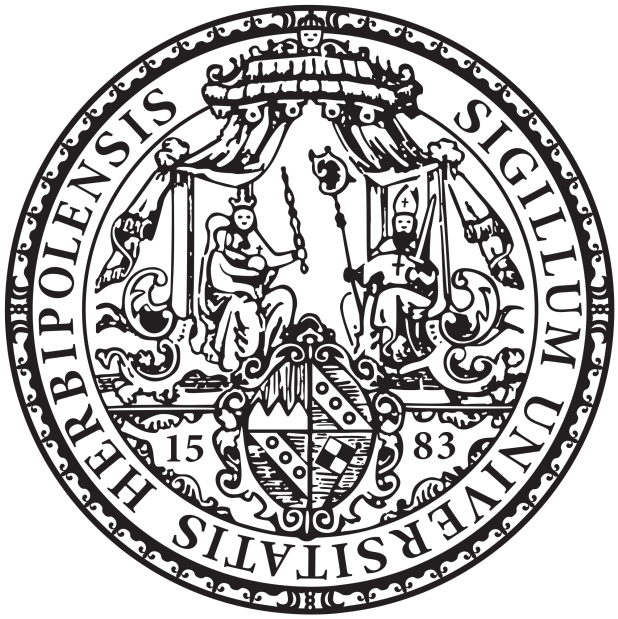
Evaluation and validation of deep learning strategies for bioimage analyses

Evaluation und Validierung von Deep learning Strategien für die Analyse biologischer Bilddaten



Doctoral thesis for a doctoral degree

at the Graduate School of Life Sciences,

Julius-Maximilians-Universität Würzburg,

Section **Neuroscience**

submitted by

Dennis Segebarth

from

Bad Kissingen

Würzburg 2020

Submitted on: . . . . . .

Office stamp

Members of the *Promotionskomitee*:

Chairperson:	Prof. Dr. Charlotte Förster
Primary Supervisor:	PD Dr. Robert Blum
Supervisor (Second):	PD Dr. Chi Wang Ip
Supervisor (Third):	Dr. Marta Andreatta
Supervisor (Fourth):	Prof. Dr. Philip Tovote

Date of Public Defence: . . . . . .

Date of Receipt of Certificates: . . . . . .

Affidavit

I hereby confirm that my thesis entitled "**Evaluation and validation of deep learning strategies for bioimage analyses**" is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Place, Date Signature

Eidesstattliche Erklärung

Hiermit erkläre Ich an Eides statt, die Dissertation "**Evaluation und Validierung von Deep learning Strategien für die Analyse biologischer Bilddaten**" eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Disserattion weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Ort, Datum Unterschrift

Contents

Abstract

Significant advances in fluorescence imaging techniques enable life scientists today to gain insights into biological systems at an unprecedented scale. The interpretation of image features in such bioimage datasets and their subsequent quantitative analysis is referred to as bioimage analysis. A substantial proportion of bioimage analyses is still performed manually by a human expert - a tedious process that is long known to be subjective. Particularly in tasks that require the annotation of image features with a low signal-to-noise ratio, like in fluorescence images of tissue samples, the inter-rater agreement drops. However, like any other scientific analysis, also bioimage analysis has to meet the general quality criteria of quantitative research, which are objectivity, reliability, and validity. Thus, the automation of bioimage analysis with computer-aided approaches is highly desirable. Albeit conventional hard-coded algorithms are fully unbiased, a human user has to set its respective feature extraction parameters. Thus, also these approaches can be considered subjective.

Recently, deep learning (DL) has enabled impressive advances in computer vision research. The predominant difference between DL and conventional algorithms is the capability of DL models to learn the respective task on base of an annotated training dataset, instead of following user-defined rules for feature extraction. This thesis hypothesized that DL can be used to increase the objectivity, reliability, and validity of bioimage analyses, thus going beyond mere automation. However, in absence of ground truth annotations, DL models have to be trained on manual and thus subjective annotations, which could cause the model to incorporate such a bias. Moreover, model training is stochastic and even training on the same data could result in models with divergent outputs. Consequently, both the training on subjective annotations and the model-to-model variability could impair the quality of DL-based bioimage analyses. This thesis systematically assessed the impacts of these two limitations experimentally by analyzing fluorescence signals of a protein called cFOS in mouse brain sections. Since the abundance of cFOS correlates with mouse behavior, behavioral analyses could be used for cross-validation of the bioimage analysis results. Furthermore, this thesis showed that pooling the input of multiple human experts during model training and integration of multiple trained models in a model ensemble can mitigate the impact of these limitations. In summary, the present study establishes guidelines for how DL can be used to increase the general quality of bioimage analyses.

Zusammenfassung

Fortschritte in den Methoden der fluoreszenz-basierten Bildgebung ermöglichen Biowissenschaftlern heutzutage noch nie dagewesene Einblicke in biologische Systeme. Die Interpretation sowie die anschließende quantitative Analyse von Bildelementen in biologischen Bilddatensätzen wird in der Wissenschaft als *bioimage analysis* bezeichnet. Ein wesentlicher Anteil der *bioimage analysis* wird noch immer von Experten per Hand durchgeführt - ein mühsamer Prozess, von dem man seit langem weiß, dass er subjektiv ist. Besonders bei Aufgabestellungen, welche die Annotierung von Bildelementen mit einem geringen Signal-Rausch-Verhältnis erfordern, wie es beispielsweise bei Fluoreszenzbildern von Gewebeproben der Fall ist, sinkt die Übereinstimmung zwischen den Bewertungen mehrerer Experten. Genauso wie jede andere wissenschaftliche Analyse, muss jedoch auch die *bioimage analysis* den generellen Qualitätskriterien quantitativer Forschung gerecht werden. Dies sind Objektivität, Zuverlässigkeit und Validität. Die Automatisierung der *bioimage analysis* mit Hilfe von computer-basierten Ansätzen ist somit erstrebenswert. Konventionelle, hartkodierte Algorithmen sind zwar vollkommen unvoreingenommen, jedoch legt ein menschlicher Benutzer jene Parameter fest, die der Algorithmus für die Extraktion der relevanten Bildelemente nutzt. Aus diesem Grund sind auch diese Ansätze zumindest partiell subjektiv.

In den letzten Jahren hat Deep learning (DL) zu beeindruckenden Fortschritten auf dem Forschungsgebiet der computer vision beigetragen. Der vorherrschende Unterschied zwischen DL und konventionellen Algorithmen besteht darin, dass DL Modelle in der Lage sind die jeweilige Aufgabe auf Grundlage eines annotierten Trainingsdatensatzes zu lernen, anstatt starr den Parametern zu folgen, die der Benutzer für die Extraktion der relevanten Bildelemente vorgegeben hat.

In dieser Dissertation wurde die Hypothese untersucht, ob DL, neben der Möglichkeit der automatischen Bildanalyse, auch dazu genutzt werden kann die Objektivität, die Zuverlässigkeit und die Validität der Bildanalyse zu verbessern. Ohne eine objektive Referenzannotierung muss das Training der DL Modelle jedoch auf händisch erstellten und somit also subjektiven Annotierungen durchgeführt werden. Theoretisch könnte dies dazu führen, dass das DL-Modell diese Vorgeingenommenheit übernimmt. Außerdem unterliegt das Training der Modelle stochastischen Prozessen und selbst Modelle, die auf den gleichen Trainingsdaten trainiert wurden, könnten sich danach in ihren aus-

gegeben Analysen unterscheiden. Demzufolge könnten also sowohl das Training auf subjektiven Annotierungen als auch die Variabilität von Modell zu Modell die Qualität der DL-basierten Analyse von biologischen Bilddaten beeinträchtigen. In dieser Dissertation werden die Einflüsse von diesen beiden Limitierungen auf Grundlage von experimentellen Daten untersucht. In den experimentellen Bilddaten werden Fluoreszenzsignale des Proteins cFOS in Hirnschnitten von Mäusen dargestellt und hier repräsentativ untersucht. Da das Vorkommen von cFOS mit dem Verhalten der Mäuse korreliert, kann die Analyse des Verhaltens der Mäuse zur Kreuzvalidierung der Analyse der biologischen Bilddaten herangezogen werden. Die Daten dieser Dissertation zeigen, dass die Integration mehrerer Experten in das Training eines Modells sowie die Integration mehrerer trainierter Modelle in ein Modell-Ensemble das Risiko einer subjektiven oder nicht reproduzierbaren Bildanalyse abschwächen können. Diese Arbeit etabliert Richtlinien dafür, wie DL verwendet werden kann, um die generelle Qualität der Analyse biologischer Bilddaten zu erhöhen.

1 Introduction

Continuous advances in image acquisition techniques enable modern research throughout the life sciences to increasingly gain information about biological systems from image data (A. Li et al., 2010; Osten and Margrie, 2013; Boutros et al., 2015; Meijering et al., 2016; Caicedo, Cooper, et al., 2017; McDole et al., 2018; Caicedo, Goodman, et al., 2019). With the concomitant optimization of fluorescent probes, it is now possible, for instance, to perform simultaneous *in vivo* calcium imaging of two brain regions in unrestrained mice (Gonzalez et al., 2019; Groot et al., 2020), to run high-throughput image-based morphological profiling (Caicedo, Cooper, et al., 2017), or to record the development of an entire mouse embryo at single cell level (McDole et al., 2018). The corresponding, ever increasing amount of acquired image data calls for the automatized and unbiased image data analysis (Danuser, 2011; McQuin et al., 2018; Moen et al., 2019; Caicedo, Goodman, et al., 2019). And yet, the development of computer-aided image analysis strategies for the bioimaging community failed for a long time to keep up with the pace of innovations in the designs of both microscopy and experimental setups (Danuser, 2011; Meijering, 2012; Meijering et al., 2016).

1.1 Bioimage analysis of fluorescence microscopy data

Today, the majority of quantitative bioimage datasets in the life sciences is based on fluorescence microscopy (Caicedo, Roth, et al., 2019), which allows the targeted imaging of fluorescently labeled macromolecules. The analysis process of such datasets is known as bioimage analysis and requires the annotation of biologically relevant image features and their subsequent quantitative analysis in order to test an underlying experimental hypothesis (Meijering et al., 2016). While the quantitative analysis of annotated features is usually straightforward and can easily be automatized (Meijering et al., 2016), the annotation of image features is challenging (Meijering, 2012; Meijering et al., 2016; Van Valen et al., 2016). Human experts integrate several criteria like morphology or fluorescence signal intensity for the annotation process and a substantial level of programming knowledge is required to transfer these criteria into a conventional, hard-coded computer-aided approach (LeCun et al., 2015; Chamier et al., 2019). While there is a plethora of computer-aided approaches for image feature segmentation available, they are often limited to a very specific task (Meijering, 2012; Van Valen et al., 2016), and their

adaptation to new datasets again necessitates computational expertise (Chamier et al., 2019). However, with the advent of deep learning (DL) algorithms that are capable of learning a specific task solely by being presented with pairs of raw input and the desired output, this is about to change (LeCun et al., 2015; Van Valen et al., 2016; Chamier et al., 2019).

1.2 The basic principles of deep learning

DL is a subclass of machine learning approaches based on representation learning methods (LeCun et al., 2015). In a DL algorithm, a multitude of non-linear modules are arranged in several layers as a neural network (LeCun et al., 2015; Chamier et al., 2019). Each module transforms its respective input into a more abstract representation of the data, which serves as input for the following module. Thus, increasingly higher levels of representations are created. Particularly high-level representations could then be responsive even to minor changes in features of the original input that are key for the correct discrimination, and yet be unaffected by major fluctuations in irrelevant features (LeCun et al., 2015). This is in stark contrast to conventional algorithms or other machine learning approaches, where such representations had to be hand-engineered as feature extractors by thoroughly designing appropriate data transformations (LeCun et al., 2015; Moen et al., 2019). Importantly, concepts like backpropagation allow the identification of such representations within the neural network. During supervised training of a DL algorithm, a loss-function is used to evaluate the deviation of the algorithms prediction from the correct result (LeCun et al., 2015; Moen et al., 2019). With the goal of minimizing this deviation, a gradient of the loss-function can be calculated which allows to change the weights of each module within one layer, propagating backwards from the output layer-by-layer towards the input (LeCun et al., 2015; Moen et al., 2019). Ultimately, the weights of modules that entail representations that contribute to the correct prediction will be increased, while the weights of less relevant modules will be decreased. Thus, DL has the potential to learn even a complex coherence that might be hidden in a high-dimensional representation of the data, solely on base of a training dataset (LeCun et al., 2015; Moen et al., 2019). Once the training of the algorithm is finalized, the trained model can be used to perform the respective task on new data (LeCun et al., 2015; Moen et al., 2019; Chamier et al., 2019).

1.3 Deep learning in the life sciences

Initiated by the triumph of a DL-based submission in the 2012 ImageNet Large Scale Visual Recognition Challenge (Krizhevsky et al., 2012), DL revolutionized the field of artificial intelligence research, including computer vision (LeCun et al., 2015). Since DL opens new possibilities to perform automatized image analysis, it also attracted the attention of the bioimaging community and is discussed to become the new state-of-the-art method for the analysis of bioimaging data, or even to enable analyses that were so far impossible (LeCun et al., 2015; Moen et al., 2019; Chamier et al., 2019). Supporting this assumption, several recent studies confirmed the remarkable capacities of DL in the field of bioimaging (a comprehensive collection can be found at `nature.com/collections/cfcdjceech`). For example, Buggenthin et al. (2017) demonstrate that DL can identify the lineage choice of primary hematopoietic progenitors from bright-field images, up to three generations prior to the expression of conventional molecular markers (Buggenthin et al., 2017). Similarly, DL can also be used to predict fluorescent labels directly from bright-field images, both in 2D and 3D (Christiansen et al., 2018; Ounkomol et al., 2018). Furthermore, DL was shown to be capable of enhancing the resolution of fluorescence images, for instance to generate super-resolution images from diffraction-limited confocal images (Wang et al., 2019). Notably, DL-based approaches were already shown to outperform conventional, hard-coded algorithms for several applications (Caicedo, Roth, et al., 2019).

However, application of DL approaches throughout the life sciences remained restricted, essentially due to the high demands for both computing power and computational expertise (Haberl et al., 2018). Embedding of DL in commonly used frameworks like Fiji (Falk et al., 2019) and the *CellProfiler* (McQuin et al., 2018), or in cloud-computing environments (Haberl et al., 2018; Nath et al., 2019) are crucial steps to lift these limitations. Nevertheless, before DL can unfold its full potential in the life sciences, biomedical researchers have to familiarize themselves with its basic principles (Moen et al., 2019). In addition, critical evaluation of DL-based workflows is essential to establish trust in the hidden computations of DL models (Chamier et al., 2019). So far, DL approaches for fluorescent feature annotations were primarily evaluated by similarity measures like precision, recall, or F1-scores (Falk et al., 2019; Haberl et al., 2018; Caicedo, Goodman, et al., 2019), while the main goal of bioimage analysis is to test a certain hypothesis (Meijering et al., 2016). The effects of DL on the objectivity and reproducibility of bioimage

analysis, however, have not yet been evaluated systematically, irrespective of their importance.

1.4 Limitations of deep learning strategies for bioimage analysis

Manual image analysis by a human expert is a pre-requisite for the training of DL algorithms. However, it is a tedious process and can introduce a subjective bias (Chamier et al., 2019; McQuin et al., 2018; Caicedo, Roth, et al., 2019; Collier et al., 2003), especially in case of image features with borders of low contrast (Niedworok et al., 2016). Computer-aided automation of the annotation process can speed up the analysis, but the implementation and the visual performance inspection of conventional algorithms remains user-based and can therefore still be considered subjective (Tadrous, 2010; Chamier et al., 2019). Likewise, the subjectivity of manual annotations also causes a critical problem for DL-based approaches, since the network is trained on pairs of fluorescence images and the corresponding manual annotations. Consequently, a subjective bias present in the annotations of the training dataset could become incorporated into the DL model (Falk et al., 2019; Chamier et al., 2019; Moen et al., 2019). Moreover, the training of DL algorithms is a stochastic process and even the output of models trained on the same training dataset could vary significantly (Dietterich, 2000). A possible reason for this could be that the models get trapped in different local optima during the training, for instance due to the random initialization of the weights prior to the training or due to the random sampling of images during the training process (Dietterich, 2000; Ronneberger et al., 2015). Intuitively, such discrepancies could also affect the reproducibility of annotations and hence of the subsequent statistical analysis. Thus, using a DL model trained on the annotations of a single human expert might therefore yield subjective and irreproducible bioimage analysis results, particularly on image datasets with low signal-to-noise ratios.

1.5 Aim of this thesis

The central hypothesis of this thesis was, that DL could also hold the potential to increase both objectivity and reproducibility of bioimage analyses. Here, *objectivity* describes the neutrality of the evidence, which is negatively affected by personal preferences, emotions, or any other limitation that could introduce a bias during data acquisi-

tion or analysis (Frambach et al., 2013). In case of fluorescent feature annotations, such limitations could also be the context in which manual annotation was performed, in addition to subjective biases and individual graphical perception capabilities (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a; Cleveland and McGill, 1985). *Reproducibility* refers to the consistency of the evidence and how likely it is to obtain the same results if a given study is repeated under slightly changed conditions (Frambach et al., 2013; Taylor and Kuyatt, 1994). In this study, reproducibility is assessed by the repetitive analysis of the same bioimage dataset with different DL models.

In order to adhere to both quality criteria, the aforementioned limitations of DL-based bioimage analyses have to be addressed. For instance, DL algorithms are not limited to be trained on the annotations of only a single human expert and the pooled input of multiple human experts could be used instead. In such a way, mutual and hence less subjective annotations might have a higher impact during training of the algorithm. Following the same reasoning as above, this could then cause the algorithm to learn more objective annotation criteria.

Similarly, the output of several trained models can be merged into a model ensemble, which is then used for the bioimage analysis. Ensemble-formation is a well-established method to gain prediction quality (Dietterich, 2000) and could reduce the impact of randomness during the training. Consequently, this would enhance the reproducibility of ensemble-based bioimage analysis results.

In summary, this thesis systematically evaluates the impact of subjective manual annotations and of model-to-model variability on the performance of DL algorithms and on the corresponding, DL-enabled bioimage analyses results. Moreover, by testing adaptations to the DL-based workflow, it derives guidelines for how DL can be used in order to obtain objective and reproducible bioimage analysis results.

2 Material and Methods

2.1 Mice

All experiments were performed in accordance with the guidelines set by the European Union and by our local veterinary authority (Veterinäramt der Stadt Würzburg) and were approved by our institutional Animal Care, the Utilization Committee, and the Regierung von Unterfranken, Würzburg, Germany (License numbers: 55.2-2531.01-95/13, 55.2.2-2532-2-558, 55.2.2-2532.2-918-15).

All mice were bred in the animal facility of the Institute of Clinical Neurobiology at the University Hospital of Würzburg. Animals were housed in groups of three to five individuals and under standard laboratory conditions, i.e. 12 hours light/dark cycle (LD 12:12), at constant temperature (21 $\pm$ 1°C), and with access to food and water *ad libitum*. Pathogen-screening was performed once per year according to the Harlan 52M profile (Harlan laboratories, Netherlands). In addition, mice were tested quarterly according to the Harlan 51M profile. All mice used for this thesis were healthy, free of pathogens, and showed no apparent behavioral phenotypes. The allocation of mice to the specific experimental groups was randomized wherever possible. The experimenter was blinded to the genotype of the mice for all behavioral experiments.

All behavioral experiments were performed during the subjective day phase of the animals and exclusively with male mice at an age of eight to twelve weeks. Prior to behavioral testing, mice were transferred into new cages and housed individually over the course of the experiments, yet with visual, olfactory, and auditory contact to each other inside a ventilated cabinet (Scantainer, Scanbur). All mice were handled twice a day for at least two consecutive days before the start of behavioral testing to habituate them to the male experimenter and to the experimental rooms. For each experimental session, mice were transported in their homecages to the experimental room.

2.2 Contextual fear conditioning

Contextual fear conditioning (also called *threat* conditioning; LeDoux, 2014) was performed with the multi conditioning setup (series 256060) by TSE (Bad Homburg, Germany). The motion of the mice during all sessions was tracked using the TSE MSC

FCS-SQ MED software. A squared conditioning chamber with a metal grid floor was used as training context and was cleaned with 70% ethanol prior to each use.

2.2.1 Acquisition session

Mice were allowed to freely explore the training context for an initial habituation phase of 60 s. Afterwards, five electric foot shocks (unconditioned stimulus, US: 1 s, 0.7 mA) were presented with a fixed inter-stimulus interval of one minute. 30 seconds after the last US presentation, i.e. after a total exploration time of 335 s, mice were transferred back into their homecage and their housing cabinet.

2.2.2 Retrieval session

Mice were re-exposed to the training context 24 hours after the acquisition session and allowed to freely explore the context for 360 s without presentations of the US.

2.2.3 Extinction sessions

To assess extinction learning, mice were re-exposed to the training context (360 s exploration, no US presentations) twice a day with an inter-session interval of three hours for three consecutive days, i.e. a total of six extinction session (Ext1 - Ext6). The first extinction session was performed 24 hours after the retrieval session and, thus, 48 hours after the acquisition session.

2.2.4 Analysis of freezing behavior

The TSE MSC FCS-SQ MED software was used to compute the freezing behavior of each mouse during each session, based on the motion tracking data. For this, freezing was defined as a period of at least two seconds of complete immobilization of the animal, as determined by the motion tracking, barring respiratory movements (Fanselow, 1980). Two freezing periods were combined, if the time-interval between the two periods was shorter than 100 ms.

2.2.5 Experimental groups for bioimage analysis of cFOS

Three groups of mice were used for the bioimage analysis of cFOS (Figure 1). A first group of mice underwent contextual fear conditioning acquisition and retrieval as described above (C+). In a second group (context control group, C-), mice were exposed for 335 s to the training context, just as C+ mice during the acquisition session, but without the presentation of electrical foot shocks. The retrieval session was identical for both groups (360 s, no US presentations). Mice of a third group (homecage control group, H), remained in their homecages within the housing cabinet during both acquisition and retrieval sessions.

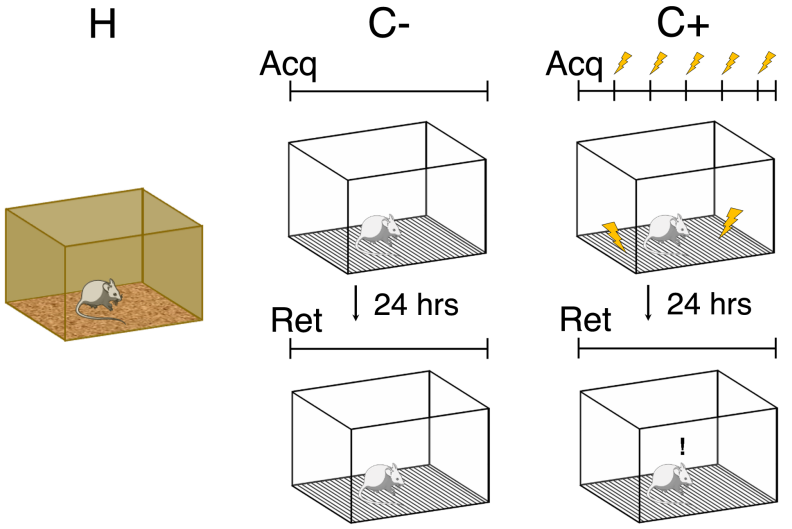


Figure 1: Experimental conditions for bioimage analysis of cFOS signals.
Three experimental groups were investigated: Mice kept in their homecage (H), mice that were trained to a context, but did not experience an electric foot shock (C-) and mice exposed to five foot shocks in the training context (C+). 24 hours after the initial training (Acq), mice were re-exposed to the training context for memory retrieval (Ret). Memory retrieval induces changes in cFOS levels. Adapted from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

2.3 Brain sample preparation

2.3.1 Anaesthesia

All mice were deeply anaesthetized prior to perfusion. For C- and C+ mice, anaesthesia was induced 90 minutes after the retrieval session and at a comparable point in time during the day for homecage controls. At first, mice were quickly anaesthetized using a rodent anaesthesia setup (Harvard Apparatus) and the volatile narcotic isoflurane

(airflow 0.4 L/min, 4% isoflurane, Iso-Vet, Chanelle). Then, deep anaesthesia was induced with a mixture of ketamine (100 mg/kg; Ursotamin, Serumwerk) and xylazine (16 mg/kg; cp-Pharma, Xylavet, Burgdorf, Germany), which was injected intraperitoneally (12 µl/g bodyweight).

2.3.2 Perfusion

After the tail reflex and the hind limb pedal reflexes were absent and the deep anaesthesia of the mice was ensured, mice were dissected and transcardially perfused using a gravity perfusion setup after puncturing the left ventricle. At first, the blood was washed out by perfusion for five minutes with 0.4% heparin (Table 1). Afterwards, the fixation of the tissue was achieved by perfusion with 4% PFA (Table 1) for another five minutes. Ultimately, brains were dissected and post-fixed in 4% PFA for two hours at 4°C. The brains were then washed using 1x PBS (Table 1) prior to embedding in 6% Agarose (in 1x PBS) for sectioning at a vibratome.

2.3.3 Serial sectioning

The embedded brains were cut in 40 µm thick coronal sections with a vibratome (Leica VT1200). Starting from Bregma -1.22 mm (according to Paxinos and Franklin, 2004), the following 30 posterior sections were considered as dorsal hippocampus. For the bioimage analysis of cFOS in the dorsal hippocampus, the 4^{th}, 14^{th}, and 24^{th} section of each brain were evaluated.

2.4 Immunohistochemistry

Immunohistochemistry was performed with up to three free floating sections per well in 24-well plates in a volume of 400 µl under constant shaking. For quenching, brain sections were incubated in quenching solution (Table 1) for one hour at RT. Sections were then permeabilized and blocked in blocking solution (Table 1) for one hour at RT, before they were incubated with the following primary antibodies at indicated dilutions in blocking solution for 48 hours at 4°C (mouse anti-Parvalbumin, SWANT, PV235, 1:5,000; rabbit anti-cFOS, SynapticSystems, 226003, 1:10,000 (lot# 226003/7); guinea-pig anti-NeuN, SynapticSystems, 266004, 1:400). The primary antibodies were washed off thrice

with washing solution (Table 1) for ten minutes at RT. The sections were then incubated with the following, fluorescently labeled secondary antibodies at a concentration of 0.5 mg/ml in blocking solution for 1.5 hours at RT (goat anti-mouse Alexa-488 conjugated, Life sciences, Thermo; donkey anti-rabbit Cy3 conjugated, Jackson ImmunoResearch; donkey anti-guinea-pig Cy5 conjugated, Jackson ImmunoResearch). Following another three washes with washing solution and one wash with 1x PBS for ten minutes at RT, the free floating sections were stained with DAPI (2 mg/ml) for five minutes. After two final washes for ten minutes with 1x PBS at RT, the sections were mounted on an object slide using Aqua-Poly/Mount and stored at 4°C.

2.5 Image acquisition

An inverted Olympus IX81 microscope equipped with an Olympus FV1000 confocal laser scanning system, an Olympus UPlan SAPO 20x/0.75 objective, a FVD10 SPD spectral detector, and diode lasers of 473, 559 and 635 nm was used for image acquisition. For the bioimage analysis of cFOS, 12-bit z-stack images covering 636 x 636 µm (1024 x 1024 pixel) and the entire thickness of the brain section with a step-size of 1.5 µm were acquired of the *dentate gyrus* (DG), the *Cornu ammonis 1* (CA1), and the CA3 region of the dorsal hippocampus. Confocal z-stack images of all examined hippocampal subregions were acquired - wherever possible - in each hemisphere of the three examined sections of each brain, resulting in a maximum of six images (n) of each hippocampal subregion per animal (N). The experimenter was blinded to the experimental condition and the genotype of the mice during image acquisition. Image acquisition parameters were kept constant within each experiment.

2.6 Image processing

ImageJ (Schneider et al., 2012) was used for all image processing steps except final figure preparation, which was performed using Adobe Photoshop (version CS5).

At first, a grey-scale maximum intensity projection was computed for each channel of the confocal image z-stack and was converted from 12-bit to 8-bit without adaptations to brightness or contrast. A total of 45 images derived from WT mice was selected for training (36 images) and testing (nine images) of the DL algorithms. The images of both subsets were selected to represent equal amounts of images of each investigated hip-

pocampal subregion (DG, CA3, and CA1) and of each examined experimental condition (H, C-, and C+). Thus, the training dataset contained four images of each experimental condition per hippocampal subregion (4 x 3 x 3), whereas the testing dataset contained one representative image of each experimental condition per hippocampal subregion (1 x 3 x 3). The 36 images that were used for the training of the algorithms were excluded from all bioimage analyses. The nine images of the test dataset were used to evaluate the annotations of the human experts and of the trained models and model ensembles, and were also included in the bioimage analyses.

2.7 Manual feature annotation

A group of five PhD-level neuroscientists with similar experience in immunofluorescence imaging was instructed to manually segment all relevant image features in the selected 45 images according to their own criteria. In general, the relevant image features were defined as *cFOS-positive nuclei* and as *Parvalbumin-positive somata* in the corresponding maximum intensity projections of the channel that showed the respective immunofluorescence signals.

