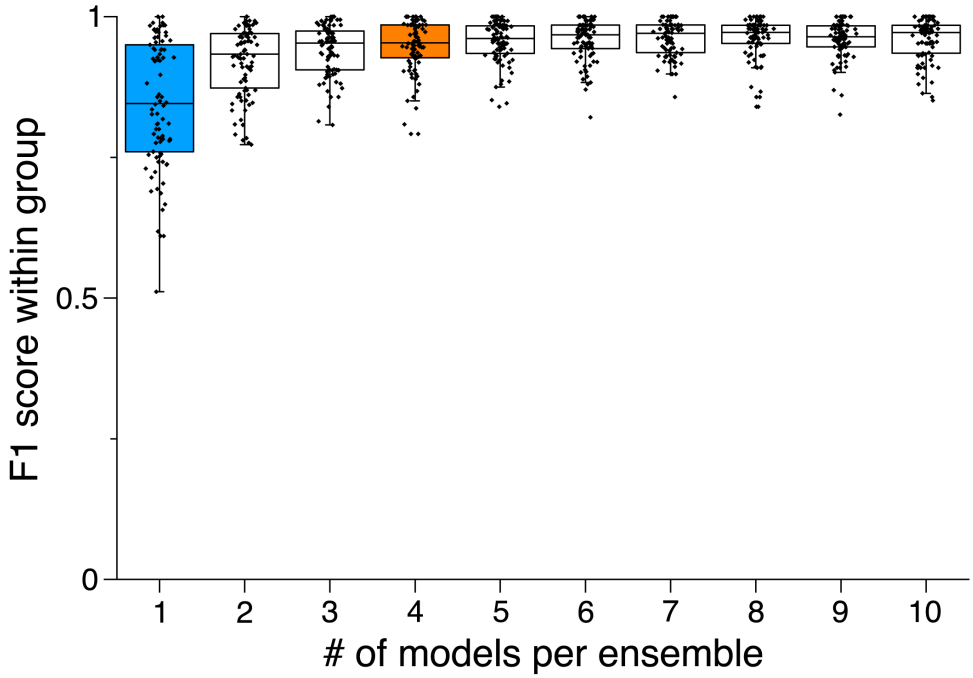


Figure S4. Ensemble size and reliability. To determine an appropriate size for the *consensus ensembles*, we analyzed the homogeneity of the results through a similarity analysis. Therefore, we calculated the $M_{\text{F1 score}}$ at an IoU matching threshold of $\iota = 0.5$ for each ensemble size $i \in \{1, ..., 10\}$ on the holdout test set (n=9 images). Stratified on the cross validation splits we randomly sampled the ensembles from a collection of trained *consensus models*. We repeated this procedure five times to mitigate the random effect of the ensemble composition ($N_{\text{ensembles}}$=5 for each i). The blue box ($i = 1$) depicts the variability between different *consensus models*. The orange box ($i = 4$) shows the variability of the finally chosen size for the *consensus ensembles*, as no substantial reduction in variation can be observed for larger i. In addition, $i = 4$ corresponds to the number of cross validation splits ($k = 4$), meaning that the ensembles have seen the entire training set.

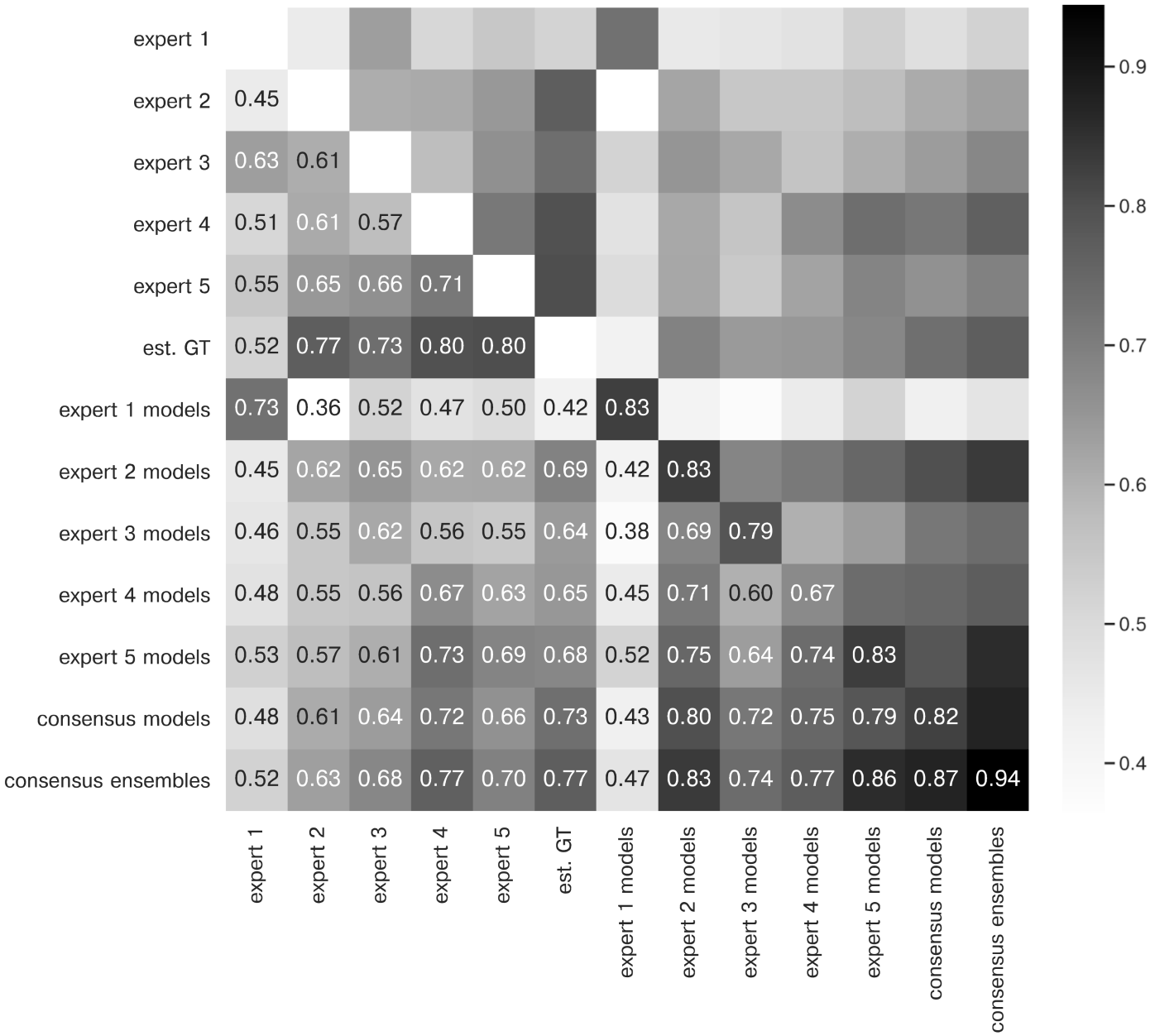


Figure S5. Extended similarity analysis: F1 score. The heatmap shows the mean of $M_{\text{F1 score}}$ at a matching IoU-threshold of $t = 0.5$ for the image feature annotations of the indicated coders. Segmentation masks of the five human experts (N_{expert}= 1 per expert), the estimated ground-truth ($N_{est.\ GT}$= 1), the respective expert models, the consensus models, and the consensus ensembles (N_{models}= 4 per model or ensemble) are compared. The diagonal values show the inter-model reliability (no data available for the human experts who only annotated the images once). The consensus ensembles show the highest reliability (0.94) and perform on par with human experts compared to the est. GT (0.77). Both expert 1 and the corresponding expert 1 models show overall low similarities to other experts and expert models, while sharing a high similarity to each other (0.73).