In addition, NeuN immunofluorescence signals were used to identify the area of the granule cell layer of the dentate gyrus and of the pyramidal cell layers of CA3 and CA1. The corresponding NeuN-positive regions were manually segmented as regions of interest (ROIs) by one expert in all images that were used for the bioimage analyses of cFOS-positive nuclei. All experts were blinded to the treatment conditions and to the annotations of the other experts.

2.8 Deep learning approach

The design, the training, and the use of all DL-based algorithms, as well as the ground truth estimations and the computation of similarity measures that are shown and discussed in this study were performed by and in close collaboration with Matthias Griebel under the supervision of Prof. Christoph M. Flath at the Department of Business and Economics at the University of Würzburg, Germany, as described in Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a). The respective source code is available in a Dryad repository (`www.doi.org/10.5061/dryad.4b8gtht9d`; Segebarth, Griebel, Stein, R. von Collenberg, et al., 2020b).

2.9 Bioimage analyses of fluorescent features

All bioimage analyses are based on the predicted annotations (binary segmentation masks) of the indicated models or model ensembles and were performed using custom written code.

For the bioimage analysis of cFOS-positive nuclei, the analysis was restricted to those features which were annotated within the NeuN-positive region of each image. To ensure the comparability of of quantified cFOS-positive nuclei across all images of the same hippocampal subregion, the number of analyzed cFOS-positive nuclei was normalized to the area of the corresponding NeuN-positive region for each image. These data were pooled within each experiment for each experimental condition (H, C-, and C+) and analyzed hippocampal subregion (the infrapyramidal blade of the DG, the suprapyramidal blade of the DG, both blades of the DG as 'DG whole', and the pyramidal cell layer of CA3 and of CA1). All data of one experiment were normalized to the corresponding mean value of the wildtypic homecage controls to enable the comparison across experiments.

For the bioimage analysis of Parvalbumin-positive (Parv-positive) interneurons, the number of Parv-positive somata and their mean signal intensity were quantified per image. Again, these data were pooled within each experiment for each experimental condition (H, C-, and C+) and for the analyzed hippocampal subregions (DG, CA3, and CA1) and the mean signal intensities were normalized to the corresponding mean value of the wildtypic homecage controls. In addition, Parv-positive somata were classified as cFOS-positive or cFOS-negative, depending on whether the predicted annotation of a Parv-positive soma contained an entire predicted annotation of a cFOS-positive nucleus. This classification was used to calculate the ratio of cFOS-positive Parv-positive somata among all Parv-positive somata within each image.

2.10 Statistical analyses

All statistical analyses were performed using custom written code (Python, version 3.7.3; SciPy, version 1.4.1; Pingouin, version 0.3.8). The data was plotted either with OriginPro (version 2019b) or with custom written code (Python, version 3.7.3; matplotlib, version 3.3.1; seaborn, version 0.11.0). The box area in boxplots was defined

as the interquartile range (IQR, 1st to 3rd quartile) and the whiskers extend to the maximal or minimal values, but no longer than 150% of the IQR.

2.10.1 Statistical analysis of bioimage analyses

In all bioimage analyses, N represents the number of animals that were investigated and n reflects the number of images that were analyzed. All datasets were tested for significant outliers using Grubb´s test. Normal distribution and homoscedasticity of the data were assessed with Shapiro-Wilk and Levene´s tests, respectively.

For the comparison of DL-based annotation strategies, an image was excluded from the bioimage analysis if it was detected as significant outlier in several DL-based quantification results and if an expert could identify any abnormalities in the image (e.g. folding of the tissue). To ensure the comparability of the statistical results, exclusively non-parametric tests (Kruskal-Wallis-ANOVA followed by two-sided Mann-Whitney-U tests with Bonferroni correction for multiple comparisons) were used for all bioimage analyses to test for significant differences between the individual groups. If not indicated otherwise, the related statistical data can be found in a Dryad repository (`www.doi.org/10.5061/dryad.4b8gtht9d`; Segebarth, Griebel, Stein, R. von Collenberg, et al., 2020b).

After *consensus ensembles* were identified as most reliable, one *consensus ensemble* was used for the annotation of cFOS-positive nuclei and another one for the annotation of Parvalbumin-positive somata in two bioimage datasets (see chapter 4). In these bioimage analyses, an image was excluded if it was detected as significant outlier. Depending on the distribution of the data and the homogeneity of variances, parametric or non-parametric tests were used. For parametric tests, One-way ANOVA followed by two-sided t-tests with Welch-correction in case of unequal sample sizes and with Bonferroni correction for multiple comparisons were used. For non-parametric tests, Kruskal-Wallis-ANOVA followed by two-sided Mann-Whitney-U tests with Bonferroni correction for multiple comparisons were used. Detailed statistical information are provided in the respective figure legend. In order to test for each analyzed parameter for a statistical difference based on the genotype (WT, c*Bdnf* KO, or *Ntrk2*$^{+/-}$), the data from all experimental conditions of the indicated hippocampal subregion was pooled for each genotype and parametric (two-sided t-test with Welch-correction in case of unequal sample sizes) or non-parametric tests (two-sided Mann-Whitney-U test) were used accordingly.

Detailed statistical information are provided in the respective figure legend.

2.10.2 Statistical analysis of behavioral analyses

Also in behavioral analyses, N represents the number of investigated animals. The travelled distance and the percentage of time spent freezing of each mouse was computed for each session based on the motion tracking data. Significant outliers (Grubbs test) were excluded from the analyses and the data was assessed for normal distribution (Shapiro-Wilk test) and for the equality of variances (Levene´s test), and parametric or non-parametric tests were used accordingly. To test for an overall effect of genotype and time (session) on the freezing levels of the mice during context extinction, a mixed-ANOVA that takes repeated measures into account, was used (within factor: session, between factor: genotype). For the discrete comparison of freezing levels in each analyzed session between two groups of mice, a parametric t-test (with Welch correction for unequal sample-sizes) or a non-parametric Mann-Whitney-U test was conducted. Detailed statistical information are provided in the respective figure legend.

2.11 Material

The following tables list the solutions (Table 1), materials (Table 2), and chemicals (Table 3) that were used in this thesis.

Table 1: Buffers and solutions with their respective composition.

Buffer / solution	Composition
0.4% heparin	0.4% heprin-sodium 25000 in 1x PBS
10x phosphate buffered saline	80g NaCl, 2g KCl, 2g KH_2PO_4, 11.75g Na_2HPO_4 x 2 H_2O in 1 liter dH_2O
1x phosphate buffered saline	100 ml 10x PBS in 900 ml dH_2O
4% paraformaldehyde	4% (by weight) paraformaldehyde (PFA) was dissolved in dH_2O (half of final volume) with few drops of 5M NaOH under constant stirring for 20-30 min at 60°C. Dissolved PFA was passed through a paper filter and phosphate buffer was added to reach final volume. The pH was adjusted to 7.4
Blocking solution	0.3% Triton X100, 0.1% Tween 20, 10% horse serum in 1x PBS
Phosphate buffer	82% (by volume) 0.2M Na_2HPO_4 x 2 H_2O in dH_2O, 18% (by volume) 0.2M NaH_2PO_4 x 2 H_2O in dH_2O
Quenching solution	100mM glycine in dH_2O, pH was adjusted to 7.4 with Tris Base

Table 2: Materials with product name and supplying company.

Material	Company
24-well plates	Sarstedt Scientific
Cage, 1264C Eurostandard TypII (267 x 207 x 140mm)	Tecniplast
Injection needles 27G x 1/2″ (0.4 x 13mm)	Braun
Object slides (70 x 26mm)	R. Langenbrick
Razorblades superior platinum double edge	Astra
Syringe (1 ml)	BD Plastipak
Venofix Safety 25 G x 3/4″ (0.5 x 19 mm, length: 30 cm)	Braun
Veterinary fluosorber	Harvard Apparatus

Table 3: Chemicals with supplying company and product number.

Chemical	Company	Product number
Agarose	Biozym	840004
Aqua-Poly/Mount	Polysciences Inc.	18606
DAPI (4′,6-diamidino-2-phenylindole)	Sigma	D9542
Di-sodium hydrogen phosphate	Merck	106580
Ethanol	Sigma	32205
Glycine	Sigma	68898
Heparin-sodium 25000	Ratiopharm	
Horse serum	Linaris	SHD3250KYA
Isoflurane	Cp-pharma	
Ursotamin (100 mg/ml)	Serumwerk	
Medical oxygen	Rießner Gase	
Paraformaldehyde	Merck	1040051000
Potassium chloride	Sigma	P5405
Potassium di-hydrogen phosphate	Merck	104873
Sodium chloride	Sigma	31434
Sodium di-hydrogen phosphate	Merck	106342
Tris Base	Applichem	A2264,1000
Triton X-100	Carl Roth	3051.2
Tween 20	Applichem	A7932,0500
Xylavet	Cp-pharma	

3 Results I - Evaluation of DL-based strategies for bioimage analysis

3.1 DL-based strategies to perform bioimage analysis

Three DL-based strategies were designed to disclose potential shortcomings of DL-based bioimage analysis and to test whether these limitations can be overcome by adequate adaptations of the workflow (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

In the most straight-forward DL-based strategy, manual annotations of a single human expert are used to train an expert-specific CNN model (Figure 2 - gray, *expert models*) (Haberl et al., 2018; Falk et al., 2019). However, subjectivity among manual segmentations is known (Collier et al., 2003; Niedworok et al., 2016; Caicedo, Roth, et al., 2019; McQuin et al., 2018), and could be incorporated in and be reproduced by such expert-specific models (Falk et al., 2019; Chamier et al., 2019; Moen et al., 2019). Consequently, this would limit the use of DL to the mere automation of a potentially subjective, labor-intensive manual analysis approach (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

Alternatively, the input of multiple experts can be pooled in one training dataset, for instance by ground truth estimation algorithms (Warfield et al., 2004). Such a consensus training strategy aims at decreasing the impact of individual annotations, while increasing the impact of mutual segmentations, which are more likely to be objectively true. Training of a CNN model on mutual annotations of multiple experts could therefore favor the model to incorporate and reproduce these rather objective criteria. Thus, the use of training datasets based on ground truth estimations from multiple expert annotations could enable the use of deep learning to increase the objectivity of fluorescent label segmentation, going beyond its mere automation (Figure 2 – blue, *consensus models*). And yet, a certain degree of randomness during the training of DL algorithms can result in a significant model-to-model variability (Dietterich, 2000). By using similarity measures for the performance evaluation of trained models, this model-to-model variability could even go unnoticed, since models can reach similar performance scores on a common reference, yet by predicting non-identical segmentations. Consequently, this could add an additional level of irreproducibility (Segebarth, Griebel, Stein, R von

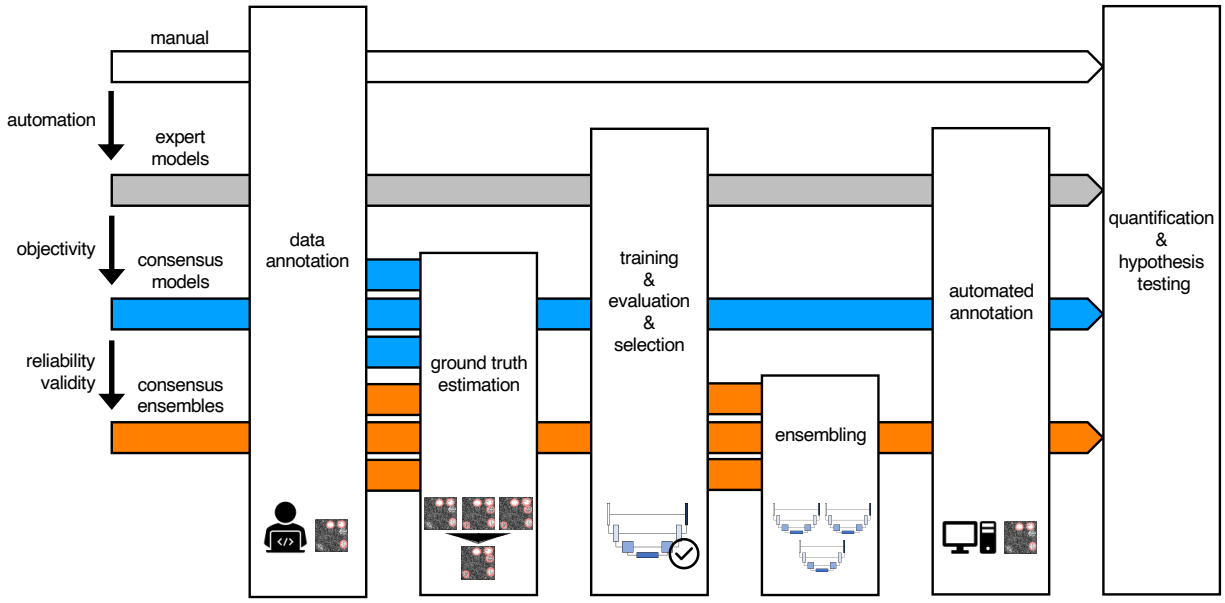


Figure 2: Schematic illustration of bioimage analysis strategies and corresponding hypotheses.
Four bioimage analysis strategies are depicted. Manual (white) refers to manual, heuristic fluorescent feature annotation by a human expert. The three DL-based strategies for automatized fluorescent feature 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. Reproduced from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a)

Collenberg, et al., 2020a).

The formation of model ensembles, for instance by averaging the output prediction of several trained models, can be effective in reducing noise, which is present in the predictions of individual models (Dietterich, 2000). In a third strategy, the output of several *consensus models* was, therefore, merged to form a *consensus ensemble*. The use of ensembles instead of individual models should decrease the discrepancies between predicted segmentations and consequently increase the reproducibility of DL-based bioimage analyses (Figure 2 – orange, *consensus ensembles*; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

3.2 Acquisition of a suitable bioimage dataset

In order to test the three DL-based strategies and the associated hypotheses, a bioimage dataset was used, which allows the quantification of changes in the abundance of the activity-related transcription factor cFOS (Greenberg and Ziff, 1984; Holtmaat and Caroni, 2016) in brain sections of mice after behavioral testing (Figure S1; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

The low signal-to-noise ratio of cFOS-positive nuclei (Shuvaev et al., 2017) fosters the presence of subjective manual annotations (Niedworok et al., 2016) and is thus well suited to test their impact on the DL models. Furthermore, manual segmentations cannot be used as a rigorously objective reference annotation (ground truth), which impedes the validation of the DL-based annotations on the level of individual images. Instead, correlations of changes in the abundance of cFOS with experimental treatment conditions that are well-established in the scientific literature, could serve as a secondary ground truth. The quantification of immediate-early genes, like cFOS, after behavioral testing is commonly used in the field of neuroscience (Gallo et al., 2018) and the results of bioimaging studies with a similar design can, therefore, be used as reference (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

Two bioimage datasets of fluorescently labeled cFOS in brain sections of mice after contextual fear conditioning were acquired in the course of this thesis. Classical "Pavlovian" conditioning refers to a type of behavioral experiments that trigger an associative learning process. During contextual fear conditioning, exposure to a neutral context (conditioned stimulus, CS) is paired with the presentation of an aversive stimulus (unconditioned stimulus, US). This leads to the formation of an associative, contextual fear memory and re-exposure to the CS alone on the following day is sufficient to elicit a species-specific defensive behavior, such as freezing in mice (LeDoux, 2000).

Each of the two bioimage datasets comprises three treatment groups (Figure 1). One group of mice underwent Pavlovian contextual fear conditioning and was re-exposed to the conditioning context 24 hours later (context with shocks, C+). A second group of mice served as context control group and was also exposed twice to the context, but without presentations of the aversive stimulus (context without shocks, C-). Mice that were directly taken out of their home-cage served as naive learning control group (H). Brain sections were prepared either 90 minutes after memory retrieval for C- and C+

mice, or at a comparable point in time for home-cage controls. The neuronal activity-related protein cFOS (Greenberg and Ziff, 1984; Holtmaat and Caroni, 2016), the calcium-binding protein Parvalbumin (Hu et al., 2014), and the neuronal marker NeuN (Fox3) were labeled by indirect immunofluorescence and images were acquired using a confocal microscope. In each of the two bioimage datasets, wildtypic mice were compared to a knock-out mouse model. The results of these comparisons are presented in detail in chapter 4 of this thesis (*Results II - Bioimage analysis of two datasets with consensus ensembles*). The following evaluation of the DL-based strategies is solely based on the data of wildtypic mice (WT), which was pooled from both datasets. Wildtype mice that underwent contextual fear conditioning (C+) showed significantly more freezing during the training session (Acq), compared to context control mice (C-) (Figure 3). Furthermore, the time in which the animals displayed freezing behavior during the retrieval session was significantly higher in C+ mice, compared to the C- group (Figure 3). Similarly, C+ mice traveled significantly shorter distances during Ret than C- mice (Figure 3).

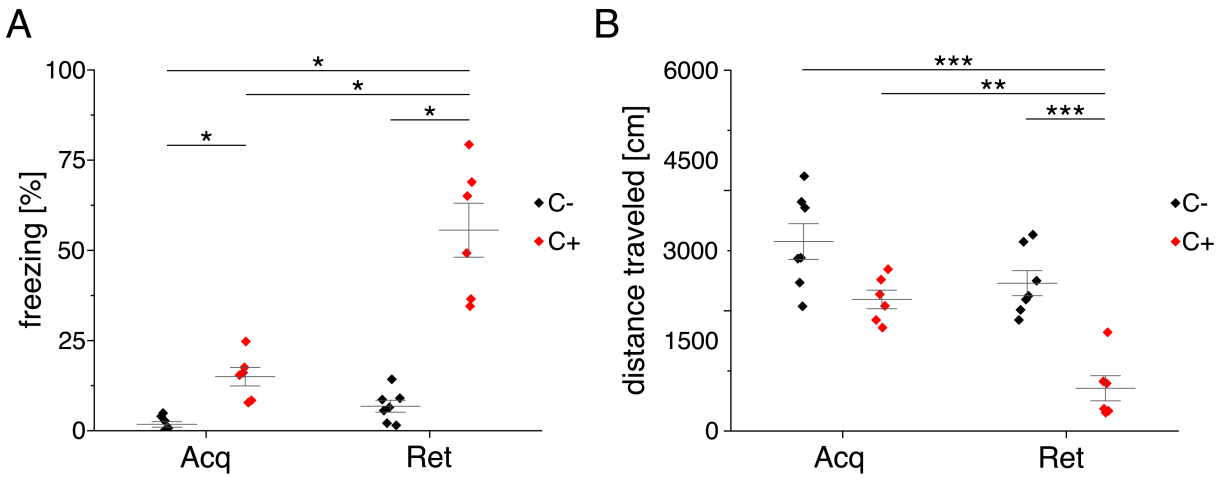


Figure 3: Behavioral analysis.
A. Fear acquisition was observed in conditioned mice (C+), while unconditioned controls (C-) did not show freezing behavior during initial context exposure (Acq). In the memory retrieval session (Ret), conditioned mice showed strong freezing behavior, while unconditioned mice did not freeze in response to the training context ($X^2(3)=20.894$, $p<0.001$, $N_{(Acq\ C-)}=7$, $N_{(Acq\ C+)}=6$, $N_{(Ret\ C-)}=7$, $N_{(Ret\ C+)}=6$, 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 ($F_{(3,\ 22)}=19.484$, $p<0.001$, $N_{(Acq\ C-)}=7$, $N_{(Acq\ C+)}=6$, $N_{(Ret\ C-)}=7$, $N_{(Ret\ C+)}=6$, one-way ANOVA followed by pairwise t-tests with Bonferroni correction, **: $p<0.01$, ***: $p<0.001$). Adapted from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

3.3 Model training, selection, and validation

We used a set of 36 images and corresponding binary segmentation masks to train the DL models (Figure S1). Depending on the strategy which was used to perform DL-based bioimage analysis (Figure 2), the binary segmentation masks resembled either the manual annotations of a single expert (for *expert models*), or the estimated ground truth (est. GT) as the result of the ground truth estimation process (for *consensus models* and *consensus ensembles*). In order to avoid any bias of the trained models due to imbalanced representations of the classes in the training dataset (Moen et al., 2019), the set of 36 images was chosen to equally represent all analyzed hippocampal regions (DG, CA3, and CA1: 12 images each) and the investigated treatment conditions (H, C-, and C+: 12 images each, four for each hippocampal region). To artificially increase the amount of training data, data augmentation using image transformations and elastic deformations was performed, as suggested by Falk et al. (2019). Finally, the augmented training dataset was split into a train and a validation set. Since the images of the validation set are withhold during training of the algorithm, they allow the evaluation of the model after each training epoch without the risk of information leakage (A. Zheng and Casari, 2018). For each model, the epoch with the highest performance on the validation set was selected for further analyses, once the training was concluded (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

Performance was assessed by comparing the predicted annotations of the model with the corresponding reference annotations (e.g. est. GT for *consensus models*), using two similarity measures that describe the quality of segmentations (IoU) and of feature detection (F1-score). First, the intersection over union (IoU) was calculated for all overlapping pairs of regions of interest (ROIs) of the two segmentation masks (Figure 4A). This comparison allows to assess the overall similarity of annotations that are represented in both segmentation masks (mean IoU). However, this metric is unaffected by ROIs of the reference annotation that are missing in the predicted annotation, or by excessive ROIs that are present solely in the predicted annotation. To account for this, the F1-score was computed. For this detection metric, only pairs of ROIs with an IoU of at least 0.5 were considered as matching, while all other ROIs, including non-overlapping ROIs, are considered as non-matching (Maška et al., 2014). The F1-score was then calculated as the harmonic mean of precision and recall (Figure 4B). Thus, the F1-score includes a comparison of segmentation accuracy and takes both excessive and missing ROIs into

account. Therefore, model selection was eventually based on the median F1-score across all validation images (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

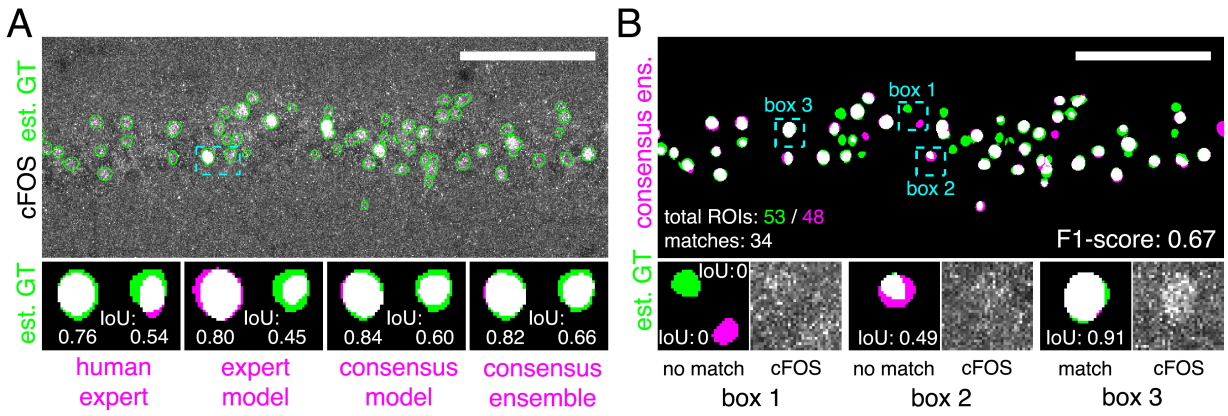


Figure 4: Illustration of the analyzed similarity measures.
A. Representative example of IoU calculations on a field of view (FOV) in a bioimage. Image raw data show the labeling of cFOS in a maximum intensity projection image of the CA1 region in the hippocampus (brightness and contrast enhanced). The 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 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. Reproduced from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

In addition, the F1-scores of the selected *consensus models* were compared to those between the manual annotations of all human experts and the est. GT annotations on the validation set. Importantly, a *consensus model* was only considered to be valid and subsequently used for the analyses, if its F1-score was higher than the lowest F1-score among all experts on each validation image. This additional performance test is only possible, if the annotations of multiple human experts are available and was, therefore, omitted for *expert models*. In total, 20 *expert models* (four for each expert) and 36 valid *consensus models* were created (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

In order to determine how many *consensus models* are to be merged into one *consensus ensemble*, the F1-scores among the predicted annotations of multiple *consensus ensembles* with incrementally increasing numbers of pooled *consensus models* were compared (Figure 5). Here, the F1-scores between the predictions of ensembles consisting only of a single *consensus model* consequently indicate the discrepancy between the predictions of individual *consensus models*. Increasing the number of pooled *consensus models* led to increased F1-scores between the annotations of the resulting *consensus ensembles*. On these data, pooling of more than four *consensus models* into a single *consensus ensemble* did not result in further increases of the F1-scores (Figure 5). Consequently, the size of *consensus*

ensembles was defined as four *consensus models* for this study (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

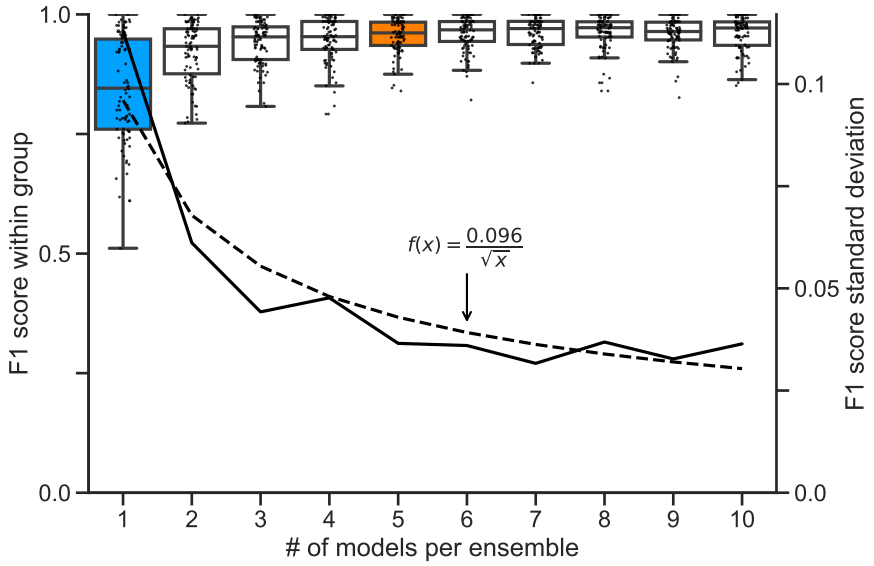


Figure 5: Ensemble size and reliability.
To determine an appropriate size for the *consensus ensembles*, the homogeneity of the results was analyzed through a similarity analysis. Therefore, the Mean F1-scores at an IoU matching threshold of $t = 0.5$ were calculated for each ensemble size $i \in \{1, ..., 10\}$ on the holdout test set (n=9 images). Stratified on the cross validation splits the ensembles were randomly sampled from a collection of trained *consensus models*. This procedure was repeated five times to mitigate the random effect of the ensemble composition ($N_{ensembles}$=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. The black line denotes the standard deviation of Mean F1-scores, which is scaled at the right y-Axis. The dashed black line denotes the best fitting function of type $f(x) = \frac{a}{\sqrt{x}}$ with $a = 0.096$ for the standard deviation. Reproduced from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

3.4 Performance evaluation on the level of similarity analysis

The initial evaluation of the three DL-based strategies was based on similarity measures. Since the models were selected on base of their performance on the validation set, another set of nine images (test set, Figure S1B) was chosen to compare their performance on new data with each other and to that of human experts. In addition, these images were used to test for potential subjectivity among the annotations of the experts. As expected, similarity analyses of the manual expert annotations revealed only modest agreement between the experts (Figure 6D; Schmitz et al., 1999; Collier et al., 2003; Niedworok et al., 2016; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a. In addition, more detailed analyses indicate, that the agreement of human experts was cor-

related with the relative intensity difference between the annotated features and their surrounding background (Figure S2; Niedworok et al., 2016; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