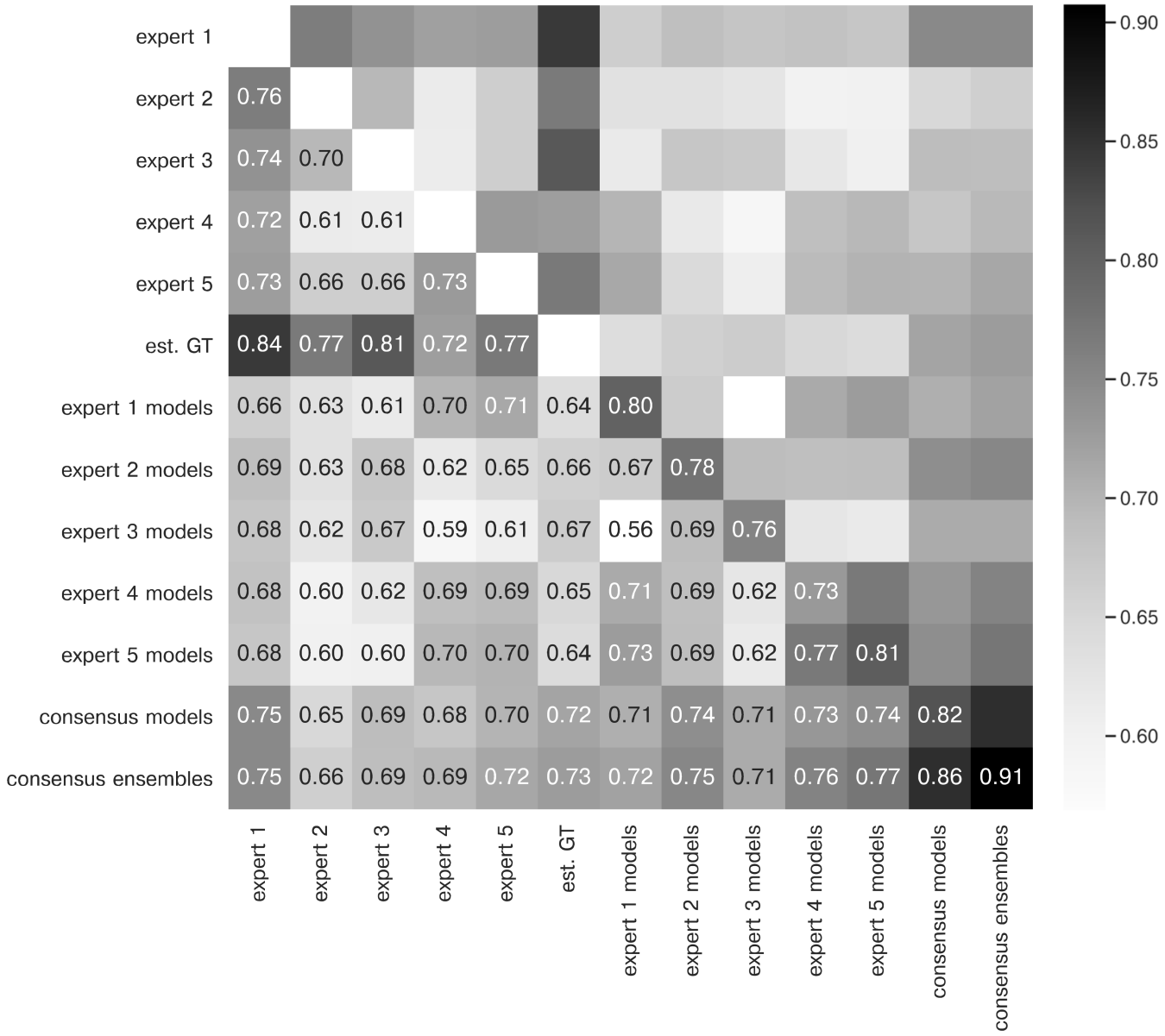


Figure S6. Extended similarity analysis: mean IoU. The heatmap shows the mean of $\bar{M}_{\text{IoU}}$ for the image feature annotations of the indicated coders. Segmentation masks of the five human experts (N_{expert}= 1 per expert), the estimated ground-truth ($N_{\text{est. GT}}$= 1), the respective expert models, the consensus models, and the consensus ensembles (N_{models}= 4 per model or ensemble) are compared. The diagonal values show the inter-model reliability (no data available for the human experts who only annotated the images once). Again, consensus ensembles show highest reliability (0.91). Est. GT annotations are directly derived from manual expert annotations, which renders this comparison favorable.

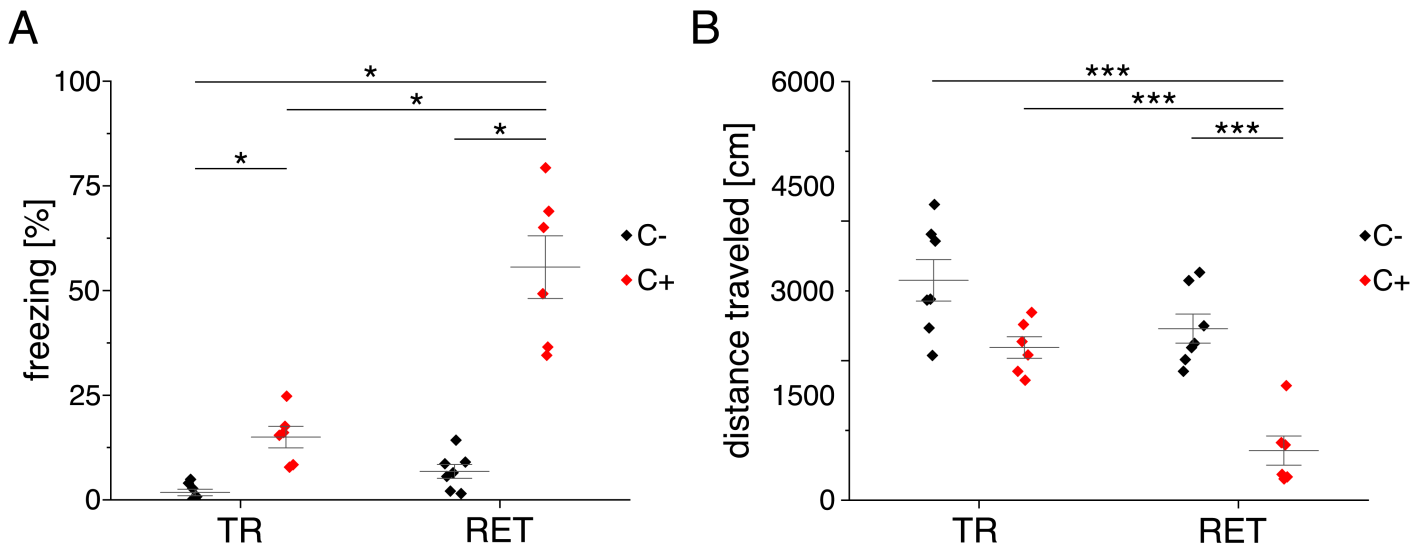


Figure S7. Behavioral analysis *Lab-Wue1*. (**A**) Fear acquisition was observed in conditioned mice (C+), while unconditioned controls (C-) did not show freezing behavior during initial context exposure (TR). In the memory retrieval session (RET), conditioned mice showed strong freezing behavior, while unconditioned mice did not freeze in response to the training context ($X^2(3) = 20.894$, $p < 0.001$, $N_{TR\ C-} = 7$, $N_{TR\ C+} = 6$, $N_{RET\ C-} = 7$, $N_{RET\ C+} = 6$, Kruskal-Wallis ANOVA followed by pairwise Mann-Whitney tests with Bonferroni correction, *: $p < 0.05$). (**B**) Distance traveled in the training context is reduced in fear conditioned mice ($F_{(3, 22)} = 19.484$, $p < 0.001$, $N_{TR\ C-} = 7$, $N_{TR\ C+} = 6$, $N_{RET\ C-} = 7$, $N_{RET\ C+} = 6$, one-way ANOVA followed by pairwise t-tests with Bonferroni correction, ***: $p < 0.001$). Source data is available as ***Figure S7 - Source data***.