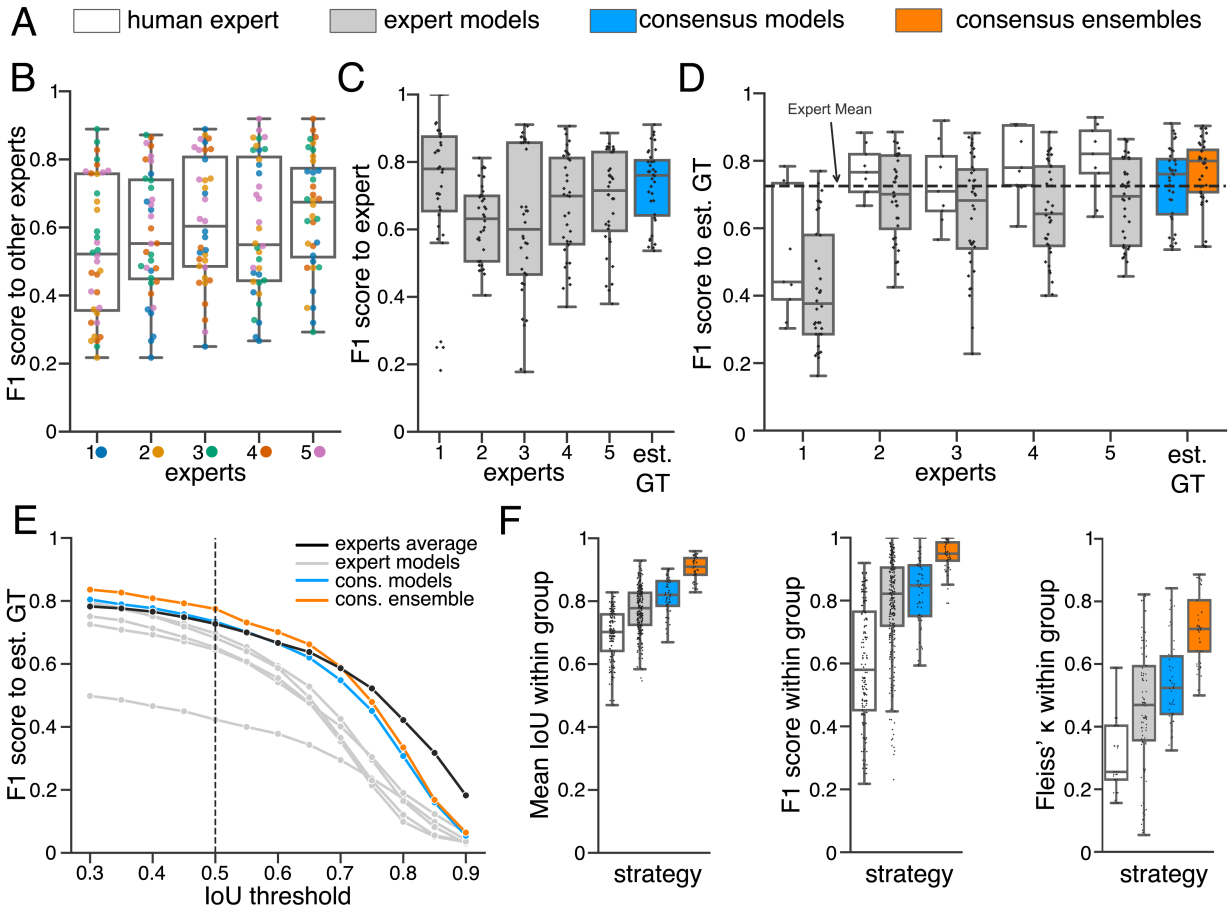


Figure 6: Similarity analysis of fluorescent feature annotations by manual or DL-based strategies.
A. Color coding refers to the individual strategies, as introduced in Figure 2 (white: manual approach, gray: *expert models*, blue: *consensus models*, orange: *consensus ensembles*).
B. Mean F1-scores between individual manual expert annotations and their overall reliability of agreement given as the mean of Fleiss' κ.
C. Mean F1-scores between annotations predicted by individual models and the annotations of the respective expert (or est. GT), whose annotations were used for training. $N_{(\text{models per expert})}$=4.
D. Mean F1-scores 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. $N_{(\text{models})}$=4, $N_{(\text{ensembles})}$=4.
E. Means of mean F1-scores 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 $t = 0.5$. $N_{(\text{models})}$=4, $N_{(\text{ensembles})}$=4.
F.) Annotation reliability of the individual strategies assessed as the similarities between annotations within the respective strategy. Mean IoU, mean F1-scores, and Fleiss' κ were calculated. $N_{(\text{experts})}$=5, $N_{(\text{models})}$=4, $N_{(\text{ensembles})}$=4. Adapted from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

As during model selection on the validation dataset, the predicted segmentations of the trained models were compared to the annotations of the respective coder (expert or est. GT), whose annotations were used to train the model. In this comparison, all models

reached similar F1-scores, indicating that all training datasets were representative of the task and in general of equal quality (Figure 6E). Notably, the median F1-scores between the predicted annotations and the reference annotations were, in most cases, higher than those between the segmentations of the human experts (Figure 6D and 6E) (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

However, only when all segmentations were compared with each other (Figure S3 and Figure S4) or to the est. GT as common reference (Figure 6F and 6G), differences among both, the experts and the DL-based strategies, became apparent. In case of manual annotations for instance, particularly those of expert 1 showed low agreement with the annotations of the other experts or with the est. GT (Figure 6, Figure S3, and Figure S4). Interestingly, *expert models* shared the same tendencies in terms of agreement with other segmentations as their respective coder, yet with overall lower F1-scores (Figure 6F and Figure S3). Consequently, the median F1-scores of all *expert models* were below the expert average when compared to the est. GT (Figure 6E and 6G). In contrast, the annotations of both *consensus models* and *consensus ensembles* were on par with human experts, even at higher IoU thresholds (Figure 6F, 6G) (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

Moreover, bioimage analysis has to be reliable and reproducible for researchers to draw valid conclusions from their experiments. Therefore, image features should be annotated with high consistency and there should, ideally, be no variability when the bioimage analysis process is reproduced. While inter- and intra-rater variability are known phenomena for manual annotations (Collier et al., 2003), each model outputs the identical segmentations once its training is concluded. However, the consistency of predicted segmentations across models trained on the identical training dataset remained elusive. Consequently, the similarities between the predicted segmentations of all models or ensembles within each DL-based strategy, i.e. 20 *expert models*, 36 *consensus models*, and 9 *consensus ensembles*, were calculated. These analyses revealed significantly higher mean IoU and F1-scores of all DL-based strategies, compared to the inter-rater agreement of manual analyses (Figure 6H). Notably, the segmentations derived from *consensus ensembles* showed the highest within-group similarities among all tested strategies (Figure 6H). Fleiss' kappa is another metric to assess the reliability of agreement between several coders, which also takes the chance of randomly overlapping ROIs into account (Fleiss and Cohen, 1973). Again, *consensus ensembles* scored highest among all DL-based strategies and compared to the human experts (Figure 6H) (Segebarth, Griebel, Stein, R

von Collenberg, et al., 2020a).

Together, the initial characterization of the manual expert annotations and of the three DL-based strategies on the level of similarity analyses supports the aforementioned hypotheses (Figure 2). First, these data show that manual annotations of human experts on these images only have a "fair agreement", according to the interpretation of Fleiss' kappa values by Landis and Koch (1977). Furthermore, comparing all expert annotations with each other and to the est. GT as a common reference indicates a substantial level of subjectivity, particularly documented in the annotations of expert 1. As hypothesized, individual *expert models* could learn and reproduce these biases, yet with overall lower similarity measures (Figure 6F, Figure S3, and Figure S4). On the contrary, the training on est. GT annotations resulted in *consensus models* and *consensus ensembles* that reached expert-level performance (Figure 6F, Figure 6G, Figure S3, and Figure S4). Notably, *consensus ensembles* significantly outperformed all other approaches in reproducibility measures (Figure 6H; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

3.5 Performance evaluation on the level of bioimage analysis

The main goal of bioimage analysis is the unbiased quantification of image features and the subsequent statistical testing of a biological hypothesis (Meijering et al., 2016). Notably, the results of the similarity analyses indicate that DL-based annotations can be limited in both objectivity and reproducibility (Figure 6, Figure S3, and Figure S4), which could impair the quality of bioimage analyses. However, it remains unclear whether performance on the level of bioimage analysis can be inferred directly from performance on the level of similarity analysis. Therefore, the final evaluation of the three DL-based strategies was performed on the level of bioimage analysis (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

For this, a total of 283 images showing cFOS signals in the hippocampus of wild type mice after behavioral testing was used (Figure S1; Segebarth, Griebel, Stein, R. von Collenberg, et al., 2020b). This bioimage dataset allows to test for significant differences in the number of cFOS-positive nuclei and their mean signal intensities, between three experimental conditions (H, C-, and C+) in a total of five hippocampal subregions (DG as a whole, suprapyramidal DG, infrapyramidal DG, CA3, and CA1) (Figure 7B-D). For

each model and model ensemble, these analyses were performed individually, based on their predicted annotations. Eventually, the DL-based bioimage analysis results were compared across all 30 pairwise comparisons, using the calculated p-values and effect sizes (η^2). Since both measures represent summary statistics, Figure 7E is dedicated to illustrate the relationship between both measures and the underlying, individual data points. Here, the number of cFOS-positive nuclei in the *stratum pyramidale* of CA1 is compared between the three treatment groups as a representative example. For each DL-based strategy, the analyses of two distinct models or model ensembles were selected and represent the minimal and maximal effect sizes reported within each strategy. All bioimage analyses reveal a significantly higher amount of cFOS-positive nuclei in CA1 in mice after retrieval of a contextual memory (C- and C+), compared to home-cage controls (H) (Figure 7E). Notably, these quantifications already indicate that the variability of effect sizes is highest among *expert models* and lowest among the bioimage analysis results based on the annotations of *consensus ensembles* (Figure 7E, Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

Next, this comparison was extended with the results of all 20 *expert models*, 36 *consensus models*, and 9 *consensus ensembles*, for all 30 pairwise comparisons, to assess the variability of bioimage analysis results within the three DL-based strategies (Figure 8). For instance, the bioimage analyses results of cFOS-positive nuclei in the stratum pyramidale of CA1 of all models and ensembles, instead of only the two most extreme examples (Figure 7E), further highlights the low variability of effect sizes among *consensus ensembles*, compared to the two alternative DL-based strategies (Figure 8A - #cFOS+ nuclei / left). In turn, a high variability of effect sizes can eventually result in differences in the reported statistical outcome between individual models or ensembles, like in the case of the analyses of the mean cFOS signal intensity in CA1 based on the annotations of *expert models* (Figure 8A - mean cFOS signal intensity / right). Here, four of the 20 *expert models* detect no significant differences in the mean cFOS signal intensity, another two *expert models* indicate only a significant difference between H and C- mice, while all other 14 *expert models* reveal a significant, context-dependent increase, in line with all *consensus models* and *consensus ensembles* (Figure 8A - mean cFOS signal intensity / right). Interestingly, all of these four models that detect no significant difference in the mean cFOS signal intensity in CA1, were trained on the annotations of expert 1 (Figure 8A). Overall, DL-based bioimage analyses revealed significant context-dependent increases in the abundance of cFOS in most of the investigated regions of the dorsal

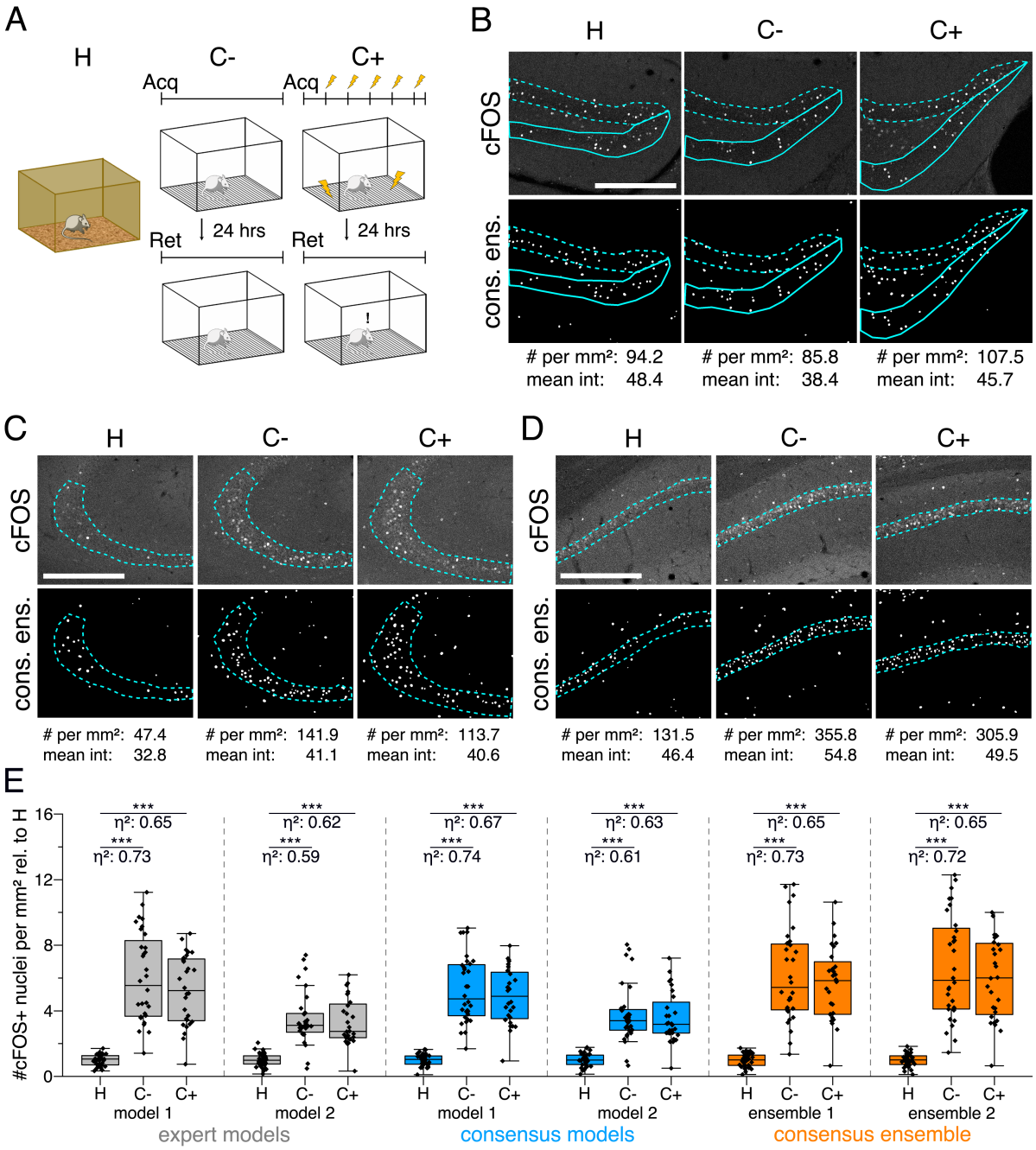


Figure 7: Application of different DL-based strategies for fluorescent feature annotation.
The figure introduces how three DL-based strategies are applied for annotation of a representative fluorescent label, here cFOS, in a representative image data set. Raw image data show behavior-related changes 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 (Acq), mice were re-exposed to the training context for memory retrieval (Ret). Memory retrieval induces changes in cFOS levels.

Figure 7 continued on next page

Figure 7 continued
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: $N_{(\text{expert models})}$=20, $N_{(\text{consensus models})}$=36, $N_{(\text{consensus ensembles})}$=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 with Mann-Whitney-U test. Statistical data are available in Segebarth, Griebel, Stein, R. von Collenberg, et al. (2020b). Adapted from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

hippocampus (Figure 8A-D). Only the analyses of the infrapyramidal blade of the DG did not show any significant differences between the three experimental groups (Figure 8E; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

In addition, for each analysis and all 30 pairwise comparisons, the respective difference between two groups was classified as *significant* ($p \leq 0.05$) or *not significant* ($p > 0.05$). These results were accumulated from all analyses within a DL-based strategy to calculate a majority vote for each pairwise comparison within each strategy. Interestingly, these majority votes were identical for all DL-based strategies, contrasting the divergence among the results of individual models and model ensembles (Figure 8) (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

The variability of results among individual models or model ensembles was analyzed in more detail as the *variation per effect* and as the *variation per model*. For the variation per effect, the standard deviation of the effect sizes within each DL-based strategy was calculated for each pairwise comparison. This revealed significantly lower standard deviations for the *consensus ensembles* compared to both alternative strategies (Figure 8F - *variation per effect*). In addition to comparing the overall reliability of each DL-based strategy, the *variation per model* was computed to assess the reliability of an individual model or model ensemble. The *variation per model* is plotted as the interaction between the number of pairwise comparisons, where the results of the corresponding model (or ensemble) differed from the congruent majority votes, and the standard deviation of centered effect sizes across all 30 analyzed effects. Strikingly, these analyses show that all of the 20 *expert models* report the statistical significance of at least one pairwise comparison differently from the majority votes (Figure 8F - *variability per model*). As indicated by the *variation per effect*, the reliability of *consensus models* is increased com-

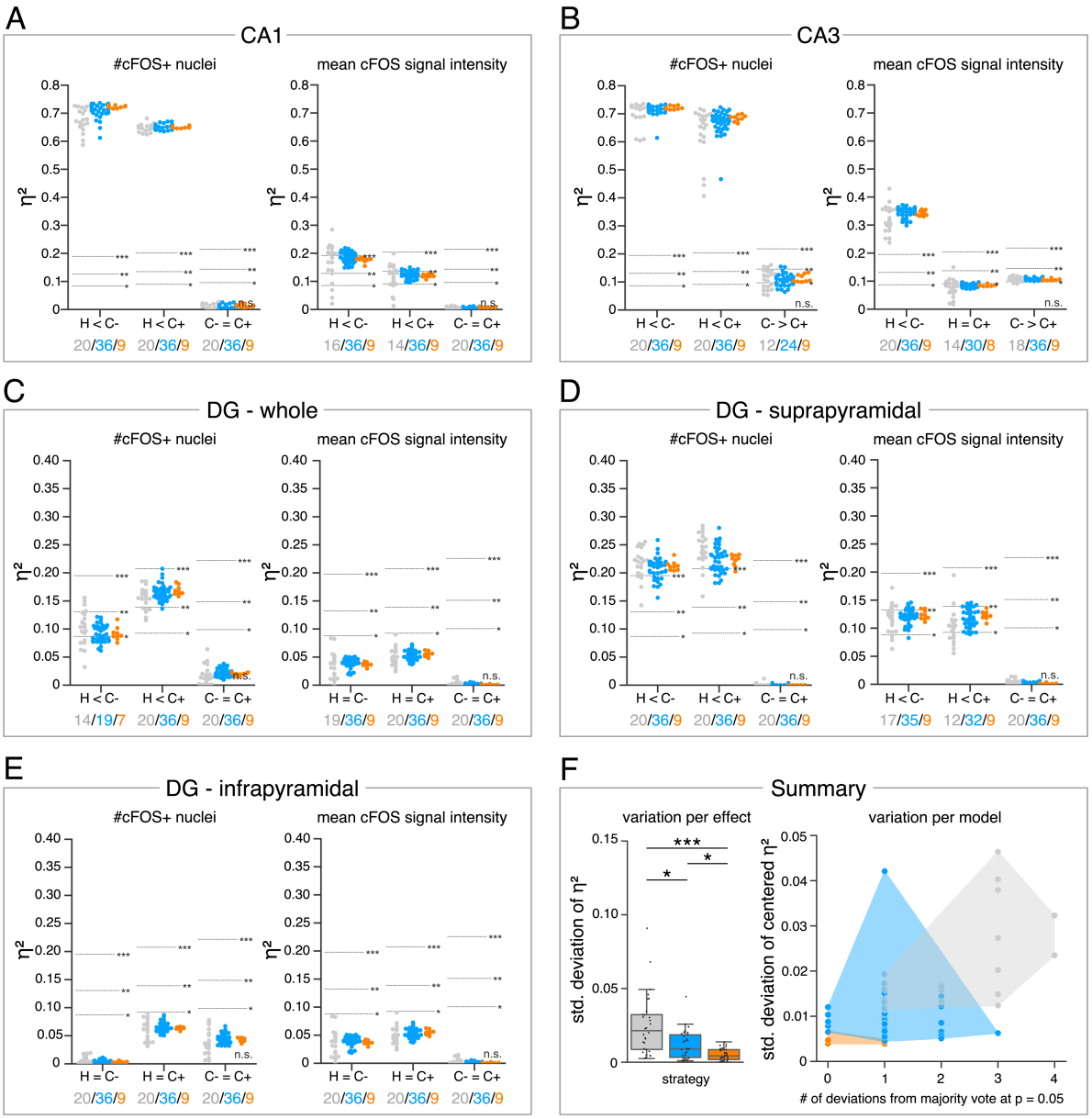


Figure 8: Consensus ensembles significantly increase reliability of bioimage analysis results.
A-E. Single data points represent the calculated effect sizes for each pairwise comparison of all individual bioimage analyses for each DL-based strategy (gray: expert models, blue: consensus models, orange: 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. Source files including source data and statistical data are available in Segebarth, Griebel, Stein, R. von Collenberg, et al. (2020b).
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.
Figure 8 continued on next page

Figure 8 continued
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*, single data points represent the standard deviation of reported effect sizes (η^2), calculated 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, $N_{(effects)}$=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≤0.05 after Bonferroni correction for multiple comparisons) were determined for each individual model and ensemble across all pairwise comparisons (x-axis). Visualizing the interaction of both measures for each model or model ensemble individually reveals that consensus ensembles show the highest reliability of all three DL-based strategies. The statistical data for the for variation per effect is available in Segebarth, Griebel, Stein, R. von Collenberg, et al. (2020b). Reproduced from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

pared to *expert models*, but is highest for the bioimage analysis results of *consensus ensembles* (Figure 8F - *variability per model*; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

Taken together, these data show that the reliability of DL-based bioimage analysis can be heavily impaired, particularly if individual models are used and training data annotations are derived from a single human expert. However, training on the pooled input of multiple experts by means of ground truth estimation in conjunction with the formation of model ensembles significantly reduced this variability and consequently increased the reliability of bioimage analysis results by a large margin (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

All together, the comprehensive evaluation of the three DL-based strategies on this dataset confirmed all initial hypotheses, both on the level of similarity analyses and, more importantly, also on the level of bioimage analyses. First, these results confirm and extend the concerns put forward by previous studies, by demonstrating that an ordinary level of subjectivity among five human expert neuroscientists is sufficient to significantly impact the subsequent bioimage analyses of *expert models* (Falk et al., 2019; Chamier et al., 2019). These analyses also show that the annotations of *consensus models*, which are trained on the pooled input of multiple experts have, on average, a higher validity than those of *expert models*. And yet, this resulted only in a modest increase in the reliability of bioimage analysis. However, the formation of *consensus model ensembles* led to a significant increase in the homogeneity of predicted segmentations and a concomitant increase in the reliability of bioimage analyses (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

4 Results II - Bioimage analysis of two datasets with *consensus ensembles*

The previous chapter (*3 Results I - Evaluation of DL-based strategies for bioimage analysis*) compared three DL-based strategies for bioimage analyses with a focus on objectivity, reliability, and validity. These data established that annotations of *consensus ensembles* and the bioimage analyses derived from these annotations are less subjective and more reliable than that of conventional DL-based strategies, such as creating individual models trained on the annotations of a single human expert (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

In this chapter, *consensus ensembles* and their applicability to large bioimage datasets are continuously assessed by using a combination of two *consensus ensembles* to analyze cFOS-positive nuclei and the somata of Parvalbumin-positive interneurons in two bioimage datasets. Together, these data comprise a total of more than 650 images per fluorescent label and were derived from two mouse models with defined genetic defects in a learning-related neurotrophin signalling cascade.

4.1 Investigated mouse models

The most important neurotrophin signaling cascade in the central nervous system (CNS) is the so-called BDNF-TrkB signaling cascade (Sasi et al., 2017). BDNF (brain-derived neurotrophic factor) was identified in 1982 as a secretory protein that is involved in the survival of a subtype of peripheral neurons (Y. A. Barde et al., 1982; Thoenen, 1995). In the CNS, however, BDNF is predominantly involved in the regulation of neuronal circuit functions (Sasi et al., 2017), and less in mediating neuronal survival (Rauskolb et al., 2010; Sairanen et al., 2005). The tropomyosin-receptor kinase B (TrkB) is the physiological, high-affinity receptor of BDNF and is the most abundant Trk receptor in the CNS (Barbacid, 1994; Klein et al., 1989; Martin-Zanca et al., 1986). The investigated mouse models allow to compare the distribution of cFOS immunofluorescence signals after genetic manipulation of either BDNF or TrkB expression.

In one of the analyzed bioimage datasets, the data of wildtypic mice is compared to that of conditional *Bdnf* knock-out mice (c*Bdnf* KO; Sasi, 2020; Rauskolb et al., 2010). The images of the second dataset were acquired from wildtypic and from heterozygous

Ntrk2 (neurotrophic tyrosine kinase receptor type 2, the TrkB-encoding gene) knock-out mice (*Ntrk2*$^{+/-}$; Rohrer et al., 1999).

4.1.1 Conditional knock-out of BDNF from a sparse population of adult dentate gyrus granule neurons

Recently, a mouse model was established to study the effects of the genetic deletion of presynaptic BDNF from adult hippocampal mossy fiber terminals (c*Bdnf* KO; Sasi, 2020). For this, the expression of Cre recombinase is restricted primarily to adult granule neurons of the DG using the CNTF::Cre knock-in mouse model (C57BL/6-Tg(CNTF-Cre)TM^{Msd}), which was created by Dr. Yasuhiro Ito under the supervision of Prof. Michael Sendtner at the Institute of Clinical Neurobiology, Würzburg. This knock-in model was then combined with a mouse model that allows the Cre-mediated deletion of the *Bdnf* gene (Rauskolb et al., 2010). This results in a robust deletion of BDNF from a sparse population of adult DG granule neurons (Sasi, 2020).

Initial behavioral characterization of this mouse model included tests that assess general behavior and appearance (SHIRPA), heat sensation via the paws (Hot Plate), neuromuscular function and muscle power (grip strength, rotor rod), hippocampal function (Morris water maze), and anxiety-like behavior (Open Field, Elevated Plus Maze, Dark-Light Box), but no apparent behavioral difference between c*Bdnf* KO mice and wildtypic controls could be observed (work by Dr. Cora Rüdt von Collenberg, Dr. Britta Wachter, Dr. Thomas Seidenbecher, and Dr. Robert Blum). However, this work also revealed that c*Bdnf* KOs showed significantly less freezing during the retrieval of a contextual fear memory, similarly to what can be observed in mice that express a human *Bdnf* polymorphism that reduces the activity-related release of BDNF (Chen et al., 2006). In addition, these mice showed delayed extinction learning after a cue fear conditioning paradigm in a background context. Moreover, *in vivo* electrophysiological recordings from freely moving mice during this extinction paradigm, revealed significantly higher neuronal activity in the CA1 region of the dorsal hippocampus, while neuronal activity in the infralimbic cortex was not altered.

Therefore, using a *consensus ensemble* for the bioimage analysis of cFOS signals could add a second evidence for the elevated neuronal activity in CA1 in these c*Bdnf* KO mice. Moreover, it could also extend these results with the investigation of all hippocampal subregions and of the population of Parvalbumin-positive interneurons. Vice versa, the

detection of such effects could also serve as additional verification of the validity of the DL-based bioimage analysis.

4.1.2 Heterozygous knock-out of TrkB

The *Ntrk2* gene encodes the the high-affinity receptor of BDNF, TrkB (Rodriguez-Tebar and Y. A. Barde, 1988). To gain more insights into the implications of aberrant BDNF-TrkB signaling and to examine potential similarities to c*Bdnf* KO mice, heterozygous *Ntrk2* knock-out mice (*Ntrk2*$^{+/-}$, Rohrer et al., 1999) were also included in these bioimage analyses. In this model, all variants of the high-affinity receptor for BDNF are knocked-out (Rohrer et al., 1999).

4.2 Behavioral analysis in a contextual fear extinction paradigm

At first, a contextual fear extinction paradigm was used to examine, whether these mice exhibit any behavioral phenotype in tests that specifically assess the processes and brain networks that are involved in the acquisition, the retrieval, and the extinction of contextual fear memories (LeDoux, 2000; Tovote et al., 2015).

For this, mice underwent contextual fear conditioning (acquisition session, Acq) in a training context and were re-exposed to this context for a total of seven additional sessions. The first re-exposure (retrieval session, Ret), took place 24 hours after the initial Acq. Another 24 hours following this retrieval session, mice were re-exposed to the training context twice per day for three consecutive days (extinction sessions 1-6, Ext1 - Ext6).

After an initial increase of the freezing levels after the Acq, all mice showed a gradual decrease of freezing behavior over time (significant main effect for time, Figure 9A and Figure 10A). These freezing levels were not different between c*Bdnf* KO mice and WT littermate controls (no significant main effect for genotype, Figure 9). Heterozygous *Ntrk2*knock-outs, on the contrary, displayed significantly higher freezing levels over the course of the experiment compared to WT controls (significant main effect for genotype, Figure 10). However, the gradual decrease of freezing levels was similar in both groups and there was no significant interaction between time and genotype, indicating effective extinction learning also in these mice (Figure 10).

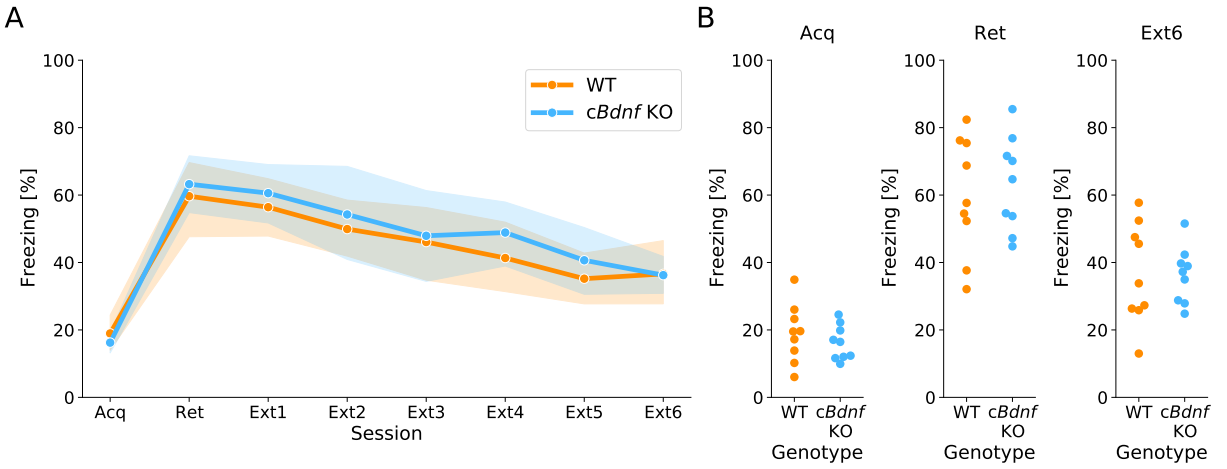


Figure 9: Freezing during contextual fear extinction of c*Bdnf* KO mice compared to WT littermates.
A. Mean freezing levels in percent (95% confidence intervals) during all sessions. Statistical analysis was performed with a mixed-ANOVA (within-factor: time, between-factor: genotype). Significant main effect of time ($F_{(7,112)}$=23.828, p<0.001, partial η^2=0.598), no significant main effect for genotype ($F_{(1,16)}$=0.388, p=0.542, partial η^2=0.024), and no significant interaction of time x genotype ($F_{(7,112)}$=0.327, p=0.941, partial η^2=0.020). $N_{(WT)}$=9, $N_{(cBdnf\ KO)}$=9.
B. Freezing levels in percent of individual mice in the indicated sessions (Acq, Ret, and Ext6) of **A**. There were no significant differences between the groups detailed statistical data are provided in Table 4.