laboratory	# total images	# training/ validation images	# test images	imaging technique	resolution	pixel dimensions	model organism	fluorescent feature
Lab-Mue	29	5	0	confocal LSM	0.8 px / µm	1024 x 1024	Mus musculus	cFOS (nuclear)
Lab-Inns1	24	5	0	epifluorescence microscopy	1 px / µm	1344 x 1024	Mus musculus	cFOS (nuclear)
Lab-Inns2	30	5	0	epifluorescence microscopy	2 px / µm	1376 x 1038	Mus musculus	cFOS (nuclear)
Lab-Wue1	319	36	9	confocal LSM	1.61 px / µm	1024 x 1024	Mus musculus	cFOS (nuclear)
Lab-Wue2	40	5	0	confocal LSM	12.87 px / µm	1024 x 1024	Danio rerio	GABA (somatic)

Figure S8. Characteristics of all five bioimage datasets. For each dataset, the number of total images, the number of images for the training and for the test dataset are depicted. LSM: laser-scanning microscopy.

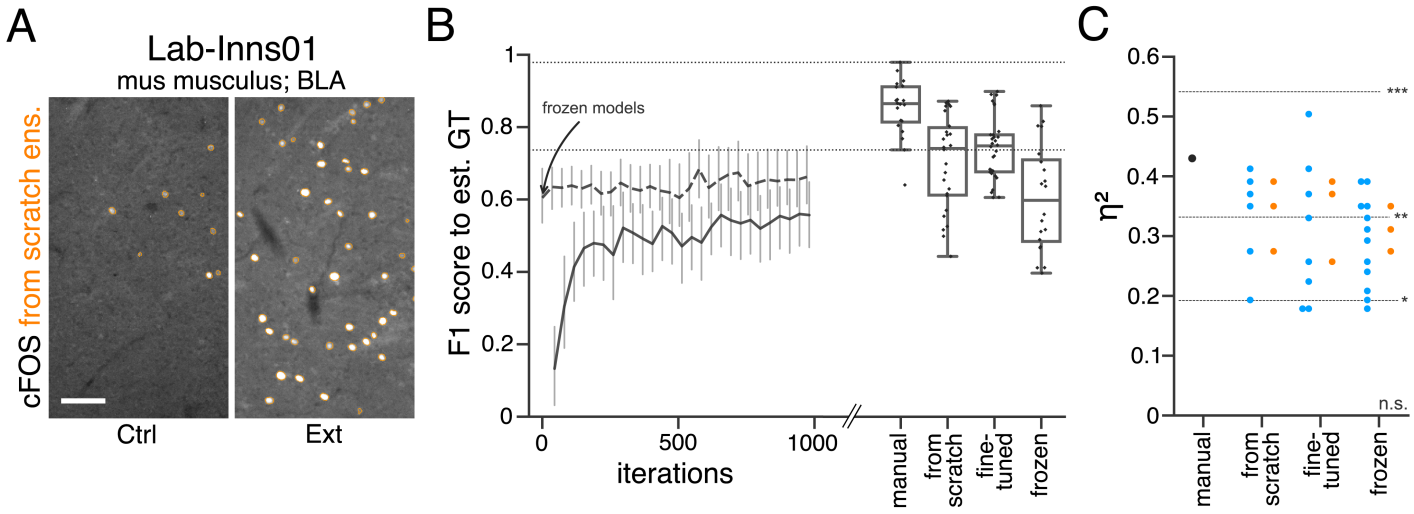


Figure S9. Performance of consensus ensembles on feature annotation in image dataset *Lab-Inns01*. (**A**) Representative microscopy images. Orange: representative annotations of a lab-specific *from scratch* consensus ensemble. BLA: basolateral amygdala, Ctrl: control, Ext: extinction. Scale bar: 80 µm. (**B**) Mean $M_{\text{F1 score}}$ of *from scratch* (solid line) and *fine-tuned* (dashed line) consensus models on the validation dataset over the course of training (iterations). Mean $M_{\text{F1 score}}$ of *frozen* consensus models are indicated with an arrow. Box plots show the $M_{\text{F1 score}}$ among the annotations of human experts as reference and the mean $M_{\text{F1 score}}$ of selected consensus models. Two dotted horizontal lines mark the whisker ends of the $M_{\text{F1 score}}$ among the human expert annotations. (**C**) Effect sizes of all individual bioimage analyses (black: manual experts, blue: consensus models, orange: consensus ensembles). Three horizontal lines separate four selected significance intervals (n.s.: not significant, *: $0.05 \geq p > 0.01$, **: $0.01 \geq p > 0.001$, ***: $p \leq 0.001$). The analyses were performed with: $N_{\text{from scratch consensus models}} = 6$, $N_{\text{fine-tuned consensus models}} = 8$, $N_{\text{frozen consensus models}} = 12$. For each initialization variant, the analyses was performed with $N_{\text{consensus ensembles}} = 3$. $N_{\text{Ctrl}} = 5$, $N_{\text{Ext}} = 5$; $n_{\text{Ctrl}} = 10$, $n_{\text{Ext}} = 9$. Statistical data can be found in ***Figure S9 - Source data***.

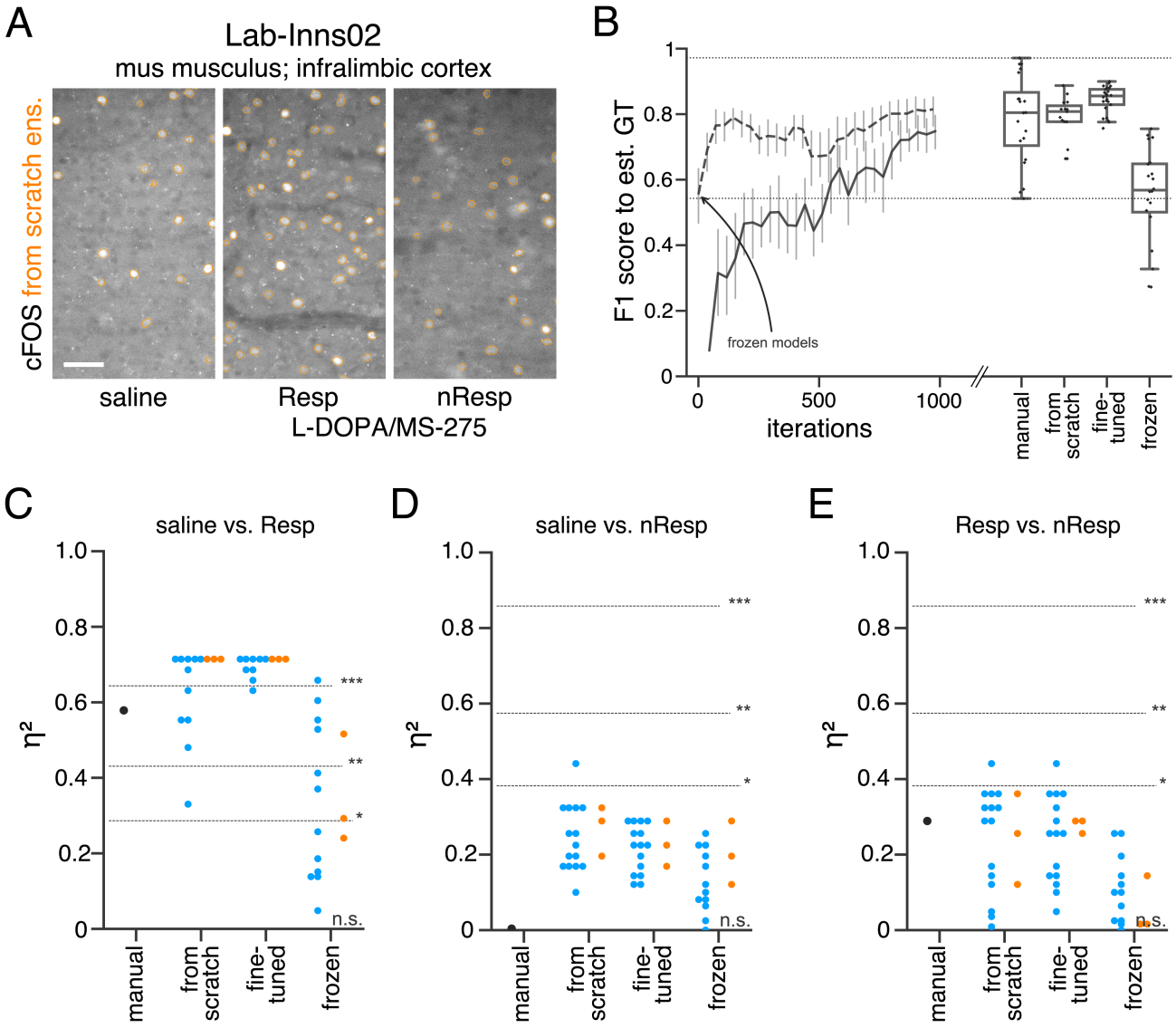


Figure S10. Performance of consensus ensembles on fluorescent feature annotation in image dataset *Lab-Inns02*. (**A**) Representative microscopy images. Orange: representative annotations of a lab-specific *from scratch* consensus ensemble. Resp: responders, nResp: non-responders. Scale bar: 40 µm. (**B**) Mean $M_{\text{F1 score}}$ of *from scratch* (solid line) and *fine-tuned* (dashed line) consensus models on the validation dataset over the course of training (iterations). Mean $M_{\text{F1 score}}$ of *frozen* consensus models are indicated with an arrow. Box plots show the $M_{\text{F1 score}}$ among the annotations of human experts as reference and the mean $M_{\text{F1 score}}$ of selected consensus models. Two dotted horizontal lines mark the whisker ends of the $M_{\text{F1 score}}$ among the human expert annotations. (**C-E**) Effect sizes of all individual bioimage analyses (black: manual experts, blue: consensus models, orange: consensus ensembles). Three horizontal lines separate four selected significance intervals (n.s.: not significant, *: $0.05 \geq p > 0.01$, **: $0.01 \geq p > 0.001$, ***: $p \leq 0.001$ after Bonferroni correction for multiple comparisons). The analyses were performed with: $N_{\text{from scratch consensus models}} = 15$, $N_{\text{fine-tuned consensus models}} = 15$, $N_{\text{frozen consensus models}} = 12$. For each initialization variant, the analyses was performed with $N_{\text{consensus ensembles}} = 3$. $N_{\text{saline}} = 6$, $N_{\text{L-DOPA/MS-275 Resp}} = 6$, $N_{\text{L-DOPA/MS-275 nResp}} = 3$; $n_{\text{saline}} = 10$, $n_{\text{L-DOPA/MS-275 Resp}} = 10$, $n_{\text{L-DOPA/MS-275 nResp}} = 5$. Statistical data can be found in ***Figure S10 - Source data***.