4.3 Bioimage analysis results

As described above (*3.2 - Acquisition of a suitable bioimage dataset*), mice were subdivided into three experimental groups (H, C-, C+) and brain sections were prepared either 90 minutes after retrieval of a contextual memory (C- and C+), or at a comparable time during the day (H). Three anatomically defined brain sections of each animal were immunofluorescently labelled for cFOS, Parvalbumin (Parv), and NeuN, and confocal image-stacks of the investigated subregions of the dorsal hippocampus were acquired. Two *consensus ensembles* were trained on the estimated ground truth annotations derived from the manual annotations of five human experts either of cFOS-positive nuclei, or of Parv-positive somata. Validation of the predicted annotations against the annotations of the five human experts confirmed expert-like performance of both ensembles. These *consensus ensembles* were then used for the unbiased analyses of cFOS-positive nuclei and Parv-positive somata.

In each bioimage dataset, wildtypic mice were included as reference, positive controls, and for normalization purposes. In order to indicate for which of the two bioimage datasets an individual WT mouse was used, a color coding within the group of WT mice is used. Orange markers represent data derived from wildtypes that were included in the bioimage dataset of c*Bdnf* KO mice, and green markers indicate that the

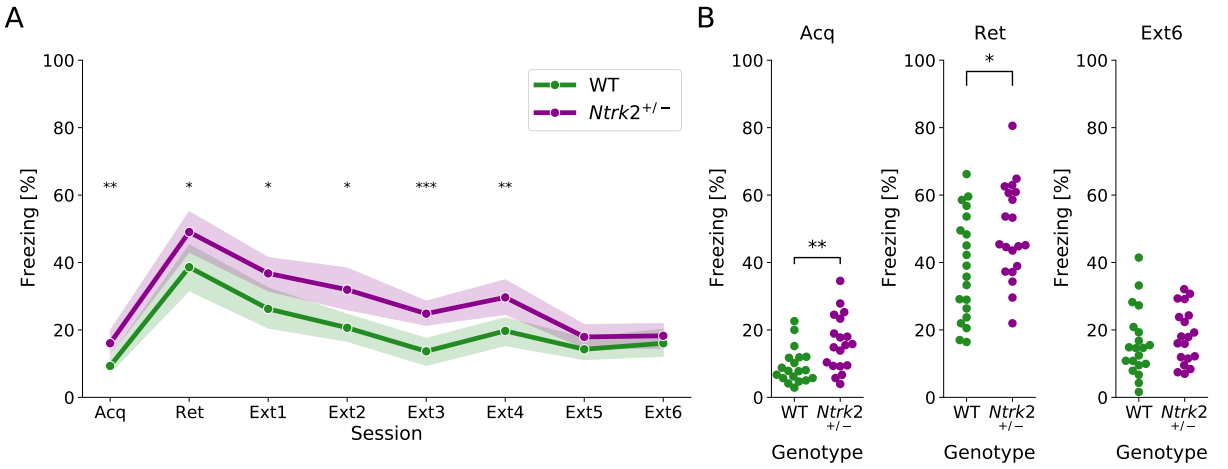


Figure 10: Freezing during contextual fear extinction of *Ntrk2*$^{+/-}$ mice compared to WT littermates.
A. Mean freezing levels in percent (95% confidence intervals) during all sessions. Statistical analysis was performed with a mixed-ANOVA (within-factor: time, between-factor: genotype). Significant main effect of time ($F_{(7,266)}$=53.433, p<0.001, partial η^2=0.584), significant main effect for genotype ($F_{(1,38)}$=13.238, p=0.008, partial η^2=0.258), and no significant interaction of time x genotype ($F_{(7,266)}$=1.770, p=0.094, partial η^2=0.045). Results of post-hoc pairwise comparisons of freezing levels per session between the two groups are indicated if significance was reached (*: p<0.05, **: p<0.01, ***: p<0.001) and detailed statistical data are provided in Table 4. $N_{(WT)}$=20, $N_{(Ntrk2^{+/-})}$=20.
B. Freezing levels in percent of individual mice in the indicated sessions (Acq, Ret, and Ext6) of **A**. *Ntrk2*$^{+/-}$ mice showed significantly more freezing during Acq and Ret, but there were no significant differences between the groups during the last extinction session (Ext6). Detailed statistical data are provided in Table 4.

data originates from a WT mouse that was used as control in the bioimage dataset of heterozygous *Ntrk2*knock-out mice.

In addition, the data of WT mice used for the analyses of DL-based strategies partially overlaps with the data used in these analyses. In fact, all data derived from WT mice that served as controls for c*Bdnf* KO mice (orange), were also part of the dataset that was used for the comparison of DL-based strategies. In addition, one WT control of each treatment condition of the *Ntrk2*$^{+/-}$ bioimage dataset was also used for the comparison of DL-based strategies, whereas the data of four wildtype mice (two H, one C-, and one C+) was exclusively used in the subsequent bioimage data analyses. Moreover, three WT mice (one of each treatment condition), which were used for the analyses of DL-based strategies, had to be excluded from the bioimage analyses of the two knock-out mouse models, since the Parvalbumin immunofluorescence staining was insufficient.

Behavioral analysis of the retrieval session (retrieval of learned fear) of all mice that were used for the following bioimage analyses shows that context conditioned mice (C+) displayed more freezing and travelled less distances compared to context controls (C-), irrespective of the genotype (Figure 11). As already observed in the analysis of the

Table 4: Statistical data of pairwise comparison of freezing levels between mice of the indicated genotypes in each session of the contextual extinction paradigm. Data were tested for normal distribution and equality of variances and a two-sided two-sample t-test or a two-sided Mann-Whitney-U-test was used, as indicated by the test statistic. For c*Bdnf* KO analysis: $N_{(WT)}$=9, $N_{(cBdnf\ KO)}$=9; for *Ntrk2*$^{+/-}$ analysis: $N_{(WT)}$=20, $N_{(Ntrk2^{+/-})}$=20.

	WT x c*Bdnf* KO		WT x *Ntrk2*$^{+/-}$	
session	test statistic	p-val	test statistic	p-val
Acq	T = 0.815	0.427	U = 86.0	0.006
Ret	T = -0.478	0.639	T = -2.218	0.033
Ext1	T = -0.629	0.538	T = -2.581	0.014
Ext2	T = -0.509	0.618	U = 107.5	0.013
Ext3	T = -0.197	0.847	T = -4.011	<0.001
Ext4	T = -0.953	0.355	T = -2.831	0.007
Ext5	T = -0.796	0.438	T = -1.478	0.148
Ext6	U = 40.0	1.0	T = -0.757	0.454

retrieval session in the contextual extinction learning paradigm (Figure 10), heterozygous *Ntrk2*knock-out mice showed comparably higher freezing rates (Figure 11A).

The combination of DL-enabled bioimage analyses of cFOS-positive nuclei and of Parv-positive somata allows the automatized quantification of seven different measures: the number of cFOS-positive nuclei within the NeuN-positive area (1), the mean cFOS-signal intensities of cFOS-positive nuclei within the NeuN-positive area (2), the number of Parv-positive somata (3), the mean Parv-signal intensities of all Parv-positive somata (4), the percentage of how many Parv-positive somata are cFOS-positive (5), the ratio of the mean Parv-signal intensities of cFOS-positive Parv-positive somata compared to cFOS-negative Parv-positive somata (6), and the mean cFOS-signal intensities of cFOS-positive nuclei within Parv-positive somata (7).

These analyses revealed, for instance, that there are no significant context-dependent differences, or differences based on the genotype for the number of detected Parv-positive somata (Figure S5, Figure S6, and Figure S7). Likewise, the mean Parv-signal intensities did not differ between the investigated genotypes or experimental conditions considering all Parv-positive somata (Figure S8, Figure S9, and S10). Calculating the ratio of Parv-signal intensities between cFOS-positive and cFOS-negative Parv-positive somata revealed no global difference between these two sub-populations (Figure S11, Figure S12, and Figure S13).

Neurons expressing cFOS as a marker for activity-related plasticity and memory pro-

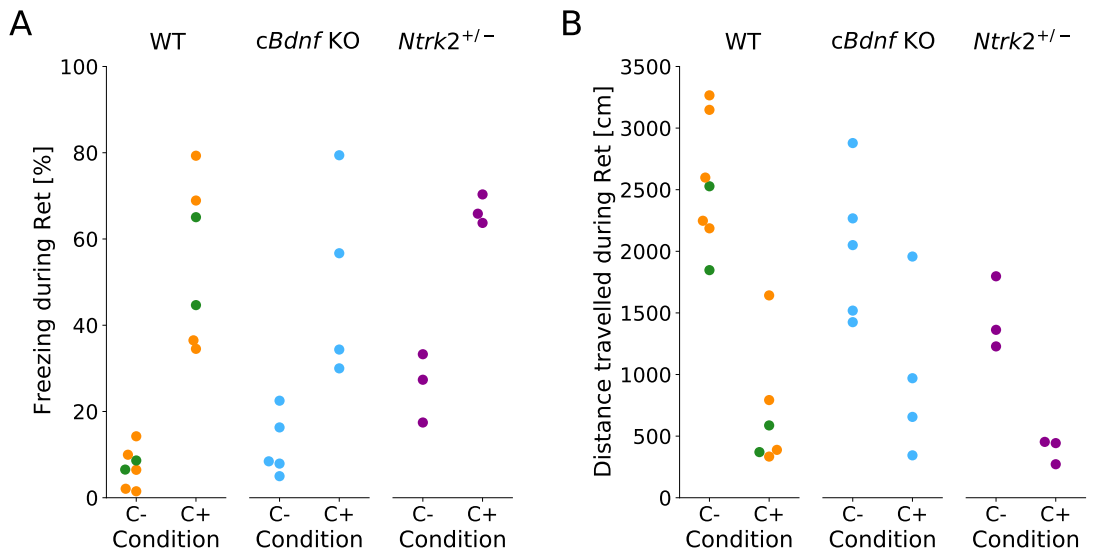


Figure 11: Behavioral analysis during context re-exposure of mice that were analyzed for cFOS signals 90 minutes later.
A. Freezing levels in percent of individual mice of the indicated genotypes and conditions. In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$). $N_{(\text{WT C- green/orange})}$=2/5, $N_{(\text{WT C+ green/orange})}$=2/4, $N_{(c\textit{Bdnf} \text{ KO C-})}$=5, $N_{(c\textit{Bdnf} \text{ KO C+})}$=4, $N_{(\textit{Ntrk2}^{+/-} \text{ C-})}$=3, $N_{(\textit{Ntrk2}^{+/-} \text{ C+})}$=3.
B. Travelled distance in cm of individual mice of the indicated genotypes and conditions. In WT animals, color coding discriminates between WT mice that were used as controls for c*Bdnf* KO mice (orange) and WT mice that served as controls for *Ntrk2*$^{+/-}$ mice (green). In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$). $N_{(\text{WT C-})}$=7, $N_{(\text{WT C+})}$=6, $N_{(c\textit{Bdnf} \text{ KO C-})}$=5, $N_{(c\textit{Bdnf} \text{ KO C+})}$=4, $N_{(\textit{Ntrk2}^{+/-} \text{ C-})}$=3, $N_{(\textit{Ntrk2}^{+/-} \text{ C+})}$=3.

cessing are typically categorized as cFOS-positive or cFOS-negative (Murawski et al., 2012; Tayler et al., 2013; Tonegawa et al., 2015; Josselyn et al., 2015; Holtmaat and Caroni, 2016; Keiser et al., 2017). However, cFOS abundance, as given by cFOS-signal intensities, also provide important information about memory traces (Ruediger et al., 2011; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

The quantifications of cFOS-positive nuclei in principal neurons of WT mice again revealed significant, context-dependent increases in the numbers of cFOS-positive nuclei in all analyzed subregions of the dorsal hippocampus (Figure S14A, Figure S15A, and Figure 12A), confirming the results of the previous analyses on partially overlapping data (Figure 8). However, a significant difference of the numbers of cFOS-positive nuclei in CA3 pyramidal neurons between C- and C+ WT mice could not be reproduced (Figure S15A, Figure 8B).

Context-dependent effects in the numbers of cFOS-positive nuclei could also be observed in the bioimage data of the two knock-out mouse models (Figure S14A, Figure S15A, and Figure 12A). Pooling the data of all conditions per genotype revealed two major differences in the amount of cFOS-positive nuclei within the NeuN-positive ar-

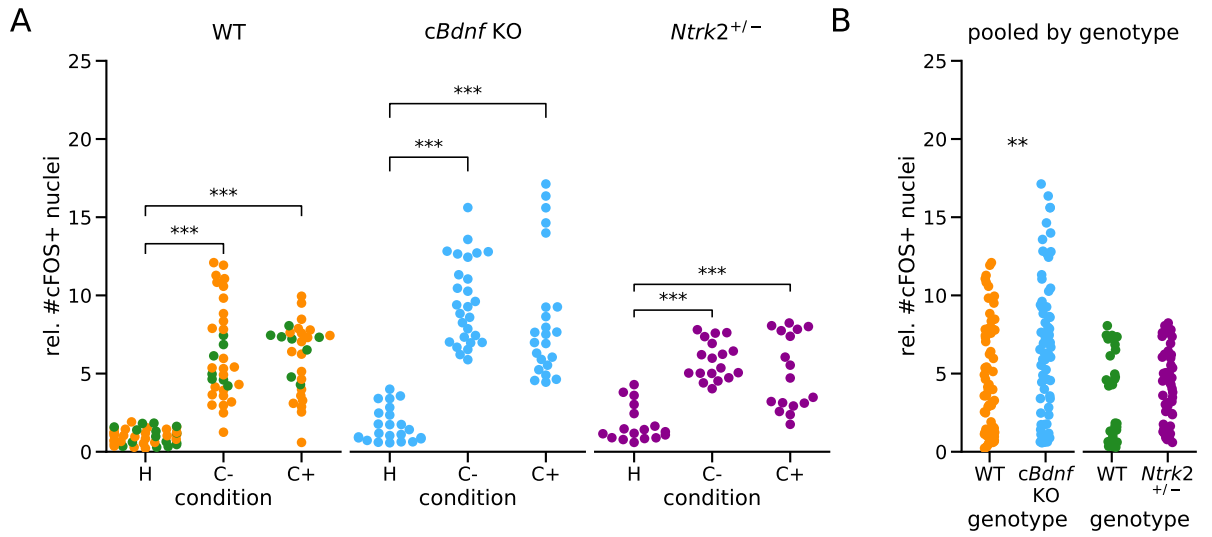


Figure 12: Quantification of cFOS-positive nuclei in CA1 in WT, *cBdnf* KO, and *Ntrk2*$^{+/-}$ mice. In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: *cBdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=70.46, P <0.001; *cBdnf* KO: H(2)=48.09, P <0.001; *Ntrk2*$^{+/-}$: H(2)=27.90, P <0.001; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (***: p <0.001).
B. Data of all conditions was pooled within the indicated genotypes. WT vs. *cBdnf* KO: U=1882.0, p = 0.001; WT vs. *Ntrk2*$^{+/-}$: U=633.0, p=0.158.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=43, $n_{(WT\ C-)}$=33, $n_{(WT\ C+)}$=29, $n_{(cBdnf\ KO\ H)}$=23, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=23, $n_{(Ntrk2^{+/-}\ H)}$=17, $n_{(Ntrk2^{+/-}\ C-)}$=18, $n_{(Ntrk2^{+/-}\ C+)}$=17.

eas between WT mice and the respective knock-out mouse model. In c*Bdnf* KO mice, significantly more cFOS-positive pyramidal neurons were detected in both CA3 and CA1 subregions of the dorsal hippocampus (Figure S15B and Figure 12B). These results are in line with the increased neuronal firing activity that was recorded *in vivo* in CA1 in this mouse model. Notably, the number of cFOS-positive neurons in the DG of c*Bdnf* KO mice was not different from those of wildtype controls (Figure S14B).

In heterozygous *Ntrk2*knock-out mice (*Ntrk2*$^{+/-}$), on the contrary, the levels of cFOS-positive pyramidal neurons in CA3 and CA1 were not different from those of WT mice (Figure S15B and Figure 12B). However, bioimage analyses revealed significantly less cFOS-positive nuclei in the granule cell layer of the DG in these mice (Figure S14). Moreover, only context conditioned, but not context control *Ntrk2*$^{+/-}$ mice showed significantly higher numbers of cFOS-positive nuclei in the DG compared to *Ntrk2*$^{+/-}$ home cage controls (Figure S14A).

The quantification of mean cFOS-signal intensities in WT mice in these bioimage analyses fully reproduced the findings described earlier on partially overlapping data (Figure S16A, Figure S17A, Figure S18A, and Figure 8). In both knock-out mouse models,

the context-dependent increase of mean cFOS-signal intensities in the dentate gyrus was abolished (Figure S16A). Likewise, in both knock-out mouse models, a context-dependent increase of the mean cFOS-signal intensities could be observerd, whereas WT mice showed significantly lower mean cFOS-signal intensities in C+ compared to C- mice (Figure S17A). Overall, no significant differences between the genotypes could be observed when the data from all conditions was pooled per genotype (Figure S16B, Figure S17B, and Figure S18B).

The use of a second *consensus ensemble* for the annotation of Parv-positive somata allowed to extend the bioimage analyses of cFOS-signals from the population of hippocampal principal neurons, as determined by the NeuN-positive area, to the population of Parv-positive interneurons, as determined by the annotations of Parv-positive somata. In WT mice, these analyses revealed a significantly higher proportion of cFOS-positive Parv-positive somata in CA3 in context control mice, compared to both, home-cage controls and context conditioned mice (Figure 13A). In c*Bdnf* KO mice, this significantly lower percentage of cFOS-positive Parv-positive somata in C+ compared to C- mice could not be observed. Instead, context conditioned mice showed a similarly increased ratio as well (Figure 13A). Moreover, the differences in the percentage of cFOS-positive Parv-positive somata between experimental conditions were completely abolished in *Ntrk2*$^{+/-}$ mice (Figure 13A). Interestingly, all mice of this genotype showed an overall elevated proportion of Parv-positive somata that were cFOS-positive, compared to WT mice (Figure 13B).

Overall, genotype-related differences in the ratio of cFOS-positive Parv-positive somata were limited to CA3. In CA1, the bioimage analyses of mice from all three genotypes revealed a significant difference of this measure only for the comparison between C- and H mice (Figure S19). In addition, Parv-positive somata in the DG were overall only rarely classified as cFOS-positive, which limits the interpretation of all measures that are derived from this classification. These are: the percentage of how many Parv-positive somata are cFOS-positive in the DG (Figure S20), the ratio of the mean Parv-signal intensities of cFOS-positive Parv-positive somata compared to cFOS-negative Parv-positive somata in the DG (Figure S11), and the mean cFOS-signal intensities of cFOS-positive nuclei within Parv-positive somata in the DG (Figure S21).

In WT mice, the mean cFOS-signal intensities of cFOS-positive nuclei within Parv-positive somata largely resembled the effects that could be observed in the ratio of

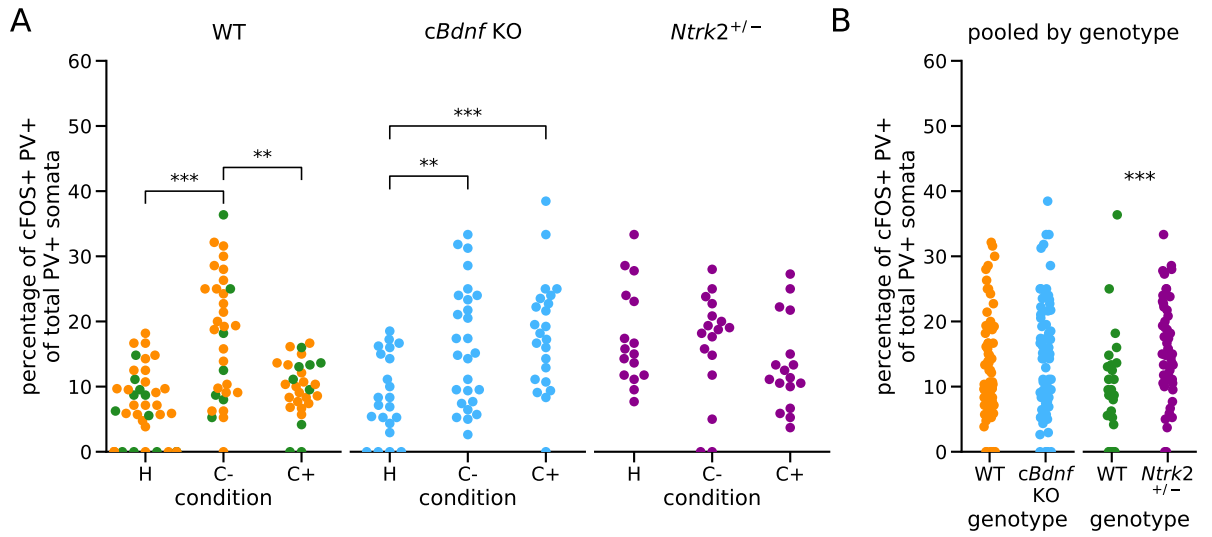


Figure 13: Percentage of cFOS-positive Parv-positive somata among all Parv-positive somata in CA3 in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice. In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=24.49, P <0.001; c*Bdnf* KO: F(2, 69)=8.86, P <0.001; *Ntrk2*$^{+/-}$: H(2)=4.04, P=0.132; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (**: p <0.01; ***: p <0.001).
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=2156.5, p = 0.082; WT vs. *Ntrk2*$^{+/-}$: U=355.5, p <0.001.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=39, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=28, $n_{(cBdnf\ KO\ H)}$=22, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=23, $n_{(Ntrk2^{+/-}\ H)}$=16, $n_{(Ntrk2^{+/-}\ C-)}$=17, $n_{(Ntrk2^{+/-}\ C+)}$=17.

cFOS-positive Parv-positive somata (Figure S22A, Figure S23A, Figure 13A, and Figure S19A). While the analyses of c*Bdnf* KOs did not show any significant differences in the mean cFOS-signal intensities of cFOS-positive nuclei within Parv-positive somata, the data of *Ntrk2*$^{+/-}$ mice revealed a context-dependent increase in CA1 (Figure S22, Figure S23).

To gain a more comprehensive overview of the detected differences between the experimental conditions and the respective genotypes on network-level, Figure 14 summarizes the results of all bioimage analyses of the two main measures, i.e. the number of cFOS-positive principal neurons and of the percentage of cFOS-positive Parv-positive interneurons, in schematic drawings of the hippocampus. The summary of all significant differences between the treatment groups of the respective genotypes highlights particularly the absence of any differences between C- and C+ mice in both genetically modified mouse models (Figure 14B). In WT mice, these groups could be differentiated by significantly lower percentages of cFOS-positive Parv-positive interneurons in CA3 of C+ mice (Figure 14B and Figure 13A). In addition, the comparison of WT mice with

c*Bdnf* KO mice revealed significantly higher amounts of cFOS-positive pyramidal neurons in both CA3 and CA1, whereas the amount of cFOS-positive DG granule cells was not affected (Figure 14C). The comparison between WT and heterozygous *Ntrk2*knockout mice showed less cFOS-positive DG granule cells and an elevated ratio of cFOS-positive Parv-positive interneurons in CA3 of *Ntrk2*$^{+/-}$ mice, irrespective of treatment condition (Figure 14D).

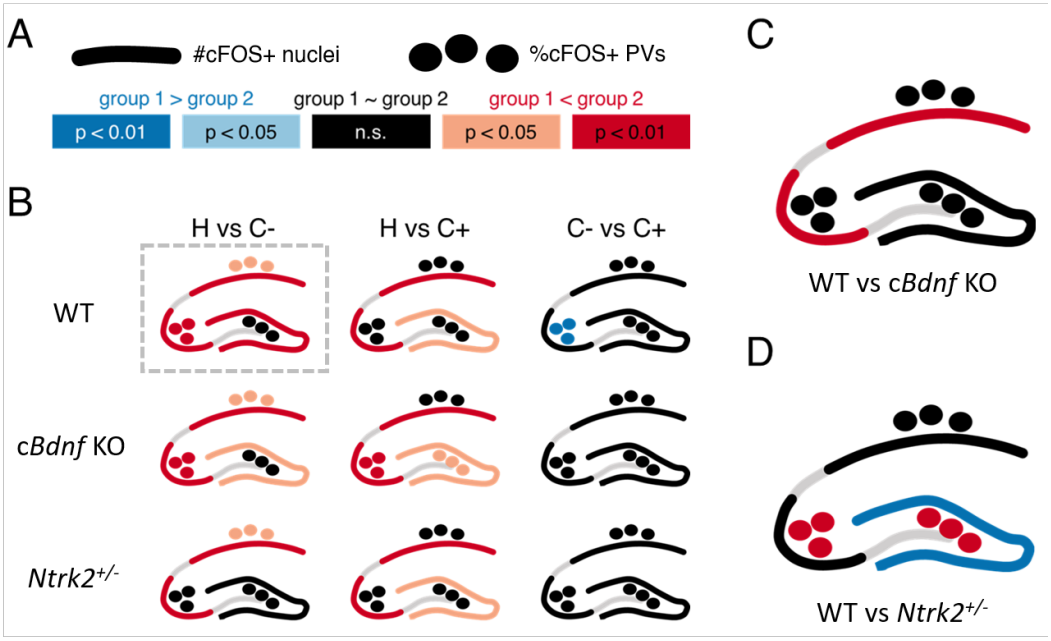


Figure 14: Network-level overview of bioimage analyses results of the two main measures in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice. Schematic drawings of the hippocampus summarize the results of the bioimage analyses of the two main measures: the number of cFOS-positive principal neurons (#cFOS+ nuclei) and the percentage of cFOS-positive Parv-positive interneurons (%cFOS+ PVs).
A. In these drawings, the population of principal neurons in the granule cell layer of the DG and in the pyramidal cell layers of CA3 and CA1 are visualized as thick lines and represent the quantifications of cFOS-positive principal neurons in the respective subregion. In addition, three oval shapes visualize the population of Parv-positive interneurons in each hippocampal subregion and represent the bioimage analysis results of the percentage of cFOS-positive Parv-positive interneurons in the respective subregion. A color coding is used to indicate the directionality and the significance of pairwise comparisons of the respective measure between indicated groups. Blue colors represent a significantly lower mean of group 2, whereas red colors indicate a significantly higher mean of group 2. The level of significance is indicated by color intensity. If the pairwise comparison did not reveal a significant difference of the respective measure between the two indicated groups, the corresponding feature is depicted in black. Grey features in the schamtic drawings represent areas of the hippocampus that were not investigated. Detailed plotting of the respective data can be found in Figures S14, S15, 12, S20, 13, and S19.
B. Pairwise comparisons of the indicated measures between two experimental conditions of mice with the indicated genotype (eg. 'WT H' as group 1 vs 'WT C-' as group 2 - grey box).
C. For these pairwise comparisons of the indicated measures between WT and c*Bdnf* KO mice, the data of all conditions was pooled per genotype.
D. For these pairwise comparisons of the indicated measures between WT and *Ntrk2*$^{+/-}$ mice, the data of all conditions was pooled per genotype.

5 Discussion

This thesis elaborates the limitations and the potentials of DL-based bioimage analyses in terms of objectivity, reliability, and validity. For the evaluation of different DL-based strategies, classical similarity measures and the comparison of bioimage analysis results were used. This allowed to investigate not only the homogeneity of image feature annotations, but, more importantly, also of the biologically relevant statistical evaluation of the image dataset. Quantitative analyses revealed significant differences between the investigated DL-based strategies and provide an empirical basis that suggests, how DL can improve the quality of bioimage data analyses.

5.1 Challenges for the analysis of fluorescence microscopy images

Recent advances in fluorescence labeling strategies and in image acquisition techniques enable life scientists to gain increasing amounts of insights into biological systems from image data (A. Li et al., 2010; Osten and Margrie, 2013; Boutros et al., 2015; McDole et al., 2018; Groot et al., 2020). This necessitates the interpretation and quantitative analysis of image features of interest throughout the entire image dataset in order to test a hypothesis that underlies the respective experiment. This process is known as bioimage analysis (Meijering et al., 2016). And just as any other empirical analysis, also bioimage analyses must adhere to the common standards of quantitative research to ensure the highest research quality possible. These standards are objectivity, reliability, and validity (Frambach et al., 2013).