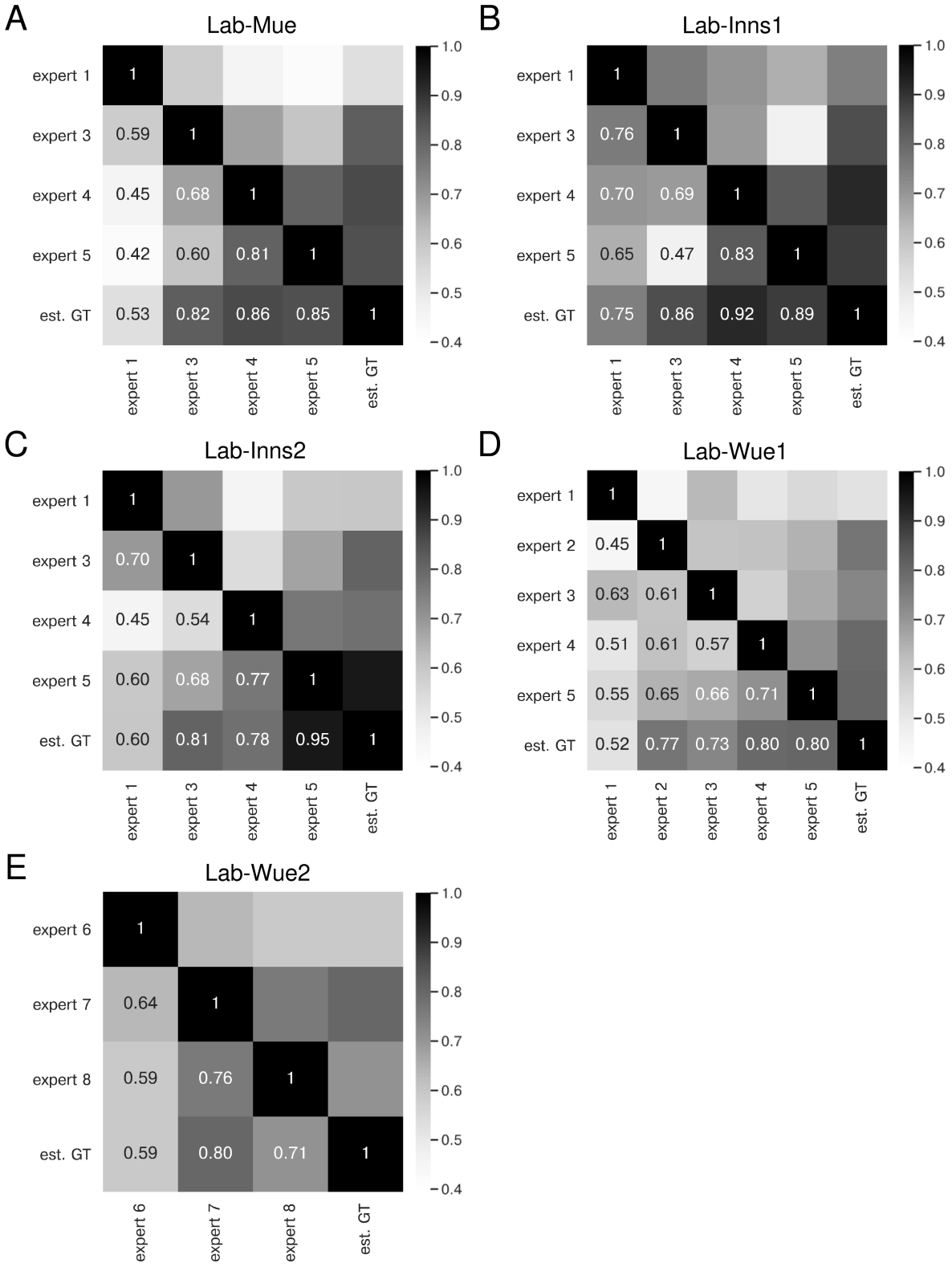


Figure S11. Expert similarity across all datasets: F1 scores. The heatmaps show the mean of $M_{\text{F1 score}}$ at a matching IoU-threshold of $t = 0.5$ for the image feature annotations of the indicated coders. The estimated ground truth (est. GT) was always calculated on all available expert annotations. The expert number refers to a unique human annotator (e.g., expert 1 is the same person across the datasets in **A-D**). The similarity scores were calculated on n=5 images for **A,B,C,** and **E** and n=9 images (test set) for **D**. The similarity between the same experts varies across the datasets (**A-D**), indicating that the heuristic bias of the annotators depends on the underlying data. However, expert 1 consistently yields the lowest similarity scores **A, C,** and **D**. The overall performance between one group of experts remains within a similar range for different datasets (**A-D**) and is comparable for a second group of experts on a different image dataset (**E**).