In the particular case of fluorescence microscopy image data, however, it is virtually impossible to define a *de facto* ground truth, meaning a reference annotation that is fully objective. Due to the physics of light, fluorescence signals do not have clear boundaries. On an even larger scale, noise can be introduced to the signal during image acquisition, for instance due to light scattering of the tissue. Nonetheless, the subsequent interpretation of relevant image features requires the definition of accurate signal-to-noise borders and these annotations can be made either manually by a human expert, or by using a computer-aided approach.

Manual image analysis by human experts is still frequently considered as the gold standard for bioimage analyses, like in histopathological analyses (Aeffner et al., 2017).

However, manual analysis is a time consuming cognitive process that is both, influenced and limited by the individual graphical perception capabilities and susceptible to visual and cognitive traps (Cleveland and McGill, 1985; Aeffner et al., 2017). As a consequence, the level of subjectivity of manual annotations increases with decreasing signal-to-noise ratios (Niedworok et al., 2016; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a). Moreover, even if the same expert repeats the annotation process, the corresponding intra-rater agreement, and thus the reliability, is limited (Collier et al., 2003).

A computer-aided approach, on the contrary, will always reproduce the identical annotation as long as its parameters are not changed. In addition, the automation of bioimage data analysis using computer-aided approaches can drastically reduce the workload of the human experimenter and is, thus, highly desirable. And yet, it remained particularly challenging to design transformations of the data that enable the computer-aided approach to extract the desired image features of interest (Meijering et al., 2016; LeCun et al., 2015). Thus, the use of conventional hard-coded approaches can require substantial computational expertise (Chamier et al., 2019). Interestingly, also these approaches are discussed as being inherently subjective, since the parameters of the algorithm are chosen by a human as user (Tadrous, 2010).

In recent years, deep learning algorithms have proven their remarkable capacities in image analysis tasks and are becoming increasingly popular, also in the life sciences (Moen et al., 2019; Chamier et al., 2019). The underlying algorithmic architecture is somewhat inspired from the organization and the computations of biological neuronal networks and enables the DL model to learn a specific task solely on base of a training dataset (LeCun et al., 2015; Moen et al., 2019; Chamier et al., 2019).

Importantly, this process does no longer require the user to develop appropriate feature extractors, addressing a central challenge of computer vision (Meijering et al., 2016). Instead, training facilitates that the DL model learns which of its data representations can serve as feature extractor and causes the model to converge as close as possible to the presented annotations (LeCun et al., 2015; Moen et al., 2019). However, this causes a critical problem if there is no ground truth data available, like in case of fluorescence images (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a). For instance, training on inconsistent or subjective annotations could impair the training or lead to the incorporation of a subjective bias into the trained model (Falk et al., 2019; Chamier et

al., 2019; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a). Moreover, model training is a stochastic process and even models that are trained on the same data can have divergent outputs (Dietterich, 2000). Ultimately, this could even cause divergent bioimage analysis results and would therefore represent a major limitation for the use of DL in bioimage analyses. Since recent efforts make DL-based image analysis tools increasingly accessible also to non-AI experts (Haberl et al., 2018; Falk et al., 2019), the need for a systematic evaluation of the impact of subjective manual annotations and of model-to-model variability on the results of DL-based bioimage analyses increases (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

5.2 Impact of DL-based strategies on fluorescent feature annotations

To address this need, this study used bioimage datasets showing cFOS signals in the hippocampus after behavioral testing of mice. Manual annotations of five P.hD.-level neurobiologists were used to train DL-models either on the annotations of individual experts (*expert models*), or on estimated ground truth annotations derived from the annotations of all five experts (*consensus models*). In a third strategy, the output of multiple *consensus models* were combined to form *consensus ensembles* (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

As expected, similarity analyses of the manual expert annotations revealed significant inter-rater variability, which was inversely correlated with the signal-to-noise ratios of the annotated features (Schmitz et al., 1999; Collier et al., 2003; Niedworok et al., 2016; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a). Yet, the training efficiency of DL models was not affected by the different annotators (experts 1-5 or est. GT), since all individual *expert models* and the *consensus models* reached similar F1-scores when compared to the annotator they were trained on (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

However, by comparing all annotations to the est. GT as a common reference, or by assessing all possible pairwise combinations, individual differences became apparent. Most strikingly, these analyses revealed pronounced deviations of the annotations of expert 1, which were closely mimicked by all four *expert 1 models*. Thus, these data provide experimental evidence for the incorporation of a subjective bias from manual annotations into a deep learning model. Moreover, despite the annotations of expert

1 were also included into the ground truth estimation process, the resulting *consensus models* did not show similar deviations. Instead, *consensus models* performed best among individual models when compared to est. GT annotations on new data. This indicates, that pooling the annotations from multiple human experts into a single training dataset by means of ground truth estimation was an effective strategy to reduce the risk of incorporation of a subjective bias into the trained models. Furthermore, these analyses also highlight the importance of using the annotations from multiple experts for the evaluation, since such effects could otherwise go unnoticed (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

Similarity analyses of the annotations within each strategy showed that *expert models* and *consensus models* had a similar annotation reliability. Notably, the annotation reliability was drastically higher in *consensus ensembles*. Therefore, these data also confirm that ensemble formation was an effective strategy in order to reduce model-to-model variability and to increase the reliability of DL-based image feature annotations (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

5.3 Impact of DL-based strategies on the results of bioimage analyses

The comparison of the three DL-based strategies was then extended to an bioimage dataset with a total of 283 images to investigate, whether subjective manual annotations and model-to-model variability can also impact the results of bioimage analyses. Notably, the majority votes for each pairwise comparison (significantly different or not at $p<0.05$) was identical across the three strategies and generally in line with the neuroscientific literature (Campeau et al., 1997; Guzowski et al., 2001; Huff et al., 2006; Ramamoorthi et al., 2011; Murawski et al., 2012; Tayler et al., 2013; Keiser et al., 2017). However, the data of individual *expert models* revealed that none of their bioimage analyses results was in full accordance with the congruent majority votes. This variance was only modestly reduced in *consensus models*. In line with the increased annotation reliability of *consensus ensembles*, the bioimage analysis results derived from their output were most often in full accordance with the congruent majority votes (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

In conclusion, the choice of how DL is trained and used to perform bioimage analyses can have a strong impact on the objectivity, reliability, and validity of the results. The

present study suggests to integrate the annotations of several human experts into the training dataset, e.g. by means of ground truth estimation. Furthermore, evaluation of the trained models should also include the annotations of multiple experts. Moreover, the formation of model ensembles proofed to be an effective and yet easily implementable feature to increase annotation reliability (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

In addition, the evaluation of the DL-based strategies on four additional bioimage datasets that were independently acquired in four different laboratories, validated the results presented in this thesis (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a). These datasets comprise various image acquisition parameters and techniques, analyze the two main cellular compartments (nuclei and somata), and are derived from two commonly used model organisms, i.e. mice and zebrafish (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a). These analyses indicate that the conclusions and recommendations derived from the extensive comparison of the DL-based strategies on the initial bioimage dataset, which is presented in this thesis, can be generalized to similar bioimage datasets and that the use of *consensus ensembles* can improve the quality of bioimage analyses results.

5.4 Extended validation of *consensus ensembles* on the bioimage analyses of two genetically modified mouse models

Moreover, this study used a combination of two *consensus ensembles* for the annotation of different fluorescent features in a large collection of bioimages. This offered insights into the performance of the *consensus ensemble* strategy under complex and variable experimental conditions. For this, cFOS fluorescence signals were analyzed in two subpopulations of hippocampal neurons (principal neurons and Parvalbumin-positive interneurons) in two genetically modified mouse models that cause an altered BDNF-TrkB signaling in the hippocampus.

5.4.1 BDNF-TrkB signaling

Brain-derived neurotrophic factor (BDNF) belongs to the family of neurotrophins, which is comprised of small secretory proteins that are predominantly involved in the regulation of neuronal cell survival, cell death, and differentiation (Y. A. Barde et al., 1982;

Yves Alain Barde, 1989; Chao, 2003; Sasi et al., 2017). BDNF was discovered as the predominant neurotrophin in the adult mammalian brain (Y. A. Barde et al., 1982; Yves Alain Barde, 1989; Thoenen, 1995). Like the other neurotrophins, BDNF promotes neuronal growth and survival, particularly in the peripheral nervous system (Y. A. Barde et al., 1982; Yves Alain Barde, 1989; Sendtner et al., 1992; Erickson et al., 1996), but to some extent also in the central nervous system (CNS; (Sairanen et al., 2005; Bergami et al., 2008). However, the main function of BDNF in the CNS is the regulation of neuronal circuit functions, like neuronal differentiation and synaptic function (Sasi et al., 2017). As a consequence, the lack of BDNF in the CNS results only in a modest loss of neurons and causes region-specific dysregulations of neuronal circuits instead (Nikoletopoulou et al., 2010; Rauskolb et al., 2010; Y. Li et al., 2012; Park and Poo, 2013; Sasi et al., 2017). In the hippocampus, the deletion of BDNF leads for instance to an impaired long-term potentiation, one of the key molecular mechanisms of learning and memory, in both CA1 and CA3 (Korte et al., 1995; Blum and Konnerth, 2005; Schildt et al., 2013).

BDNF signalling is mediated via two types of receptors, the high-affinity tropomyosin-receptor-kinase B (TrkB, Rodriguez-Tebar and Y. A. Barde, 1988; Barbacid, 1994; Klein et al., 1989; Martin-Zanca et al., 1986) and the low-affinity p75 pan-neurotrophin receptor (Chao, 2003; Blum and Konnerth, 2005). Upon binding of a BDNF homodimer to its physiological receptor TrkB, the receptor dimerizes, which triggers an autophosphorylation cascade and consequently initiates downstream signalling pathways that mediate the physiological effects of BDNF (Blum and Konnerth, 2005; Park and Poo, 2013; Sasi et al., 2017).

BDNF and TrkB are both abundant in the hippocampus (Conner et al., 1997; Minichiello et al., 1999). One source of BDNF in the hippocampus are granule cells of the dentate gyrus, whose axons, the the so-called mossy fibers, project to CA3 (Conner et al., 1997; Deng et al., 2010; Wiera and Mozrzymas, 2015). In CA3, they form several large mossy-fiber terminals (LMTs) that represent huge bouton-like structures, which comprise multiple active-zones and engulf the dendritic structures of CA3 pyramidal neurons. In addition, there are filopdia protruding from LMTs. These filopodia are, in turn, able to form synapses with CA3 pyramidal neurons, but also with inhibitory interneurons, like Parvalbumin-positive interneurons (Wiera and Mozrzymas, 2015; Martin et al., 2017). Parv-positive interneurons and CA3 pyramidal neurons both express the TrkB receptor and BDNF was detected in LMTs and the protruding filopodia (Danzer and McNAmara, 2004; Huang et al., 2008; K. Zheng et al., 2011; Schildt et al., 2013; Sasi et al., 2017).

5.4.2 Analyses of *cBdnf* KO mice

Previous work by Dr. Manju Sasi, Dr. Cora Rüdt von Collenberg, Dr. Britta Wachter, Dr. Thomas Seidenbecher, and Dr. Robert Blum revealed that the deletion of *Bdnf* from a sparse population of DG granule cells does not cause apparent behavioral deficits in these mice, but results in an increased neuronal activity in CA1 during fear extinction learning. Using a *consensus ensemble* to analyze cFOS labelings in these mice revealed elevated numbers of cFOS-positive nuclei in CA1 and, thus, confirmed these results. Moreover, this effect was also observed in CA3, while activity-dependent cFOS labels in the DG were not altered. Consequently, these findings support the role of BDNF as an anterograde signalling molecule in the hippocampus (Conner et al., 1997; Dieni et al., 2012; Andreska et al., 2014).

Since BDNF signaling acts excitatory (Sasi et al., 2017), these increases in neuronal activity of principal neurons upon conditional loss of BDNF could be explained best by a loss of excitation of inhibitory neurons in *cBdnf* KO mice. For this, the analysis of Parv-positive interneurons was included, since these inhibitory interneurons express TrkB and have already been shown to be important for the precision of contextual memory encoding (Ruediger et al., 2011; K. Zheng et al., 2011).

Interestingly, bioimage analyses of WT mice showed significantly more cFOS-positive Parv-positive interneurons in CA3 only of context-control mice, but not of context-conditioned mice, when compared to homecage controls. In fact, the pairwise comparison of C- and C+ wildtype mice revealed a significantly smaller proportion of cFOS-positive Parv-positive interneurons in C+ mice. Overall, the levels of cFOS-positive Parv-positive somata were not reduced in *cBdnf* KO mice. Instead, the significant difference of cFOS-positive Parv-positive interneurons between C+ and C- in CA3 was even absent in these mice and levels of cFOS-positive Parv-positive interneurons were elevated also in C+ *cBdnf* KO mice. However, Parvalbumin-positive interneurons represent only a sub-population of all inhibitory neurons in the hippocampus and the lack of BDNF in these mice could also affect another class of inhibitory neurons. In addition to deficits in acute molecular signalling, the conditional deletion of BDNF could also induce morphological changes in the network, for instance alterations in the connectivity of DG granule neurons. Ongoing work using viral tracing tools will address this question and compare the morphology of LMT filopodia after retrieval of a contextual memory between WT and *cBdnf* KO mice.

5.4.3 Analyses of *Ntrk2*$^{+/-}$ mice

In a second mouse model, one copy of the TrkB-encoding gene *Ntrk2* was deleted (Rohrer et al., 1999). In contrast to the conditional knock-out of *Bdnf*, *Ntrk2*$^{+/-}$ mice did not show significantly elevated numbers of cFOS-positive principal neurons. Instead, activity-dependent cFOS labels were significantly less in the DG of *Ntrk2*$^{+/-}$ mice compared to wildtype controls. However, somewhat similar to the *cBdnf* KO mice, *Ntrk2*$^{+/-}$ mice displayed no significant reduction of the proportion of cFOS-positive Parv-positive interneurons in CA3 between C- and C+ mice. Instead, also homecage control mice of this genotype showed activation levels of CA3 Parv-positive interneurons that were comparable to the significantly elevated levels that were found in WT C-, *cBdnf* KO C-, or *cBdnf* KO C+ mice. The small groups of mice that were used for the investigation of cFOS (N=3 per experimental condition) demands caution in terms of reliability and reproducibility of the results and make additional experiments necessary. However, behavioral analysis of a bigger cohort of these mice in a contextual fear extinction paradigm revealed significantly higher freezing levels of *Ntrk2*$^{+/-}$ mice and thus strengthens the hypothesis, that BDNF-TrkB signalling constitutes a role in the processes of fear and anxiety.

In summary, the combined use of two *consensus ensembles* for the annotation of cFOS-positive nuclei and of Parv-positive somata confirmed their reliable and consistent performance also in this large bioimage dataset. Moreover, the correlation of significantly increased numbers of cFOS-positive nuclei with the increased neuronal firing activity in CA1 of *cBdnf* KO, which was revealed with *in vivo* electrophysiological recordings, provides a second line of evidence for the validity of these DL-based bioimage analyses.

5.5 Conclusion and outlook

Implemented in the right way, deep learning has the potential to improve objectivity, reliability, and validity of bioimage analyses, going beyond the mere automation of the image annotation process (Figure 15; Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a). Current work in progress will provide a user-friendly, open access toolbox that comprises all core elements of the suggested *consensus ensemble* workflow, like ground truth estimation or ensemble formation (`www.github.com/matjesg/deepflash2`). Moreover, the explorative analyses of four bioimage datasets provided by independent laboratories also confirm the advantages of transfer learning on these data (Segebarth,

Griebel, Stein, R von Collenberg, et al., 2020a; Falk et al., 2019). Transfer learning describes a concept, in which a model that was already trained on a specific task is used as starting point for the training on a similar task (Moen et al., 2019). In fact, using a model with pre-trained weights can significantly reduce the computational effort during training, compared to starting with randomly initialized weights (Falk et al., 2019). Importantly, bioimage analyses are often focused on highly similar image features, like the analysis of nuclear labels. Consequently, an open access library that contains validated *consensus ensembles* that were pre-trained on a specific cellular feature, would enable other researchers with similar image data to re-use them. Ultimately, this would foster the use of DL by reducing the efforts required for model training. Furthermore, sharing of pre-trained model ensembles may represent a form of sharing annotation expertise, such that it becomes freely available and accessible for the entire life science community (Segebarth, Griebel, Stein, R von Collenberg, et al., 2020a).

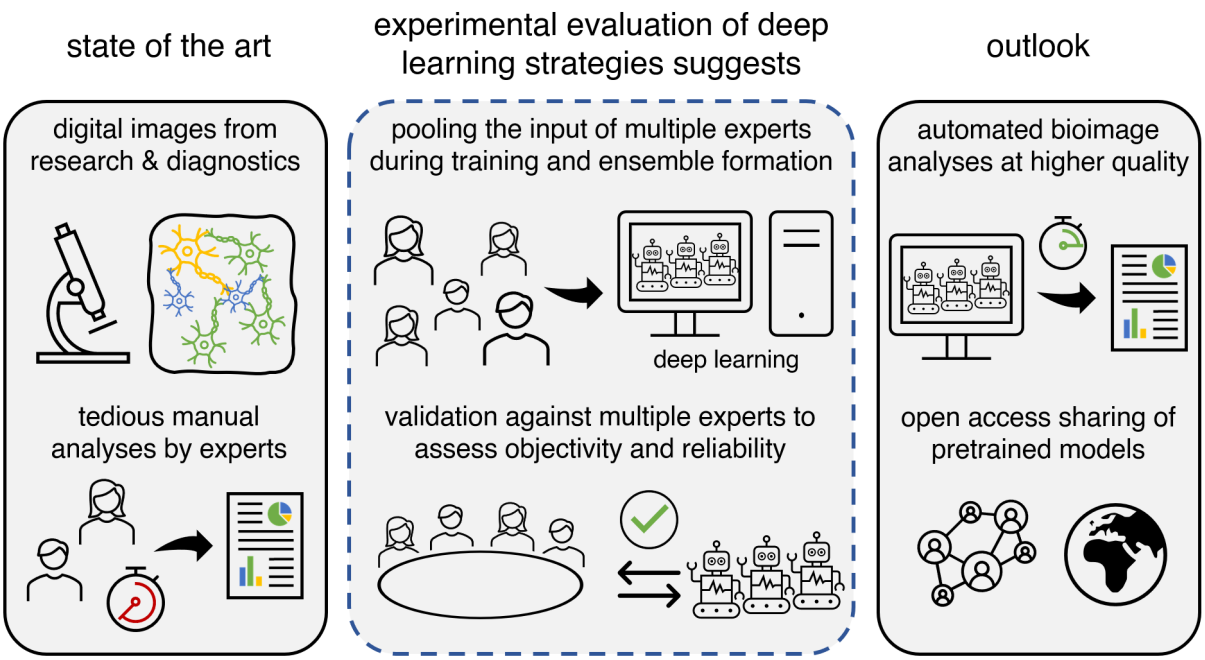


Figure 15: Schematic conclusion. Today, tedious manual analysis of bioimages by human experts frequently resembles the current state-of-the-art of bioimage analyses in research and diagnostics. This thesis hypothesized, that DL could be used to automatize this process with an concomitant increase of objectivity, reliability, and validity. Experimental evaluation of three DL-based strategies suggests to integrate the pooling of the input of multiple experts for the training (e.g. via ground truth estimation) and the formation of model ensembles in any DL-based approach for bioimage analyses to establish objectivity and reliability. Moreover, model ensemble performance should be validated on base of the annotations of multiple experts. Implemented in such a way, DL has the potential to increase the quality of bioimage analyses, going beyond mere automation. Researchers can share their annotation expertise with the entire life science community via sharing of validated pretrained models and model ensembles in open access libraries. Here, the pictogram of a robot symbolizes a deep learning model and a group of robots symbolizes a model ensemble.

6 References

Aeffner, Famke, Kristin Wilson, Nathan T. Martin, Joshua C. Black, Cris L. Luengo Hendriks, Brad Bolon, Daniel G. Rudmann, Roberto Gianani, Sally R. Koegler, Joseph Krueger, and G. Dave Young (Sept. 2017). "The Gold Standard Paradox in Digital Image Analysis: Manual Versus Automated Scoring as Ground Truth". In: *Archives of Pathology & Laboratory Medicine* 141.9, pp. 1267–1275. ISSN: 0003-9985. DOI: 10.5858/arpa.2016-0386-RA.

Andreska, Thomas, Sarah Aufmkolk, Markus Sauer, and Robert Blum (2014). "High abundance of BDNF within glutamatergic presynapses of cultured hippocampal neurons". In: *Frontiers in cellular neuroscience* 8, p. 107.

Barbacid, Mariano (Nov. 1994). "The Trk family of neurotrophin receptors". In: *Journal of Neurobiology* 25.11, pp. 1386–1403. ISSN: 0022-3034. DOI: 10.1002/neu.480251107.

Barde, Y. A., D. Edgar, and H. Thoenen (1982). "Purification of a new neurotrophic factor from mammalian brain." In: *The EMBO journal* 1.5, pp. 549–553. ISSN: 02614189. DOI: 10.1002/j.1460-2075.1982.tb01207.x.

Barde, Yves Alain (1989). "Trophic factors and neuronal survival". In: *Neuron* 2, pp. 1525–1534. ISSN: 08966273. DOI: 10.1016/0896-6273(89)90040-8.

Bergami, Matteo, Roberto Rimondini, Spartaco Santi, Robert Blum, Magdalena Götz, and Marco Canossa (2008). "Deletion of TrkB in adult progenitors alters newborn neuron integration into hippocampal circuits and increases anxiety-like behavior". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.40, pp. 15570–15575. ISSN: 00278424. DOI: 10.1073/pnas.0803702105.

Blum, Robert and Arthur Konnerth (2005). "Neurotrophin-Mediated Rapid Signaling in the Central Nervous System: Mechanisms and Functions". In: *Physiology* 20, pp. 70–78.

Boutros, Michael, Florian Heigwer, and Christina Laufer (2015). "Microscopy-based high-content screening". In: *Cell* 163.6, pp. 1314–1325.

Buggenthin, Felix, Florian Buettner, Philipp S. Hoppe, Max Endele, Manuel Kroiss, Michael Strasser, Michael Schwarzfischer, Dirk Loeffler, Konstantinos D. Kokkaliaris, Oliver Hilsenbeck, Timm Schroeder, Fabian J. Theis, and Carsten Marr (2017). "Prospective identification of hematopoietic lineage choice by deep learning". In: *Nature Methods* 14.4, pp. 403–406. ISSN: 15487105. DOI: 10.1038/nmeth.4182.

Caicedo, Juan C., Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, et al. (2017). "Data-analysis strategies for image-based cell profiling". In: *Nature methods* 14.9, pp. 849–863.

Caicedo, Juan C., Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Jeanelle Ackerman, Marzieh Haghighi, Cher Keng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Rohban, Shantanu Singh, and Anne E. Carpenter (2019). "Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl". In: *Nature Methods* 16.12, pp. 1247–1253. ISSN: 15487105. DOI: 10.1038/s41592-019-0612-7.

Caicedo, Juan C., Jonathan Roth, Allen Goodman, Tim Becker, Kyle W Karhohs, Matthieu Broisin, Csaba Molnar, Claire McQuin, Shantanu Singh, Fabian J. Theis, and Anne E. Carpenter (Sept. 2019). "Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images". In: *Cytometry Part A* 95.9, pp. 952–965. ISSN: 1552-4922. DOI: 10.1002/cyto.a.23863.

Campeau, S, WA Falls, WE Cullinan, DL Helmreich, M Davis, and SJ Watson (1997). "Elicitation and reduction of fear: behavioural and neuroendocrine indices and brain induction of the immediate-early gene c-fos". In: *Neuroscience* 78.4, pp. 1087–1104.

Chamier, Lucas von, Romain F Laine, and Ricardo Henriques (2019). "Artificial intelligence for microscopy: what you should know". In: *Biochemical Society Transactions* 47.4, pp. 1029–1040.

Chao, Moses (2003). "Neurotrophins and their receptors: a convergence point for many signalling pathways". In: *Nature Reviews Neurology* 4, pp. 299–309.

Chen, Z.-Y., D. Jing, K. G. Bath, A. Ieraci, T. Khan, C.-J. Siao, D. G. Herrera, M. Toth, C. Yang, B. S. McEwen, B. L. Hempstead, and F. S. Lee (2006). "Genetic Variant BDNF (Val66Met) Polymorphism Alters Anxiety-Related Behavior". In: *Science* 314, pp. 140–143.

Christiansen, Eric M, Samuel J Yang, D Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison O'Neil, Kevan Shah, Alicia K Lee, et al. (2018). "In silico labeling: predicting fluorescent labels in unlabeled images". In: *Cell* 173.3, pp. 792–803.

Cleveland, William S and Robert McGill (1985). "Graphical perception and graphical methods for analyzing scientific data". In: *Science* 229.4716, pp. 828–833.

Collier, Dawn C., Stuart S.C. Burnett, Mayankkumar Amin, Stephen Bilton, Christopher Brooks, Amanda Ryan, Dominique Roniger, Danny Tran, and George Starkschall

(2003). "Assessment of consistency in contouring of normal-tissue anatomic structures." In: *Journal of applied clinical medical physics / American College of Medical Physics* 4.1, pp. 17–24. ISSN: 15269914. DOI: 10.1120/jacmp.v4i1.2538.

Conner, James M, Julie C Lauterborn, Qiao Yan, Christine M Gall, and Silvio Varon (1997). "Distribution of Brain-Derived Neurotrophic Factor (BDNF) Protein and mRNA in the Normal Adult Rat CNS : Evidence for Anterograde Axonal Transport". In: *The Journal of Neuroscience* 17.7, pp. 2295–2313.

Danuser, Gaudenz (2011). "Computer vision in cell biology". In: *Cell* 147.5, pp. 973–978. ISSN: 10974172. DOI: 10.1016/j.cell.2011.11.001.

Danzer, Steve C. and James O. McNAmara (2004). "Localization of Brain-Derived Neurotrophic Factor to Distinct Terminals of Mossy Fiber Axons Implies Regulation of Both Excitation and Feedforward Inhibition of CA3 Pyramidal Cells". In: *Journal of Neuroscience* 24.50, pp. 11346–11355. ISSN: 0270-6474.

Deng, Wei, James B. Aimone, and Fred H. Gage (2010). "New neurons and new memories: how does adult hippocampal neurogenesis affect learning and memory?" In: *Nature Reviews Neuroscience* 11.5, pp. 339–350.

Dieni, Sandra, Tomoya Matsumoto, Martijn Dekkers, Stefanie Rauskolb, Mihai S Ionescu, Ruben Deogracias, Eckart D Gundelfinger, Masami Kojima, Sigrun Nestel, Michael Frotscher, and Yves-Alain Barde (Mar. 2012). "BDNF and its pro-peptide are stored in presynaptic dense core vesicles in brain neurons". In: *Journal of Cell Biology* 196.6, pp. 775–788. ISSN: 1540-8140. DOI: 10.1083/jcb.201201038.

Dietterich, Thomas G (2000). "Ensemble methods in machine learning". In: *International workshop on multiple classifier systems*. Springer, pp. 1–15.

Erickson, Jeffery T., Joanne C. Conover, Veronique Borday, Jean Champagnat, Mariano Barbacid, George Yancopoulos, and David M. Katz (1996). "Mice lacking brain-derived neurotrophic factor exhibit visceral sensory neuron losses distinct from mice lacking NT4 and display a severe developmental deficit in control of breathing". In: *Journal of Neuroscience* 16.17, pp. 5361–5371. ISSN: 02706474. DOI: 10.1523/jneurosci.16-17-05361.1996.

Falk, Thorsten et al. (2019). "U-Net: deep learning for cell counting, detection, and morphometry". In: *Nature Methods* 16.1, pp. 67–70. ISSN: 15487105. DOI: 10.1038/s41592-018-0261-2.

Fanselow, M. S. (1980). "Conditioned and unconditional components of post-shock freezing". In: *Pavlov J Biol Sci* 15.4, pp. 177–82. ISSN: 0093-2213 (Print) 0093-2213 (Linking).

Fleiss, Joseph L and Jacob Cohen (1973). "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability". In: *Educational and psychological measurement* 33.3, pp. 613–619.

Frambach, Janneke M, Cees PM van der Vleuten, and Steven J Durning (2013). "AM last page: Quality criteria in qualitative and quantitative research". In: *Academic Medicine* 88.4, p. 552.

Gallo, Francisco T, Cynthia Katche, Juan F Morici, Jorge H Medina, and Noelia V Weisstaub (2018). "Immediate early genes, memory and psychiatric disorders: focus on c-Fos, Egr1 and Arc". In: *Frontiers in behavioral neuroscience* 12, p. 79.

Gonzalez, Walter G, Hanwen Zhang, Anna Harutyunyan, and Carlos Lois (2019). "Persistence of neuronal representations through time and damage in the hippocampus". In: *Science* 365.6455, pp. 821–825.

Greenberg, Michael E. and Edward B. Ziff (1984). "Stimulation of 3T3 cells induces transcription of the c-fos proto-oncogene". In: *Nature* 311.5985, pp. 433–438. ISSN: 00280836. DOI: 10.1038/311433a0.

Groot, Andres de, Bastijn J.G. van den Boom, Romano M. van Genderen, Joris Coppens, John van Veldhuijzen, Joop Bos, Hugo Hoedemaker, Mario Negrello, Ingo Willuhn, Chris I. De Zeeuw, and Tycho M. Hoogland (2020). "Ninscope, a versatile miniscope for multi-region circuit investigations". In: *eLife* 9, pp. 1–24. ISSN: 2050084X. DOI: 10.7554/eLife.49987.

Guzowski, John F, Barry Setlow, Edward K Wagner, and James L McGaugh (2001). "Experience-dependent gene expression in the rat hippocampus after spatial learning: a comparison of the immediate-early genesArc, c-fos, and zif268". In: *Journal of Neuroscience* 21.14, pp. 5089–5098.

Haberl, Matthias G., Christopher Churas, Lucas Tindall, Daniela Boassa, Sébastien Phan, Eric A. Bushong, Matthew Madany, Raffi Akay, Thomas J. Deerinck, Steven T. Peltier, and Mark H. Ellisman (2018). "CDeep3M—Plug-and-Play cloud-based deep learning for image segmentation". In: *Nature Methods* 15.9, pp. 677–680. ISSN: 15487105. DOI: 10.1038/s41592-018-0106-z.

Holtmaat, Anthony and Pico Caroni (2016). "Functional and structural underpinnings of neuronal assembly formation in learning". In: *Nature Neuroscience* 19.12, pp. 1553–1562.

Hu, Hua, Jian Gan, and Peter Jonas (2014). "Fast-spiking, Parvalbumin+ GABAergic interneurons: From cellular design to microcircuit function". In: *Science* 345.6195.

Huang, Yang Z, Enhui Pan, ZhiQi Xiong, and James O. McNamara (Feb. 2008). "Zinc-Mediated Transactivation of TrkB Potentiates the Hippocampal Mossy Fiber-CA3 Pyramid Synapse". In: *Neuron* 57.4, pp. 546–558. ISSN: 08966273. DOI: 10.1016/j.neuron.2007.11.026.

Huff, Nicole C, Matthew Frank, Karli Wright-Hardesty, David Sprunger, Patricia Matus-Amat, Emily Higgins, and Jerry W Rudy (2006). "Amygdala regulation of immediate-early gene expression in the hippocampus induced by contextual fear conditioning". In: *Journal of Neuroscience* 26.5, pp. 1616–1623.

Josselyn, Sheena A., Stefan Köhler, and Paul W. Frankland (2015). "Finding the engram". In: *Nature Reviews Neuroscience* 16, pp. 521–534.

Keiser, Ashley A, Lacie M Turnbull, Mara A Darian, Dana E Feldman, Iris Song, and Natalie C Tronson (2017). "Sex Differences in Context Fear Generalization and Recruitment of Hippocampus and Amygdala during Retrieval". In: *Neuropsychopharmacology* 42, pp. 397–407.

Klein, R., L.F. Parada, F. Coulier, and M. Barbacid (Dec. 1989). "trkB, a novel tyrosine protein kinase receptor expressed during mouse neural development." In: *The EMBO Journal* 8.12, pp. 3701–3709. ISSN: 02614189. DOI: 10.1002/j.1460-2075.1989.tb08545.x.

Korte, M, P Carroll, E Wolf, G Brem, H Thoenen, and T Bonhoeffer (1995). "Hippocampal long-term potentiation is impaired in mice lacking brain-derived neurotrophic factor." In: *Proceedings of the National Academy of Sciences of the United States of America* 92.19, pp. 8856–60.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances In Neural Information Processing Systems*, pp. 1–9. ISSN: 10495258. DOI: http://dx.doi.org/10.1016/j.protcy.2014.09.007. arXiv: 1102.0183.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, pp. 159–174.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444.

LeDoux, Joseph E (2000). "Emotion Circuits in the Brain". In: *Annu. Rev. Neurosci* 23, pp. 155–184.

LeDoux, Joseph E (2014). "Coming to terms with fear". In: *Proceedings of the National Academy of Sciences of the United States of America* 11.8, pp. 2871–2878. DOI: 10.1073/pnas.1400335111.

Li, Anan, H. Gong, B. Zhang, Q. Wang, C. Yan, J. Wu, L. Qian, S. Zeng, and Q. Luo (2010). "Micro-Optical Sectioning Tomography to Obtain a HIgh-Resolution Atlas of the Mouse Brain". In: *Science* 1404.2010, pp. 1404–1408. ISSN: 10959203. DOI: 10.1126/science.1191776.

Li, Yun, Daishi Yui, Bryan W. Luikart, Renée M. McKay, Yanjiao Li, John L. Rubenstein, and Luis F. Parada (2012). "Conditional ablation of brain-derived neurotrophic factor-TrkB signaling impairs striatal neuron development". In: *Proceedings of the National Academy of Sciences of the United States of America* 109.38, pp. 15491–15496. ISSN: 00278424. DOI: 10.1073/pnas.1212899109.

Martin, E Anne, Derek Woodruff, Randi L Rawson, Megan E Williams, and Megan Williams (2017). "Examining Hippocampal Mossy Fiber Synapses by 3D Electron Microscopy in Wildtype and Kirrel3 Knockout Mice 3D EM of mossy fiber synapses in Kirrel3 KO mice". In: *eNeuro* 10.June, pp. 88–17.

Martin-Zanca, Dionisio, Stephen H. Hughes, and Mariano Barbacid (Feb. 1986). "A human oncogene formed by the fusion of truncated tropomyosin and protein tyrosine kinase sequences". In: *Nature* 319.6056, pp. 743–748. ISSN: 0028-0836. DOI: 10.1038/319743a0.

Maška, Martin, Vladimır Ulman, David Svoboda, Pavel Matula, Petr Matula, Cristina Ederra, Ainhoa Urbiola, Tomás España, Subramanian Venkatesan, Deepak MW Balak, et al. (2014). "A benchmark for comparison of cell tracking algorithms". In: *Bioinformatics* 30.11, pp. 1609–1617.

McDole, Katie, Léo Guignard, Fernando Amat, Andrew Berger, Grégoire Malandain, Loïc A. Royer, Srinivas C. Turaga, Kristin Branson, and Philipp J. Keller (2018). "In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level". In: *Cell*, pp. 1–18. ISSN: 00928674. DOI: 10.1016/j.cell.2018.09.031.

McQuin, Claire, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A. Cimini, Kyle W Karhohs, Minh Doan, Liya Ding, Susanne M Rafelski, Derek Thirstrup, Winfried Wiegraebe, Shantanu Singh, Tim Becker, Juan C. Caicedo, and Anne E Carpenter (July 2018). "CellProfiler 3.0: Next-generation image processing for biology". In:

PLOS Biology 16.7. Ed. by Tom Misteli, e2005970. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.2005970.

Meijering, Erik (2012). "Cell segmentation: 50 Years down the road [life Sciences]". In: *IEEE Signal Processing Magazine* 29.5, pp. 140–145. ISSN: 10535888. DOI: 10.1109/MSP.2012.2204190.

Meijering, Erik, Anne E. Carpenter, Hanchuan Peng, Fred A. Hamprecht, and Jean-Christophe Olivo-Marin (Dec. 2016). "Imagining the future of bioimage analysis". In: *Nature Biotechnology* 34.12, pp. 1250–1255. ISSN: 1087-0156. DOI: 10.1038/nbt.3722.

Minichiello, Liliana, Martin Korte, David Wolfer, Ralf Kühn, Klaus Unsicker, Vincenzo Cestari, Clelia Rossi-Arnaud, Hans Peter Lipp, Tobias Bonhoeffer, and Rüdiger Klein (1999). "Essential role for TrkB receptors in hippocampus-mediated learning". In: *Neuron* 24.2, pp. 401–414. ISSN: 08966273. DOI: 10.1016/S0896-6273(00)80853-3.

Moen, Erick, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen (2019). "Deep learning for cellular image analysis". In: *Nature Methods*. ISSN: 1548-7091. DOI: 10.1038/s41592-019-0403-1.

Murawski, Nathen J, Anna Y Klintsova, and Mark E Stanton (2012). "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". In: *Neuroscience* 206, pp. 89–99.

Nath, Tanmay, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis (2019). "Using DeepLabCut for 3D markerless pose estimation across species and behaviors". In: *Nature Protocols* 14.7, pp. 2152–2176. ISSN: 17502799. DOI: 10.1038/s41596-019-0176-0.

Niedworok, Christian J., Alexander P.Y. Brown, M. Jorge Cardoso, Pavel Osten, Sebastien Ourselin, Marc Modat, and Troy W. Margrie (2016). "AMAP is a validated pipeline for registration and segmentation of high-resolution mouse brain data". In: *Nature Communications* 7.May, pp. 1–9. ISSN: 20411723. DOI: 10.1038/ncomms11879.

Nikoletopoulou, Vassiliki, Heiko Lickert, José Maria Frade, Chantal Rencurel, Patrizia Giallonardo, Lixin Zhang, Miriam Bibel, and Yves-Alain Barde (Sept. 2010). "Neurotrophin receptors TrkA and TrkC cause neuronal death whereas TrkB does not". In: *Nature* 467.7311, pp. 59–63. ISSN: 0028-0836. DOI: 10.1038/nature09336.

Osten, Pavel and Troy W. Margrie (2013). "Mapping brain circuitry with a light microscope". In: *Nature Methods* 10.6, pp. 515–523. ISSN: 15487091. DOI: 10.1038/nmeth.2477. eprint: NIHMS150003.

Ounkomol, Chawin, Sharmishtaa Seshamani, Mary M Maleckar, Forrest Collman, and Gregory R Johnson (2018). "Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy". In: *Nature methods* 15.11, pp. 917–920.

Park, Hyungju and Muming Poo (2013). "Neurotrophin regulation of neural circuit development and function". In: *Nature Reviews Neuroscience* 14, pp. 7–23. DOI: 10.1038/nrn3379.

Paxinos, George and Keith B. J. Franklin (2004). *The mouse brain in stereotaxic coordinates*. Compact 2nd. Amsterdam ; Boston: Elsevier Academic Press. ISBN: 012547640X (alk. paper).

Ramamoorthi, Kartik, Robin Fropf, Gabriel M Belfort, Helen L Fitzmaurice, Ross M McKinney, Rachael L Neve, Tim Otto, and Yingxi Lin (2011). "Npas4 regulates a transcriptional program in CA3 required for contextual memory formation". In: *Science* 334.6063, pp. 1669–1675.

Rauskolb, Stefanie, Marta Zagrebelsky, Anita Dreznjak, Rubén Deogracias, Tomoya Matsumoto, Stefan Wiese, Beat Erne, Michael Sendtner, Nicole Schaeren-Wiemers, Martin Korte, and Yves-Alain Barde (2010). "Global Deprivation of Brain-Derived Neurotrophic Factor in the CNS Reveals an Area-Specific Requirement for Dendritic Growth". In: *Journal of Neuroscience* 30.5. DOI: 10.1523/JNEUROSCI.5100-09.2010.

Rodriguez-Tebar, A. and Y. A. Barde (1988). "Binding characteristics of brain-derived neurotrophic factor to its receptor on neurons from the chick embryo". In: *Journal of Neuroscience* 8.9, pp. 3337–3342. ISSN: 02706474. DOI: 10.1523/jneurosci.08-09-03337.1988.

Rohrer, Baerbel, Juan I. Korenbrot, Matthew M. LaVail, Louis F. Reichardt, and Baoji Xu (Oct. 1999). "Role of Neurotrophin Receptor TrkB in the Maturation of Rod Photoreceptors and Establishment of Synaptic Transmission to the Inner Retina". In: *The Journal of Neuroscience* 19.20, pp. 8919–8930. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.19-20-08919.1999.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241.

Ruediger, Sarah, Claudia Vittori, Ewa Bednarek, Christel Genoud, Piergiorgio Strata, Benedetto Sacchetti, and Pico Caroni (May 2011). "Learning-related feedforward inhibitory connectivity growth required for memory precision". In: *Nature* 473.7348, pp. 514–518.

Sairanen, Mikko, Guilherme Lucas, Patrik Ernfors, Maija Castrén, and Eero Castrén (2005). "Brain-derived neurotrophic factor and antidepressant drugs have different but coordinated effects on neuronal turnover, proliferation, and survival in the adult dentate gyrus". In: *Journal of Neuroscience* 25.5, pp. 1089–1094. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.3741-04.2005.

Sasi, Manju (2020). "A mouse model for genetic deletion of presynaptic BDNF from adult hippocampal mossy fiber terminals". Doctoral Thesis. Universität Würzburg. DOI: 10.25972/OPUS-18625.

Sasi, Manju, Beatrice Vignoli, Marco Canossa, and Robert Blum (2017). "Neurobiology of local and Intracellular BDNF signaling". In: *European Journal of Physiology*.

Schildt, Sandra, Thomas Endres, Volkmar Lessmann, and Elke Edelmann (2013). "Acute and chronic interference with BDNF/TrkB-signaling impair LTP selectively at mossy fiber synapses in the CA3 region of mouse hippocampus". In: *Neuropharmacology* 71, pp. 247–254.

Schmitz, Christoph, Hubert Korr, and Helmut Heinsen (1999). "Design-based counting techniques: the real problems". In: *Trends in neurosciences* 22.8, p. 345.

Schneider, C. A., W. S. Rasband, and K. W. Eliceiri (2012). "NIH Image to ImageJ: 25 years of image analysis". In: *Nat Methods* 9.7, pp. 671–5.

Segebarth, Dennis, Matthias Griebel, Nikolai Stein, Cora R von Collenberg, Corinna Martin, Dominik Fiedler, Lucas B Comeras, Anupam Sah, Victoria Schoeffler, Teresa Lüffe, Alexander Dürr, Rohini Gupta, Manju Sasi, Christina Lillesaar, Maren D Lange, Ramon O Tasan, Nicolas Singewald, Hans-Christian Pape, Christoph M Flath, and Robert Blum (Oct. 2020a). "On the objectivity, reliability, and validity of deep learning enabled bioimage analyses". In: *eLife* 9. ISSN: 2050-084X. DOI: 10.7554/eLife.59780.

Segebarth, Dennis, Matthias Griebel, Nikolai Stein, Cora R. von Collenberg, Corinna Martin, Dominik Fiedler, Lucas B. Comeras, Anupam Sah, Victoria Schoeffler, Theresa Lüffe, Alexander Dürr, Rohini Gupta, Manju Sasi, Christine Lillesaar, Maren D. Lange, Ramon O. Tasan, Nicolas Singewald, Hans-Christian Pape, Christoph M. Flath, and

Robert Blum (2020b). *data from: On the objectivity, reliability, and validity of deep learning enabled bioimage analyses*. Dataset.

Sendtner, Michael, B. Holtmann, R. Kolbeck, Hans Thoenen, and Yves-Alain Barde (1992). "Brain-derived neurotrophic factor prevents the death of motorneurons in newborn rats after nerve section". In: *Nature* 360, pp. 757–759.

Shuvaev, Sergey A, Alexander A Lazutkin, Alexander V Kedrov, Konstantin V Anokhin, Grigori N Enikolopov, and Alexei A Koulakov (2017). "Dalmatian: An algorithm for automatic cell detection and counting in 3d". In: *Frontiers in neuroanatomy* 11, p. 117.

Tadrous, Paul J. (2010). "On the concept of objectivity in digital image analysis in pathology". In: *Pathology* 42.3, pp. 207–211. ISSN: 14653931. DOI: 10.3109/00313021003641758.

Tayler, Kaycie K, Kazumasa Z Tanaka, Leon G Reijmers, and Brian J Wiltgen (2013). "Reactivation of neural ensembles during the retrieval of recent and remote memory". In: *Current Biology* 23.2, pp. 99–106.

Taylor, Barry N and Chris E Kuyatt (1994). *Guidelines for evaluating and expressing the uncertainty of NIST measurement results*. Tech. rep. US Department of Commerce, Technology Administration, National Institute of Standards and Technology.

Thoenen, Hans (Oct. 1995). "Neurotrophins and Neuronal Plasticity". In: *Science* 270.5236, pp. 593–598. ISSN: 0036-8075. DOI: 10.1126/science.270.5236.593.

Tonegawa, Susumu, Xu Liu, Steve Ramirez, and Roger Redondo (2015). "Memory Engram Cells Have Come of Age". In: *Neuron* 87.5, pp. 918–931.

Tovote, Philip, Jonathan Paul Fadok, and Andreas Lüthi (2015). "Neuronal circuits for fear and anxiety". In: *Nature Reviews Neuroscience* 16, p. 317329.

Van Valen, David A., Takamasa Kudo, Keara M. Lane, Derek N. Macklin, Nicolas T. Quach, Mialy M. DeFelice, Inbal Maayan, Yu Tanouchi, Euan A. Ashley, and Markus W. Covert (2016). "Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments". In: *PLoS Computational Biology*. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1005177.

Wang, Hongda, Yair Rivenson, Yiyin Jin, Zhensong Wei, Ronald Gao, Harun Günaydın, Laurent A. Bentolila, Comert Kural, and Aydogan Ozcan (2019). "Deep learning enables cross-modality super-resolution in fluorescence microscopy". In: *Nature Methods* 16.1, pp. 103–110. ISSN: 15487105. DOI: 10.1038/s41592-018-0239-0.

Warfield, Simon K, Kelly H Zou, and William M Wells (2004). "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation". In: *IEEE transactions on medical imaging* 23.7, pp. 903–921.

Wiera, Grzegorz and Jerzy W. Mozrzymas (2015). “Extracellular proteolysis in structural and functional plasticity of mossy fiber synapses in hippocampus”. In: *Frontiers in Cellular Neuroscience* 9.November, pp. 1–21. ISSN: 1662-5102. DOI: 10.3389/fncel.2015.00427.

Zheng, Alice and Amanda Casari (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. " O’Reilly Media, Inc."

Zheng, K., J. J. An, F. Yang, W. Xu, Z.-Q. D. Xu, J. Wu, T. G. M. Hokfelt, A. Fisahn, B. Xu, and B. Lu (Oct. 2011). “TrkB signaling in parvalbumin-positive interneurons is critical for gamma-band network synchronization in hippocampus”. In: *Proceedings of the National Academy of Sciences* 108.41, pp. 17201–17206. ISSN: 0027-8424. DOI: 10.1073/pnas.1114241108.

7 Supplementary Figures

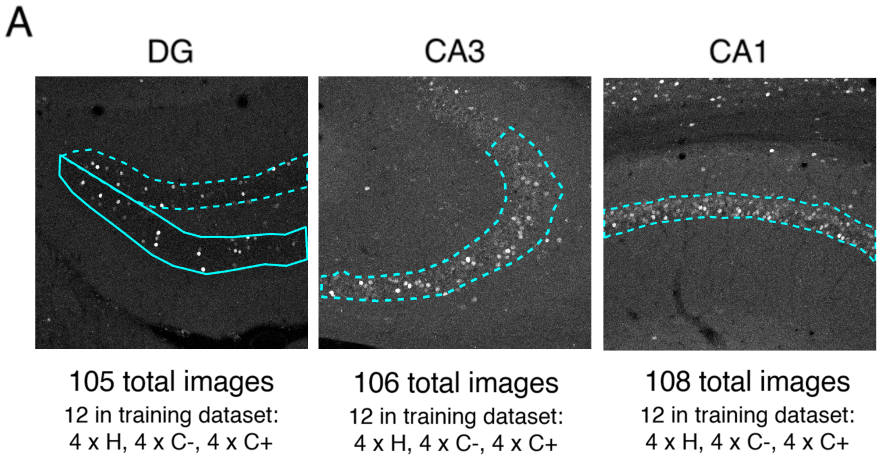


Figure S1: Illustration of the bioimage dataset which was used for the comparison of DL-based strategies.
A. A total of 319 images showing cFOS immunoreactivity in the dorsal hippocampus of wildtype mice was split up in 105 images of the Dentate gyrus, 106 images of CA3 and 108 images of CA1. To create 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 $\sim$ 10,000 cFOS-positive feature within the NeuN-positive areas in all 283 images used for bioimage analysis, which is equivalent to $\sim$ 35 features per image. Reproduced from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

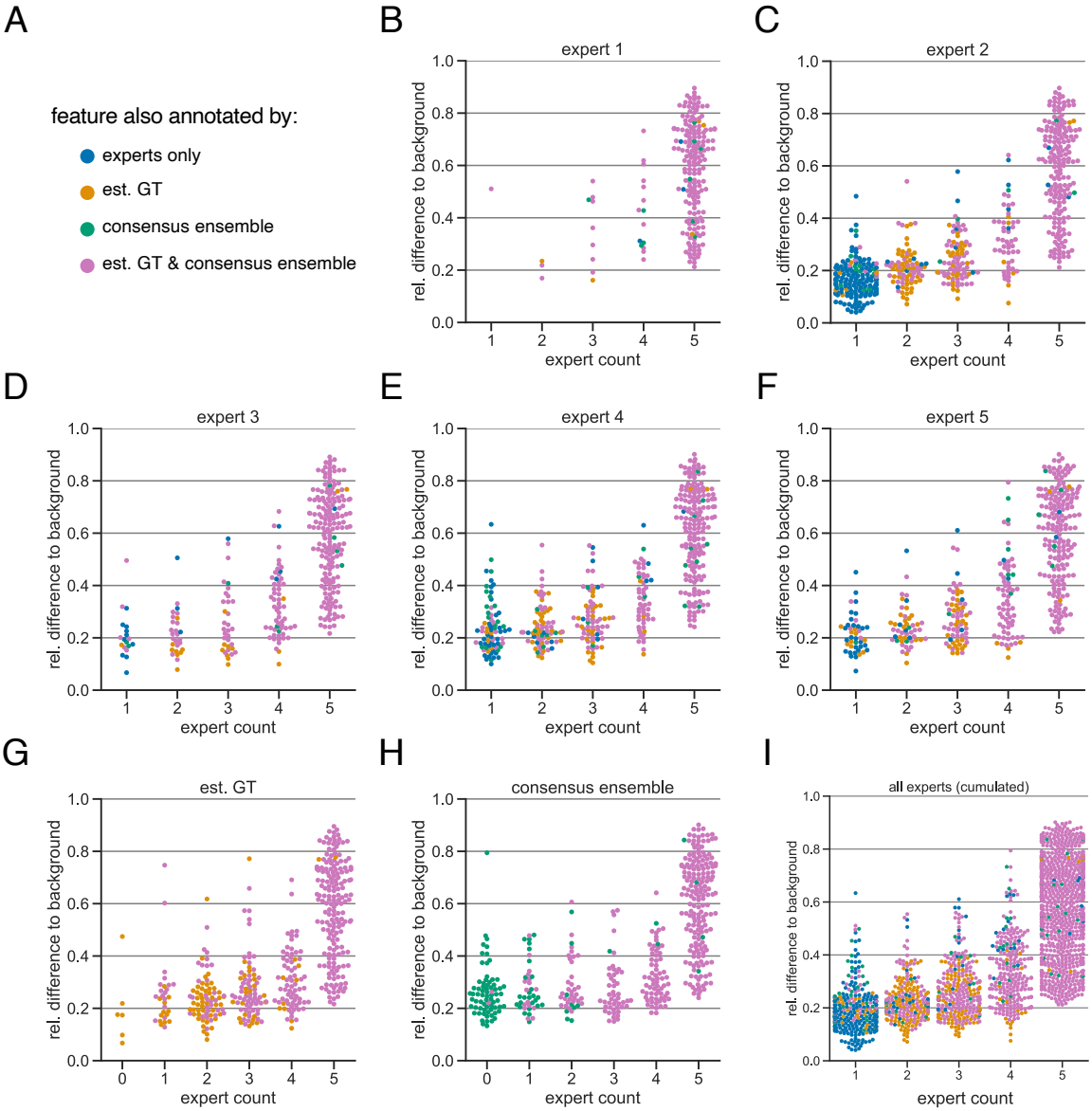


Figure S2: Extended subjectivity analysis.
The subjectivity analysis depicts the relationship between the relative intensity difference of a florescent feature (ROI) to the background and the annotation count of human experts. A visual interpretation indicates that the annotation probability of a ROI 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. An IoU threshold of $t = 0.5$ was used for ROI matching. The expert 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 were 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 revealed 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 showed that human annotators may have missed several ROIs (green) even with a large relative difference to the background.
I. Cumulative summary of **B-F**. Reproduced from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

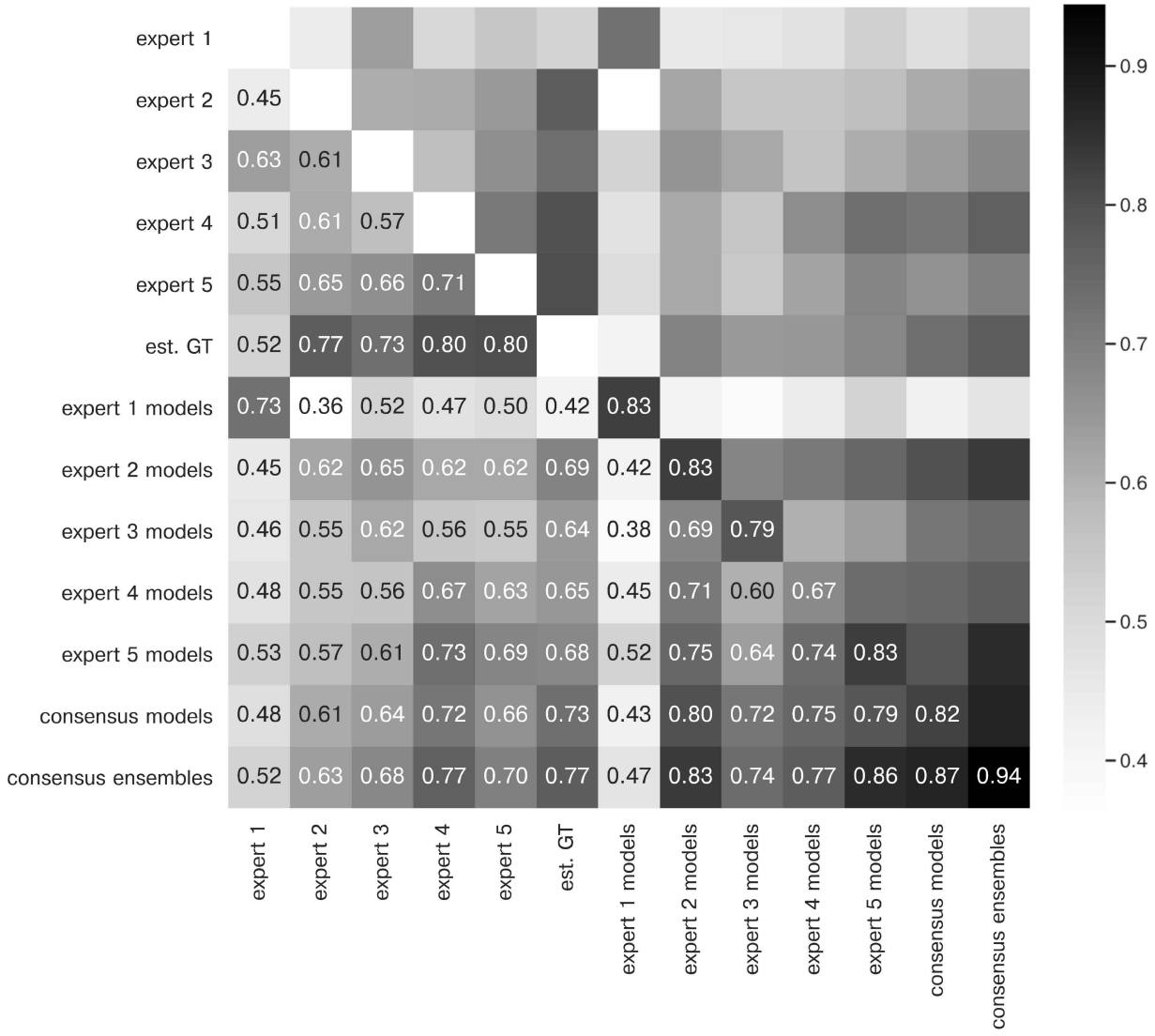


Figure S3: *Extended similarity analysis: F1 score.* The heatmap shows the mean F1-scores at a matching IoU-threshold of $t = 0.5$ for the image feature annotations of the indicated experts. Segmentation masks of the five human experts ($N_{(expert)}$=1 per expert), the estimated ground-truth ($N_{(est.\ GT)}$=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* showed 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). Reproduced from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

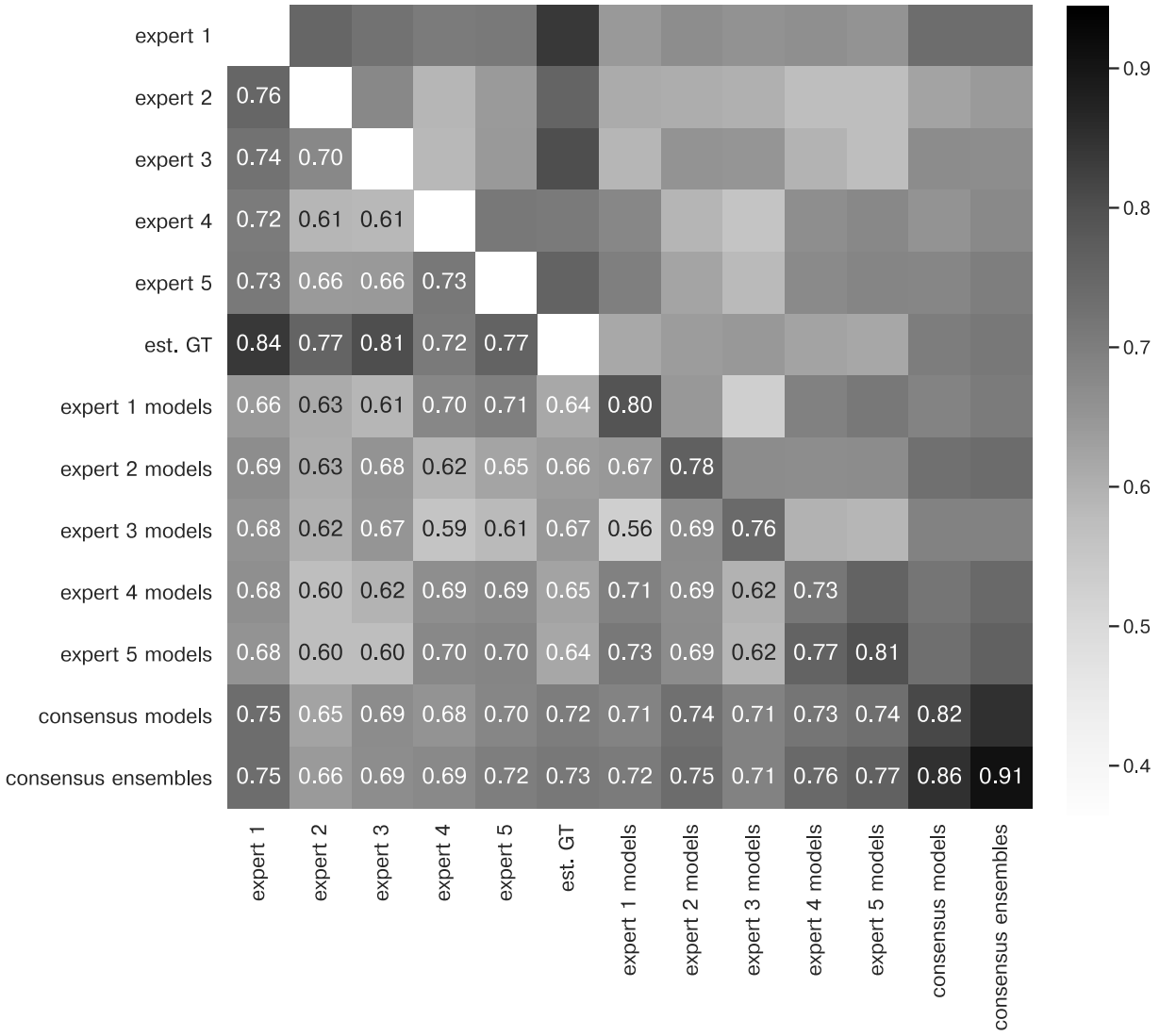


Figure S4: *Extended similarity analysis: mean IoU.* The heatmap shows the mean of mean IoU for the image feature annotations of the indicated experts. Segmentation masks of the five human experts ($N_{(expert)}$= 1 per expert), the estimated ground-truth ($N_{(est.\ GT)}$= 1), the respective *expert 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). Again, *consensus ensembles* showed highest reliability (0.91). Est. GT annotations are directly derived from manual expert annotations, which renders this comparison favorable. Reproduced from Segebarth, Griebel, Stein, R von Collenberg, et al. (2020a).