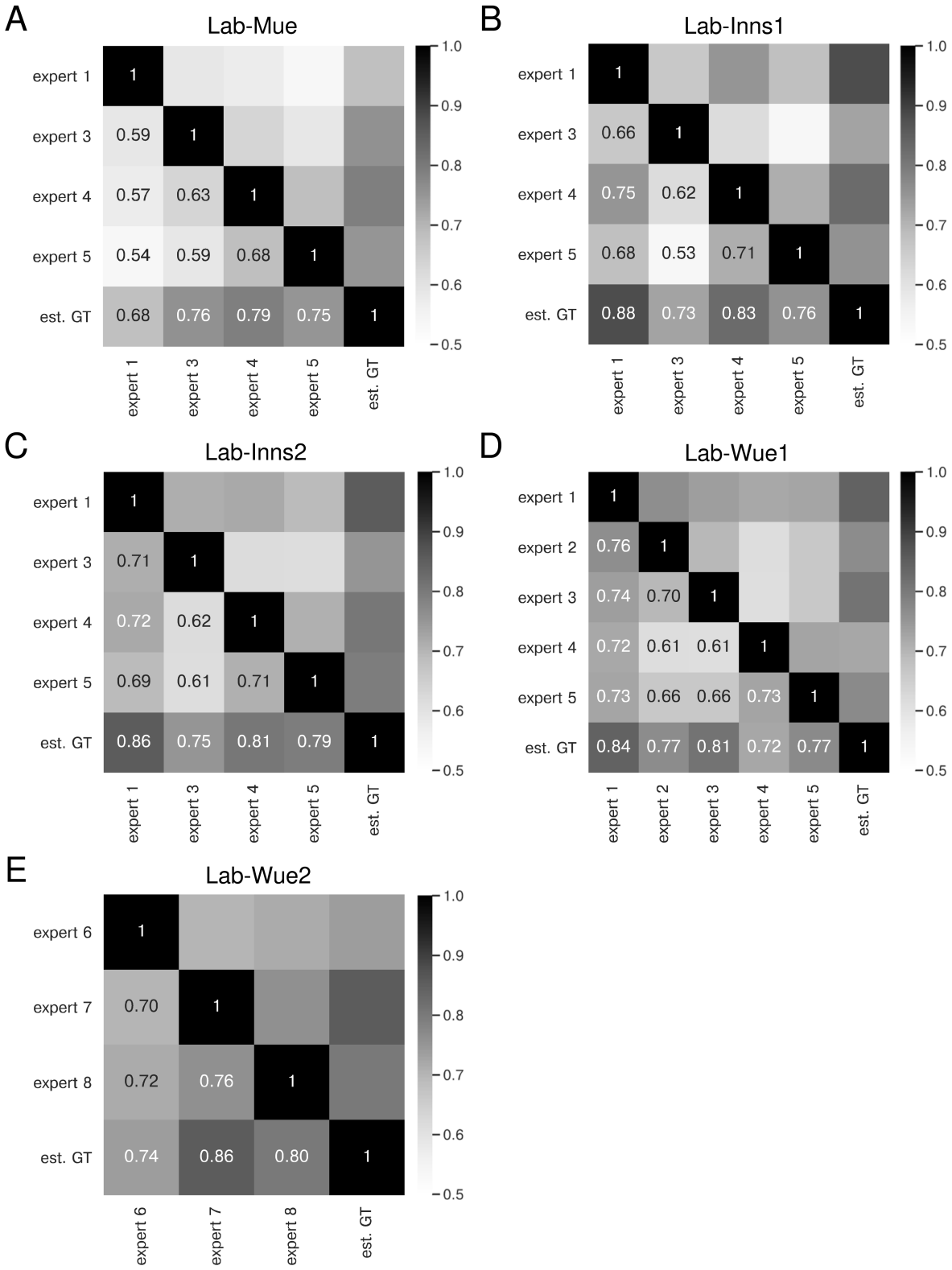


Figure S12. Expert similarity across all datasets: mean IoU. The heatmaps show the mean of $\bar{M}_{\text{IoU}}$ for the image feature annotations of the indicated coders. The estimated ground truth (est. GT) was always calculated on all available expert annotations. The expert number refers to a unique human annotator (e.g., expert 1 is the same person across the datasets in **A-D**). The similarity scores were calculated on n=5 images for **A,B,C, and E** and n=9 images (test set) for **D**. The similarity between the same experts varies across the datasets (**A-D**), indicating that the heuristic bias of the annotators depends on the underlying data. However, the overall performance between the coders remains within a similar range.