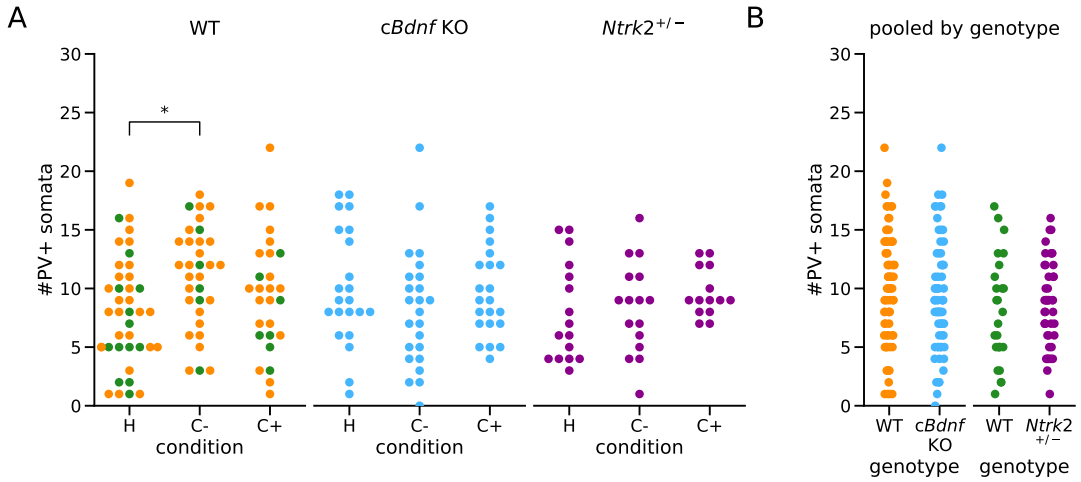


Figure S5: Quantification of Parv-positive somata in DG in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice.
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: F(2, 96)=3.85, P=0.025; c*Bdnf* KO: F(2, 65)=1.00, P=0.375; *Ntrk2*$^{+/-}$: H(2)=2.27, P=0.320; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (*: p <0.05).
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: T(138.74)=0.77, p=0.445; WT vs. *Ntrk2*$^{+/-}$: T(42.97)=-0.61, p=0.544.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=40, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=27, $n_{(cBdnf\ KO\ H)}$=22, $n_{(cBdnf\ KO\ C-)}$=24, $n_{(cBdnf\ KO\ C+)}$=22, $n_{(Ntrk2^{+/-}\ H)}$=16, $n_{(Ntrk2^{+/-}\ C-)}$=16, $n_{(Ntrk2^{+/-}\ C+)}$=14.

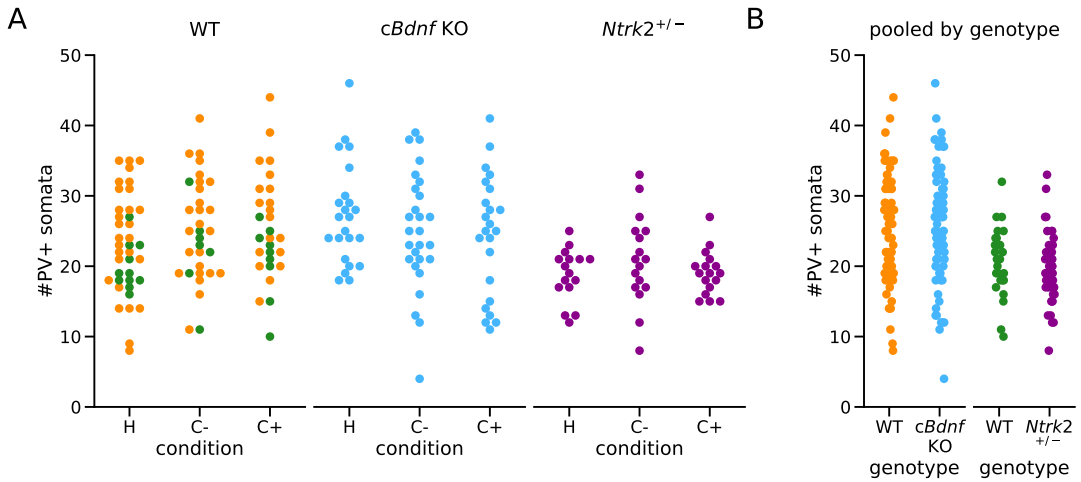


Figure S6: Quantification of Parv-positive somata in CA3 in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: F(2, 96)=1.18, P=0.311; c*Bdnf* KO: F(2, 70)=0.83, P=0.439; *Ntrk2*$^{+/-}$: H(2)=1.61, P=0.448.
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: T(142.76)=0.21, p=0.836; WT vs. *Ntrk2*$^{+/-}$: T(51.91)=1.05, p=0.296.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=39, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=28, $n_{(cBdnf\ KO\ H)}$=23, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=23, $n_{(Ntrk2^{+/-}\ H)}$=16, $n_{(Ntrk2^{+/-}\ C-)}$=17, $n_{(Ntrk2^{+/-}\ C+)}$=17.

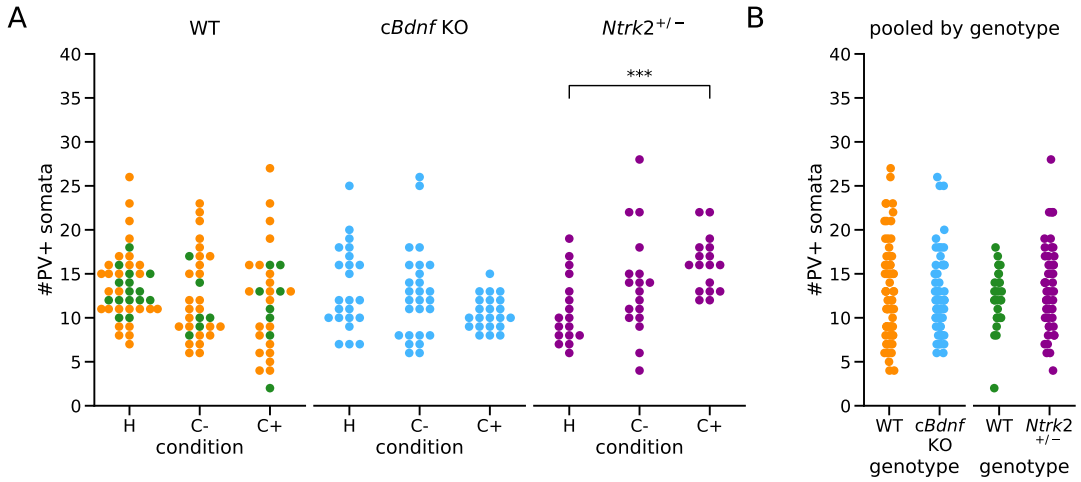


Figure S7: Quantification of Parv-positive somata in CA1 in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=2.90, P=0.234; c*Bdnf* KO: H(2)=4.17, P=0.125; *Ntrk2*$^{+/-}$: F(2, 49)=5.86, P=0.005; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (***: p<0.001).
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=2811.0, p=0.469; WT vs. *Ntrk2*$^{+/-}$: U=647.5, p=0.295.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=42, $n_{(WT\ C-)}$=31, $n_{(WT\ C+)}$=29, $n_{(cBdnf\ KO\ H)}$=23, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=22, $n_{(Ntrk2^{+/-}\ H)}$=17, $n_{(Ntrk2^{+/-}\ C-)}$=18, $n_{(Ntrk2^{+/-}\ C+)}$=17.

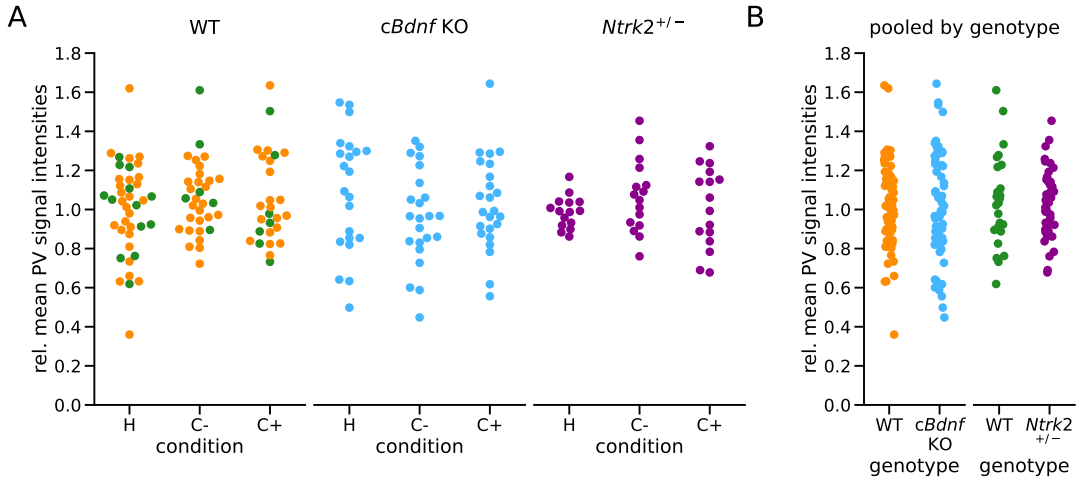


Figure S8: Mean Parv-signal intensities of Parv-positive somata in DG in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice.
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=0.63, P=0.729; c*Bdnf* KO: F(2, 66)=1.46, P=0.241; *Ntrk2*$^{+/-}$: H(2)=1.62, P=0.444.
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=2528.0, p=0.971; WT vs. *Ntrk2*$^{+/-}$: T(39.40)=0.42, p=0.677.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=40, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=27, $n_{(cBdnf\ KO\ H)}$=22, $n_{(cBdnf\ KO\ C-)}$=23, $n_{(cBdnf\ KO\ C+)}$=24, $n_{(Ntrk2^{+/-}\ H)}$=15, $n_{(Ntrk2^{+/-}\ C-)}$=16, $n_{(Ntrk2^{+/-}\ C+)}$=16.

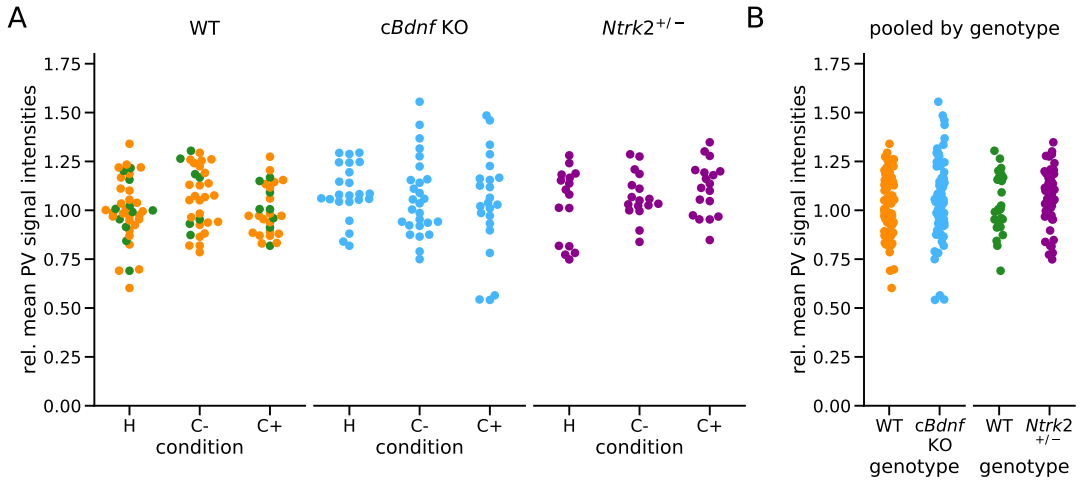


Figure S9: Mean Parv-signal intensities of Parv-positive somata in CA3 in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: F(2, 96)=1.84, P=0.163; c*Bdnf* KO: F(2, 70)=0.46, P=0.635; *Ntrk2*$^{+/-}$: H(2)=1.49, P=0.475.
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: T(144)=-1.63, p=0.104; WT vs. *Ntrk2*$^{+/-}$: T(49.91)=-1.18, p=0.243.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=39, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=28, $n_{(cBdnf\ KO\ H)}$=23, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=23, $n_{(Ntrk2^{+/-}\ H)}$=16, $n_{(Ntrk2^{+/-}\ C-)}$=17, $n_{(Ntrk2^{+/-}\ C+)}$=18.

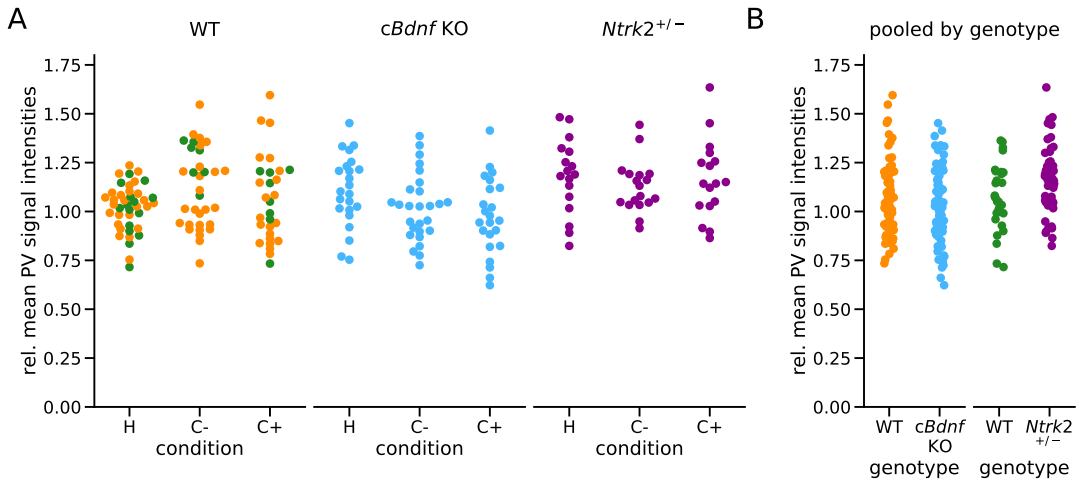


Figure S10: Mean Parv-signal intensities of Parv-positive somata in CA1 in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=3.76, P=0.152; c*Bdnf* KO: F(2, 70)=3.28, P=0.043; *Ntrk2*$^{+/-}$: F(2, 49)=0.43, P=0.651.
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=2787.0, p=0.633; WT vs. *Ntrk2*$^{+/-}$: T(62.20)=-2.00, p=0.050.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=41, $n_{(WT\ C-)}$=33, $n_{(WT\ C+)}$=29, $n_{(cBdnf\ KO\ H)}$=23, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=23, $n_{(Ntrk2^{+/-}\ H)}$=17, $n_{(Ntrk2^{+/-}\ C-)}$=18, $n_{(Ntrk2^{+/-}\ C+)}$=17.

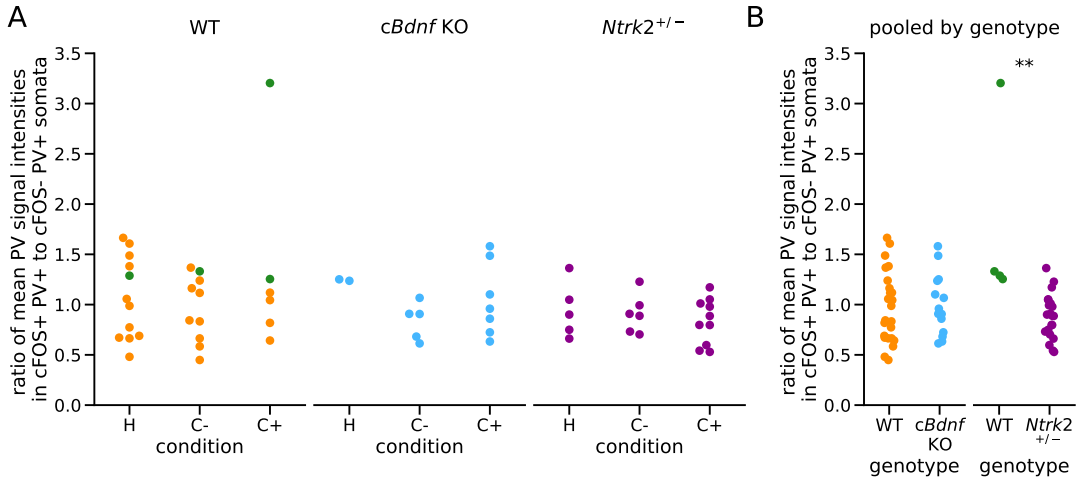


Figure S11: Ratio of mean Parv-signal intensities of cFOS-positive Parv-positive somata compared to cFOS-negative Parv-positive somata in DG in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice.
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=0.52, P=0.771; c*Bdnf* KO: F(2, 11)=1.60, P=0.246; *Ntrk2*$^{+/-}$: F(2, 18)=0.43, P=0.659.
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: T(30.70)=-0.26, p=0.800; WT vs. *Ntrk2*$^{+/-}$: M=81.0, p=0.544.
$N_{(WT\ H)}$=6, $N_{(WT\ C-}$=6, $N_{(WT\ C+)}$=5, $N_{(cBdnf\ KO\ H)}$=3, $N_{(cBdnf\ KO\ C-)}$=4, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=12, $n_{(WT\ C-)}$=10, $n_{(WT\ C+)}$=6, $n_{(cBdnf\ KO\ H)}$=2, $n_{(cBdnf\ KO\ C-)}$=5, $n_{(cBdnf\ KO\ C+)}$=7, $n_{(Ntrk2^{+/-}\ H)}$=5, $n_{(Ntrk2^{+/-}\ C-)}$=6, $n_{(Ntrk2^{+/-}\ C+)}$=10.

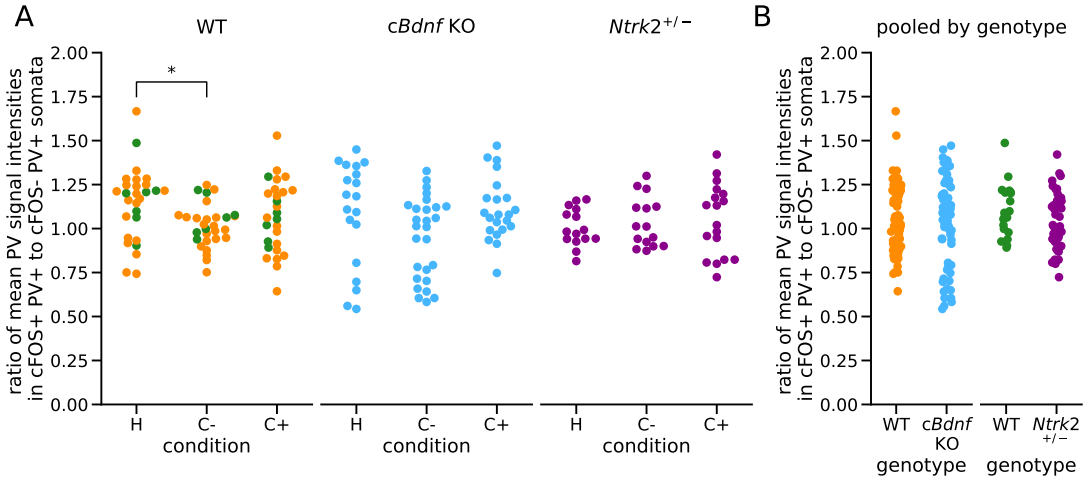


Figure S12: Ratio of mean Parv-signal intensities of cFOS-positive Parv-positive somata compared to cFOS-negative Parv-positive somata in CA3 in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=7.06, P=0.029; c*Bdnf* KO: H(2)=5.18, P=0.075; *Ntrk2*$^{+/-}$: H(2)=0.57, P=0.751; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (*: p <0.05).
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=2200.0, p=0.791; WT vs. *Ntrk2*$^{+/-}$: T(41.06)=1.72, p=0.092.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=28, $n_{(WT\ C-)}$=29, $n_{(WT\ C+)}$=27, $n_{(cBdnf\ KO\ H)}$=19, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=22, $n_{(Ntrk2^{+/-}\ H)}$=15, $n_{(Ntrk2^{+/-}\ C-)}$=15, $n_{(Ntrk2^{+/-}\ C+)}$=18.

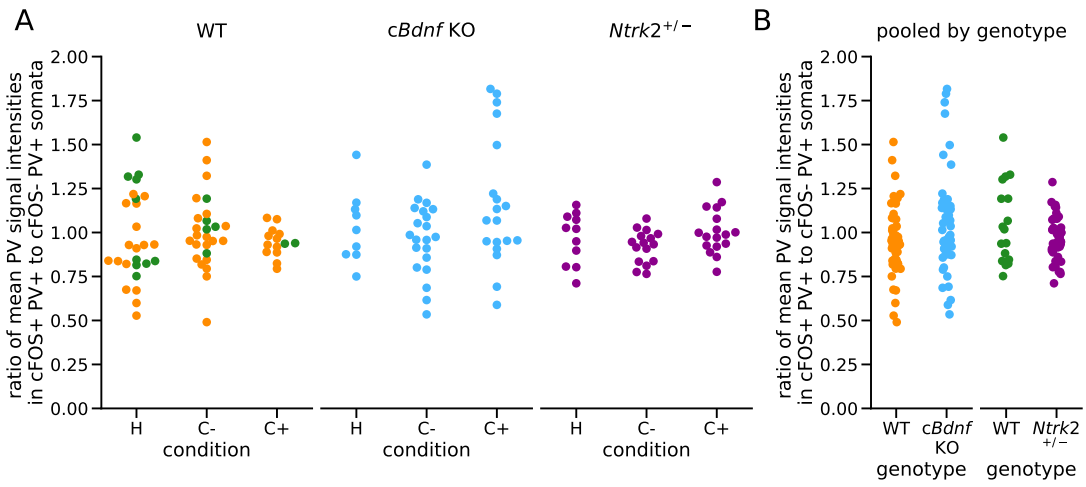


Figure S13: Ratio of mean Parv-signal intensities of cFOS-positive Parv-positive somata compared to cFOS-negative Parv-positive somata in CA1 in WT, *cBdnf* KO, and *Ntrk2*$^{+/-}$ mice

In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).

A. Comparisons of the treatment conditions within each genotype. WT: H(2)=1.41, P=0.494; c*Bdnf* KO: H(2)=2.60, P=0.272; *Ntrk2*$^{+/-}$: F(2, 41)=2.01, P=0.146

B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=948.0, p=0.073; WT vs. *Ntrk2*$^{+/-}$: U=439.0, p=0.299.

$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=3, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=26, $n_{(WT\ C-)}$=26, $n_{(WT\ C+)}$=14, $n_{(cBdnf\ KO\ H)}$=9, $n_{(cBdnf\ KO\ C-)}$=21, $n_{(cBdnf\ KO\ C+)}$=19, $n_{(Ntrk2^{+/-}\ H)}$=11, $n_{(Ntrk2^{+/-}\ C-)}$=16, $n_{(Ntrk2^{+/-}\ C+)}$=17.

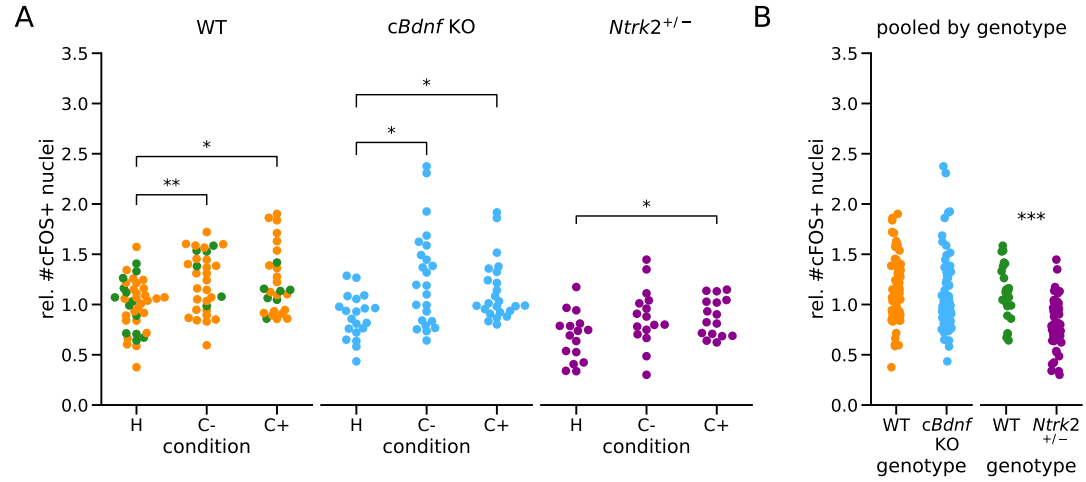


Figure S14: Quantification of cFOS-positive nuclei in DG in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice.

In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).

A. Comparisons of the treatment conditions within each genotype. WT: H(2)=10.45, P=0.005; c*Bdnf* KO: H(2)=10.65, P=0.005; *Ntrk2*$^{+/-}$: F(2, 45)=3.54, P=0.037; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (*: p<0.05, **: p<0.01).

B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=2828.0, p=0.154; WT vs. *Ntrk2*$^{+/-}$: T(49.20)=4.69, p<0.001.

$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=40, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=27, $n_{(cBdnf\ KO\ H)}$=20, $n_{(cBdnf\ KO\ C-)}$=24, $n_{(cBdnf\ KO\ C+)}$=24, $n_{(Ntrk2^{+/-}\ H)}$=16, $n_{(Ntrk2^{+/-}\ C-)}$=16, $n_{(Ntrk2^{+/-}\ C+)}$=16.