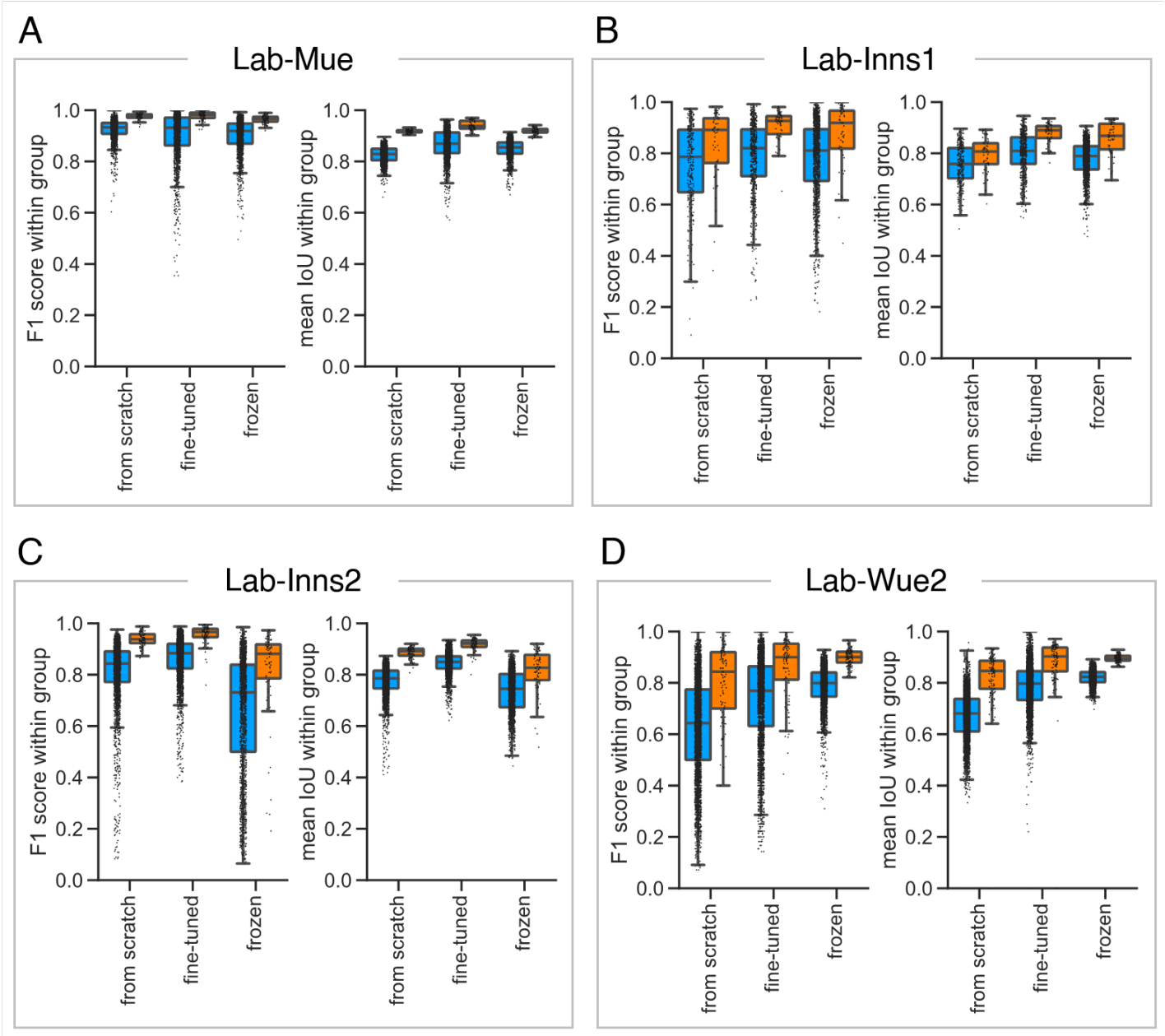


Figure S13. Reliability of the consensus approaches across different datasets. To asses the reliability and reproducibility of the *consensus models* and *consensus ensembles* strategies, we illustrate the within group $M_{\text{F1 score}}$ at a matching IoU-threshold of $t = 0.5$ and the within group $\bar{M}_{\text{IoU}}$ across different datasets. Color coding refers to the different DL strategies: blue to *consensus models*; orange to *consensus ensembles*. All images used for training were excluded from the analysis. The differences between the datasets highlight the difficulties in establishing a unified approach for automated fluorescent label annotation. (**A**) The *Lab-Mue* analysis comprises n=24 images and the following models: $N_{\text{consensus models}}$=12 and $N_{\text{consensus ensembles}}$=3 for all initialization variants. (**B**) The *Lab-Inns1* analysis comprises n=20 images and the following models: $N_{\text{from scratch consensus models}}$=6, $N_{\text{fine-tuned consensus models}}$=8, $N_{\text{frozen consensus models}}$=12, and $N_{\text{consensus ensembles}}$=3 for all initialization variants. (**C**) The *Lab-Inns2* analysis comprises n=25 images and the following models: $N_{\text{from scratch consensus models}}$=15, $N_{\text{fine-tuned consensus models}}$=15, $N_{\text{frozen consensus models}}$=12, and $N_{\text{consensus ensembles}}$=3 for all initialization variants. (**D**) The *Lab-Wue2* analysis comprises n=25 images and the following models: $N_{\text{from scratch consensus models}}$=15, $N_{\text{fine-tuned consensus models}}$=15, $N_{\text{frozen consensus models}}$=12, and $N_{\text{consensus ensembles}}$=3 for each initialization variant.