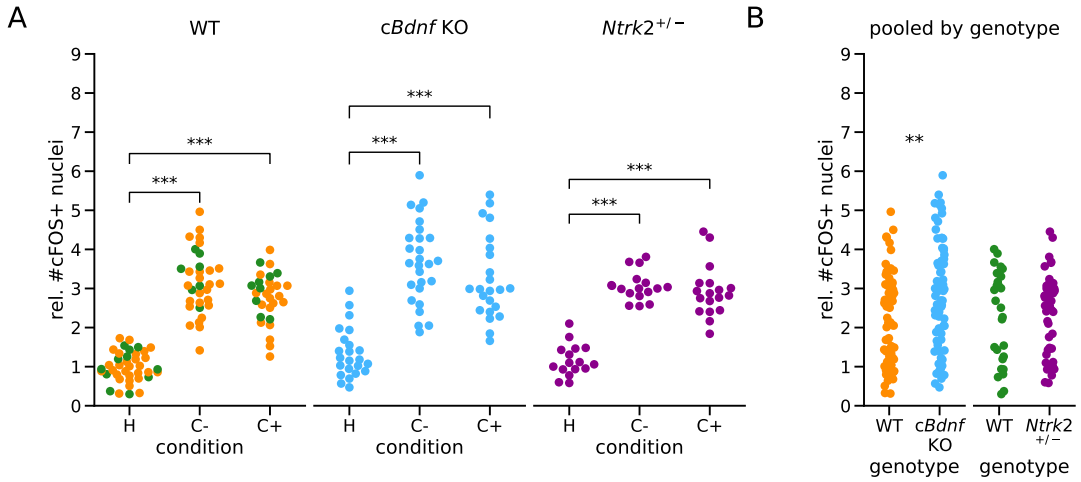


Figure S15: Quantification of cFOS-positive nuclei in CA3 in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=69.37, P<0.001; c*Bdnf* KO: F(2, 70)=41.73, P<0.001; *Ntrk2*$^{+/-}$: F(2, 48)=73.03, P<0.001; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (***: p<0.001).
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=1981.0, p=0.008; WT vs. *Ntrk2*$^{+/-}$: U=675.0, p=0.891.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=39, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=29, $n_{(cBdnf\ KO\ H)}$=23, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=23, $n_{(Ntrk2^{+/-}\ H)}$=16, $n_{(Ntrk2^{+/-}\ C-)}$=17, $n_{(Ntrk2^{+/-}\ C+)}$=18.

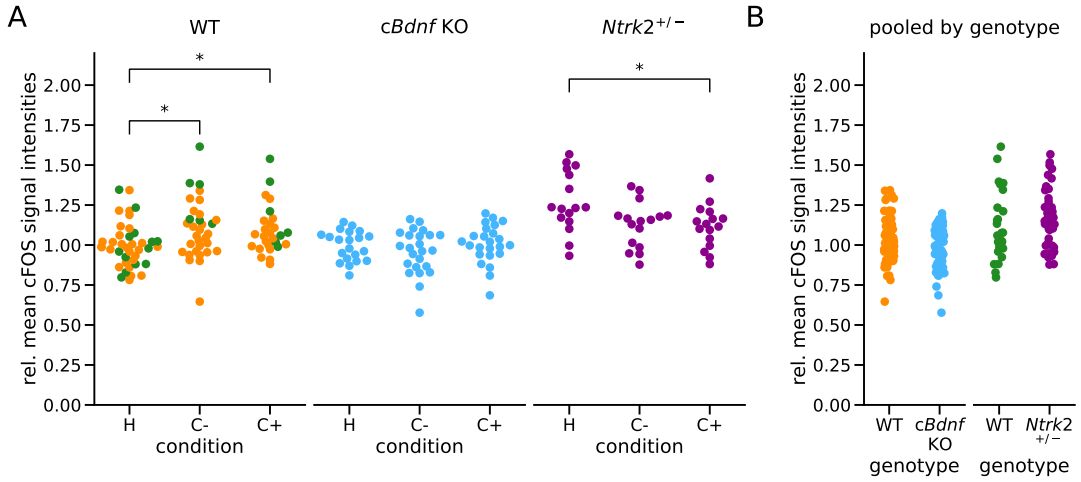


Figure S16: Mean cFOS-signal intensities of cFOS-positive nuclei in DG in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice.
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=9.49, P=0.009; c*Bdnf* KO: F(2, 66)=1.08, P=0.346; *Ntrk2*$^{+/-}$: F(2, 45)=4.85, P=0.012; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (*: p <0.05).
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=2925.0, p=0.097; WT vs. *Ntrk2*$^{+/-}$: T(42.17)=-1.04, p=0.306.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=40, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=27, $n_{(cBdnf\ KO\ H)}$=21, $n_{(cBdnf\ KO\ C-)}$=24, $n_{(cBdnf\ KO\ C+)}$=24, $n_{(Ntrk2^{+/-}\ H)}$=16, $n_{(Ntrk2^{+/-}\ C-)}$=16, $n_{(Ntrk2^{+/-}\ C+)}$=16.

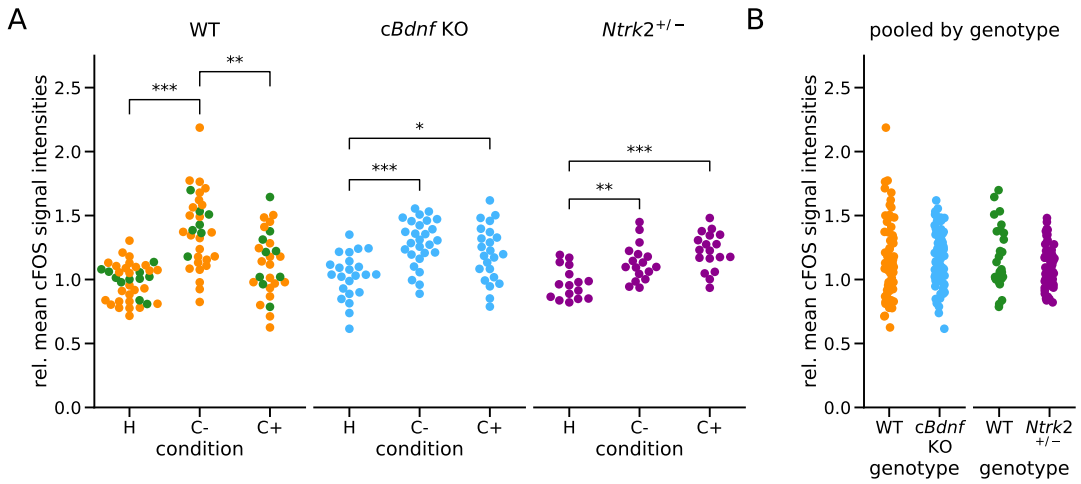


Figure S17: Mean cFOS-signal intensities of cFOS-positive nuclei in CA3 in WT, *cBdnf* KO, and *Ntrk2*$^{+/-}$ mice
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=33.56, P<0.001; c*Bdnf* KO: F(2, 70)=11.70, P<0.001; *Ntrk2*$^{+/-}$: F(2, 48)=13.79, P<0.001; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (*: p<0.05, **: p<0.01, ***: p<0.001).
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=2228.0, p=0.147; WT vs. *Ntrk2*$^{+/-}$: U=762.0, p=0.443.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=38, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=28, $n_{(cBdnf\ KO\ H)}$=23, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=23, $n_{(Ntrk2^{+/-}\ H)}$=16, $n_{(Ntrk2^{+/-}\ C-)}$=17, $n_{(Ntrk2^{+/-}\ C+)}$=18.

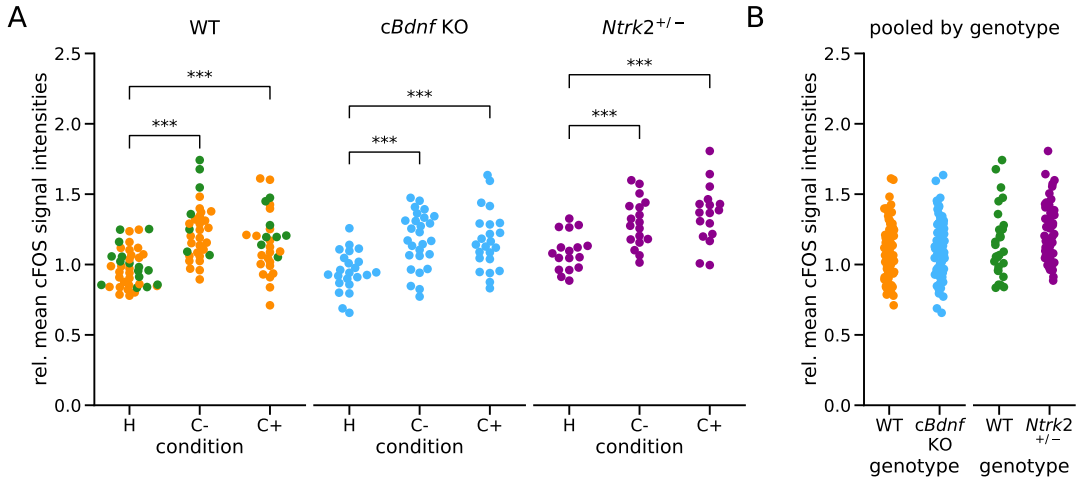


Figure S18: Mean cFOS-signal intensities of cFOS-positive nuclei in CA1 in WT, *cBdnf* KO, and *Ntrk2*$^{+/-}$ mice
In WT animals, marker color encodes the bioimage dataset for which the respective mouse was used as control (orange: c*Bdnf* KO; green: *Ntrk2*$^{+/-}$).
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=30.41, P<0.001; c*Bdnf* KO: F(2, 70)=11.16, P<0.001; *Ntrk2*$^{+/-}$: F(2, 49)=11.17, P<0.001; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (***: p<0.001).
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: T(144)=-0.40, p=0.691; WT vs. *Ntrk2*$^{+/-}$: T(52.94)=-1.75, p=0.087.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=42, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=29, $n_{(cBdnf\ KO\ H)}$=23, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=23, $n_{(Ntrk2^{+/-}\ H)}$=17, $n_{(Ntrk2^{+/-}\ C-)}$=18, $n_{(Ntrk2^{+/-}\ C+)}$=17.

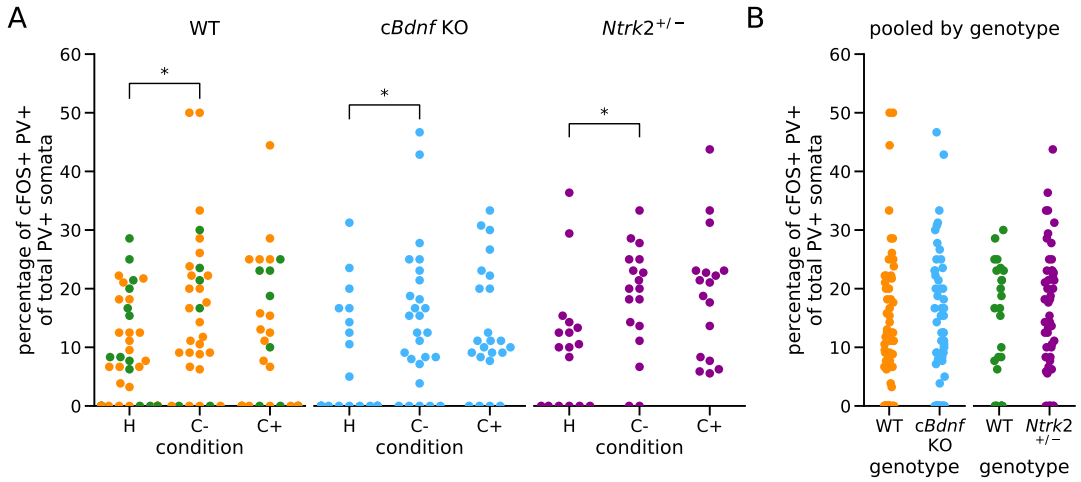


Figure S19: Percentage of cFOS-positive Parv-positive somata among all Parv-positive somata in CA1 in WT, *cBdnf* KO, and *Ntrk2*$^{+/-}$ mice

A. Comparisons of the treatment conditions within each genotype. WT: H(2)=7.65, P=0.022; *cBdnf* KO: H(2)=8.08, P=0.018; *Ntrk2*$^{+/-}$: H(2)=7.99, P=0.018; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (*: p<0.05).

B. Data of all conditions was pooled within the indicated genotypes. WT vs. *cBdnf* KO: U=2629.0, p=0.673; WT vs. *Ntrk2*$^{+/-}$: U=621.5, p=0.191.

$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=43, $n_{(WT\ C-)}$=32, $n_{(WT\ C+)}$=29, $n_{(cBdnf\ KO\ H)}$=23, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=23, $n_{(Ntrk2^{+/-}\ H)}$=17, $n_{(Ntrk2^{+/-}\ C-)}$=18, $n_{(Ntrk2^{+/-}\ C+)}$=17.

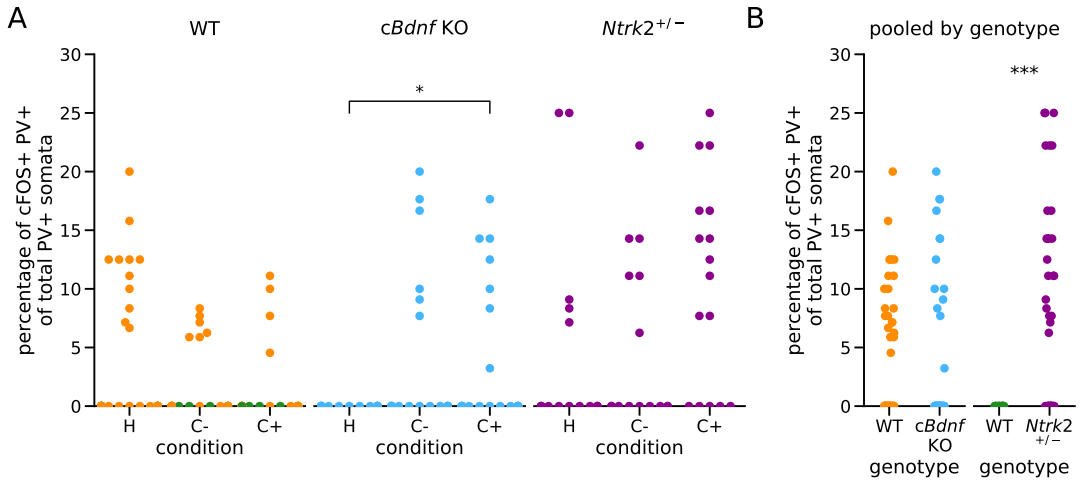


Figure S20: Percentage of cFOS-positive Parv-positive somata among all Parv-positive somata in DG in WT, *cBdnf* KO, and *Ntrk2*$^{+/-}$ mice.

A. Comparisons of the treatment conditions within each genotype. WT: H(2)=2.06, P=0.357; *cBdnf* KO: H(2)=6.27, P=0.044; *Ntrk2*$^{+/-}$: H(2)=5.38, P=0.068; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (*: p<0.05).

B. Data of all conditions was pooled within the indicated genotypes. WT vs. *cBdnf* KO: U=2514.5, p=0.246; WT vs. *Ntrk2*$^{+/-}$: U=275.0, p<0.001.

$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=39, $n_{(WT\ C-)}$=28, $n_{(WT\ C+)}$=24, $n_{(cBdnf\ KO\ H)}$=19, $n_{(cBdnf\ KO\ C-)}$=24, $n_{(cBdnf\ KO\ C+)}$=24, $n_{(Ntrk2^{+/-}\ H)}$=16, $n_{(Ntrk2^{+/-}\ C-)}$=15, $n_{(Ntrk2^{+/-}\ C+)}$=16.

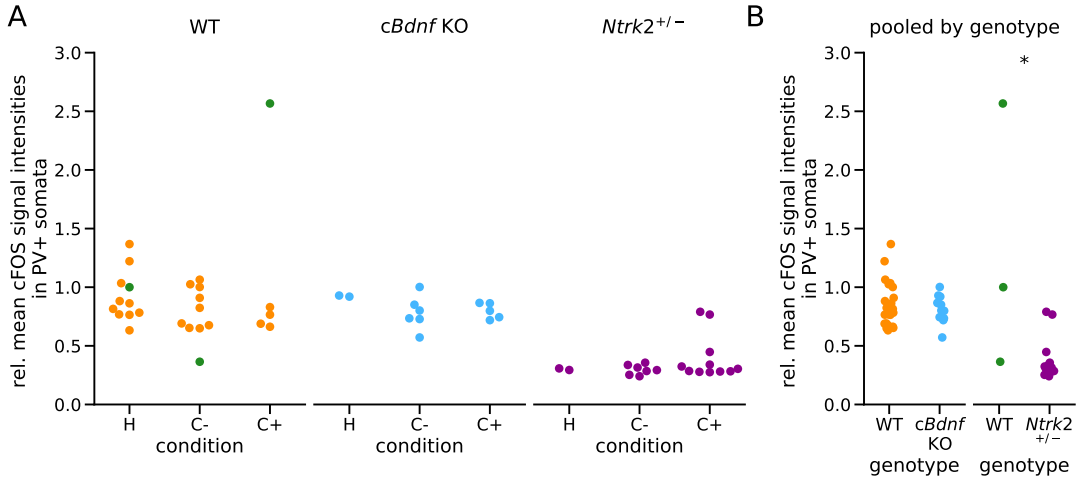


Figure S21: Mean cFOS-signal intensities of cFOS-positive nuclei in Parv-positive somata in DG in WT, *cBdnf* KO, and *Ntrk2*$^{+/-}$ mice.
A. Comparisons of the treatment conditions within each genotype. WT: H(2)=1.28, P=0.527; c*Bdnf* KO: F(2, 10)=1.31, P=0.312; *Ntrk2*$^{+/-}$: H(2)=0.57, P=0.752.
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=153.0, p=0.921; WT vs. *Ntrk2*$^{+/-}$: U=57.0, p=0.016.
$N_{(WT\ H)}$=6, $N_{(WT\ C-}$=6, $N_{(WT\ C+)}$=5, $N_{(cBdnf\ KO\ H)}$=3, $N_{(cBdnf\ KO\ C-)}$=4, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=11, $n_{(WT\ C-)}$=10, $n_{(WT\ C+)}$=5, $n_{(cBdnf\ KO\ H)}$=2, $n_{(cBdnf\ KO\ C-)}$=6, $n_{(cBdnf\ KO\ C+)}$=5, $n_{(Ntrk2^{+/-}\ H)}$=2, $n_{(Ntrk2^{+/-}\ C-)}$=7, $n_{(Ntrk2^{+/-}\ C+)}$=11.

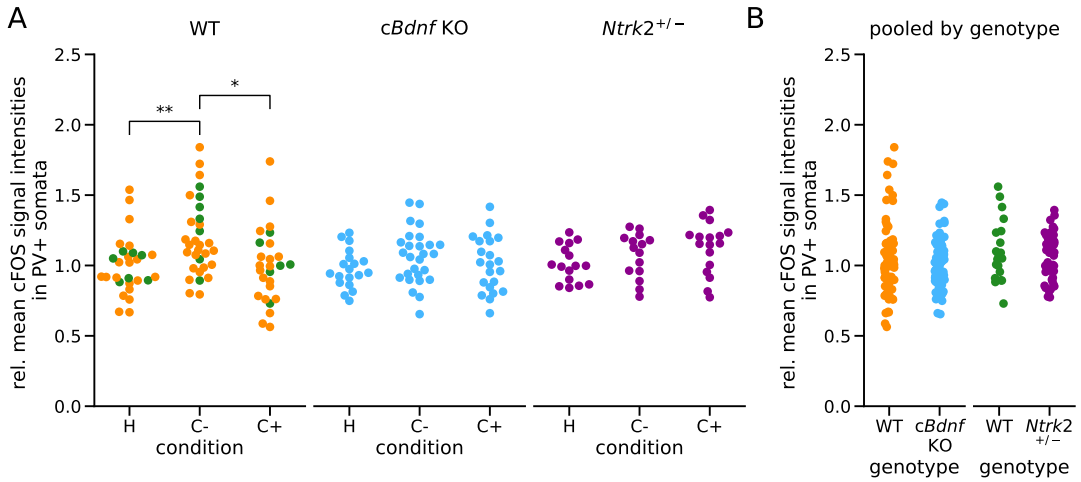


Figure S22: Mean cFOS-signal intensities of cFOS-positive nuclei in Parv-positive somata in CA3 in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice
A. Comparisons of the treatment conditions within each genotype. WT: F(2, 82)=5.95, P=0.004; c*Bdnf* KO: F(2, 65)=1.30, P=0.279; *Ntrk2*$^{+/-}$: F(2, 45)=2.15, P=0.129; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (*: p<0.05, **: p<0.01).
B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=2274.0, p=0.775; WT vs. *Ntrk2*$^{+/-}$: T(28.02)=0.57, p=0.571.
$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=4, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=28, $n_{(WT\ C-)}$=31, $n_{(WT\ C+)}$=26, $n_{(cBdnf\ KO\ H)}$=19, $n_{(cBdnf\ KO\ C-)}$=27, $n_{(cBdnf\ KO\ C+)}$=22, $n_{(Ntrk2^{+/-}\ H)}$=16, $n_{(Ntrk2^{+/-}\ C-)}$=15, $n_{(Ntrk2^{+/-}\ C+)}$=17.

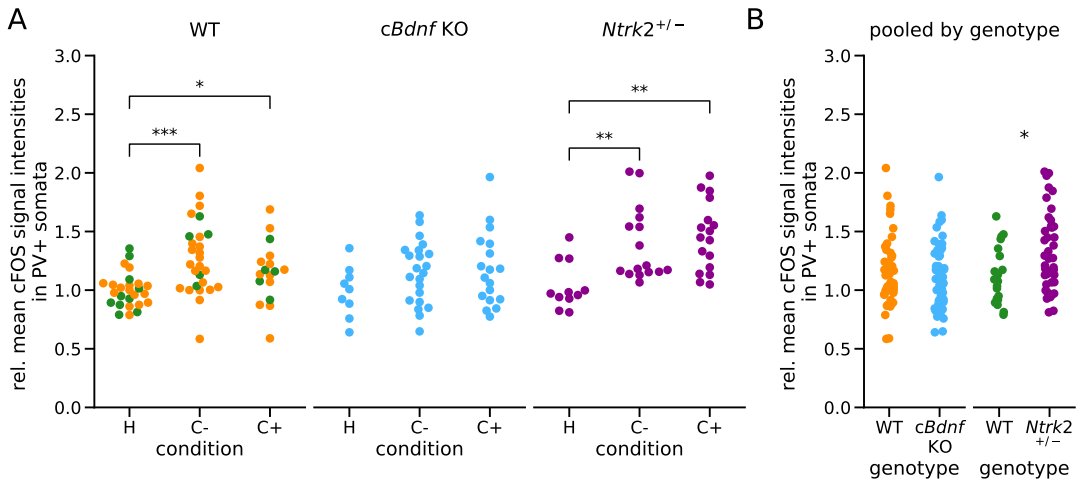


Figure S23: Mean cFOS-signal intensities of cFOS-positive nuclei in Parv-positive somata in CA1 in WT, c*Bdnf* KO, and *Ntrk2*$^{+/-}$ mice

A. Comparisons of the treatment conditions within each genotype. WT: H(2)=15.32, P<0.001; c*Bdnf* KO: F(2, 46)=1.48, P=0.238; *Ntrk2*$^{+/-}$: H(2)=13.62, P=0.001; post-hoc pairwise comparisons with Bonferroni correction for multiple comparisons (*: p<0.05, **: p<0.01, ***: p<0.001).

B. Data of all conditions was pooled within the indicated genotypes. WT vs. c*Bdnf* KO: U=1263.0, p=0.793; WT vs. *Ntrk2*$^{+/-}$: U=275.0, p=0.017.

$N_{(WT\ H)}$=8, $N_{(WT\ C-}$=7, $N_{(WT\ C+)}$=6, $N_{(cBdnf\ KO\ H)}$=3, $N_{(cBdnf\ KO\ C-)}$=5, $N_{(cBdnf\ KO\ C+)}$=4, $N_{(Ntrk2^{+/-}\ H)}$=3, $N_{(Ntrk2^{+/-}\ C-)}$=3, $N_{(Ntrk2^{+/-}\ C+)}$=3; $n_{(WT\ H)}$=26, $n_{(WT\ C-)}$=27, $n_{(WT\ C+)}$=17, $n_{(cBdnf\ KO\ H)}$=9, $n_{(cBdnf\ KO\ C-)}$=22, $n_{(cBdnf\ KO\ C+)}$=18, $n_{(Ntrk2^{+/-}\ H)}$=11, $n_{(Ntrk2^{+/-}\ C-)}$=16, $n_{(Ntrk2^{+/-}\ C+)}$=17.

I Abbreviations

Acq	acquisition session
BDNF	brain-derived neurotrophic factor (protein)
Bdnf	brain-derived neurotrophic factor (gene)
C-	context control (no shocks)
C+	context conditioned
CA	*Cornu ammonis*
c*Bdnf* KO	conditional *Bdnf* knock-out
CS	conditioned stimulus
DG	*Dentate gyrus*
dH_2O	distilled water
DL	deep learning
Ext	extinction session
H	homecage control
LD	light-dark cycle
LMT	large mossy fiber terminal
n	number of analyzed images
N	number of analyzed animals
Ntrk2	Neurotrophic tyrosine kinase receptor type 2 (gene)
Ntrk2$^{+/-}$	heterozygous TrkB knockout
PBS	phosphate buffered saline
PFA	paraformaldehyde
Parv	parvalbumin
Ret	retrieval session
ROI	region of interest
RT	room temperature
TrkB	tropomyosin-receptor kinase B (protein)
WT	wildtype

II List of Figures

III List of Tables

IV Acknowledgements

Mein größter Dank gebührt dem Betreuer meiner Doktorarbeit & Erstgutachter dieser Arbeit, Robert Blum. Nichts von dieser Arbeit wäre möglich gewesen, wenn Du mir nicht von Anfang an Deine volle Unterstützung hättest zuteil werden lassen und mir nicht die Freiheit eingeräumt hättest, diese Ideen zu entwickeln. Vielen Dank für die vielen intensiven Diskussionen, dafür dass Du mir immer wieder einen neuen Blickwinkel eröffnet hast und dafür, dass Du mir auch über das wissenschaftliche Denken und Arbeiten hinaus viele essentielle Eigenschaften näher gebracht hast, die einen guten Wissenschaftler und Menschen ausmachen. Vielen Dank!

Natürlich möchte ich mich auch bei den anderen Mitgliedern meines Thesiskomitees - Marta Andreatta, Philip Tovote und ganz besonders auch bei Chi Wang Ip als Zweitgutachter - herzlich bedanken. Neben den Komitee-Meetings hatte ich mit Euch allen viele anregende Diskussionen, die mir nicht nur beim Erstellen dieser Arbeit geholfen haben.

Für den schlussendlichen Erfolg dieses Projektes war jedoch auch die Kollaboration mit vielen Gruppen entscheidend, allen voran natürlich die enge Zusammenarbeit mit Matthias Griebel, Nikolai Stein und Christoph Flath. Trotz einigen Rückschlägen auf dem Weg haben wir es immer wieder geschafft in diesem interdisziplinären Team neue Wege und Lösungen zu finden. Ein großes Danke also auch an Euch! Gleiches gilt für alle anderen Kooperationspartner aus dem SFB TRR-58 unter dem Vorstand von Prof. Hans-Christian Pape. Für mich als jungen Wissenschaftler war dies wirklich ein prototypisches Beispiel dafür, wie ein CRC bzw. wie Community-basierte Wissenschaft im Allgemeinen funktionieren kann und sollte.

Ganz unmittelbar von diesem Projekt waren auch meine PhD-Kolleginnen betroffen, die jedoch zu keiner Zeit das manuelle Annotieren eines weiteren Bildes gescheut haben! Vielen Dank für Euern Einsatz und noch viel mehr dafür, dass Ihr die Zeit für mich im Institut immer zu einer Positiven gemacht habt & dass wir auch abseits der Arbeit gute Freunde geworden sind! Vielen Dank, liebe Annemarie, liebe Cora, liebe Corinna und liebe Manju - und ein ganz besonderes Danke an Euch zwei, liebe Rohini & liebe Nina! Ohne Euch hätte es sehr viel mehr triste Tage gegeben und ich habe jede Pause und jedes Stück Kuchen mit Euch genossen! Allen voran nochmal ein extra Danke an Dich, Nina, dafür dass Du Dir wirklich immer Zeit für mich genommen hast wenn selbst Nussschnecken nichts mehr ausrichten konnten. Ein liebes Dankeschön auch an Dich,

liebe Michi, für die unermüdliche Unterstützung & Hilfe im Labor!

Nicht weniger hierzu beigetragen haben aber auch meine Familie und meine engsten Freunde, auf deren Unterstützung und auf deren Verständnis ich viel zu oft zählen musste und immer wieder bauen konnte. Ihr habt mir den Halt, die Unterstützung und auch die Ablenkung geboten, ohne welche ich nie so weit gekommen wäre. Mama, Papa, Anna, Oma, Opa, Nelly, Paul, Nati, Bastian & ja auch euch: Jonas, Stephan, Jens, Basti & Hübi - VIELEN DANK!!!!

Zum Schluss aber geht mein herzlichster Dank an Dich, liebste Julia! Kein Mensch auf dieser Welt könnte mich mehr unterstützen als Du es tagtäglich tust, niemand besser für mich da sein wenn ich Zuspruch oder ein offenes Ohr brauche und mit niemand anders kann ich die wirklich schönen Dinge im Leben mehr genießen, als mit Dir. Vielen Dank, dass Du immer an meiner Seite bist :*