Machine assisted annotation in neuroanatomy

Kui Qian, Beth Friedman, David Kleinfeld, Yoav Freund

September 14, 2023

Abstract

One routine and necessary, yet time-consuming task in neuroanatomy is the annotation of labeled cells relative to the background. Currently, staining and imaging techniques enable the marking of specific cell groups with fluorescent dyes. Modern high throughput scanning microscopes allow high resolution multi-channel imaging of the sectioned brain. However, manual identification of labeled cells is prohibitively time consuming.

We present a methodology for developing digital assistants that significantly reduce the labor of the anatomist while improving the consistency of the annotation. Machine learning methods are combined with a rigorous way to measure the confidence of the predictions.

We compare the error rate of our method to the disagreement rate between human anatomists. This comparison demonstrates that our method can reduce the time to annotate as much as ten-fold without significantly increasing the error rate.

1 Main

We present an adaptive system designed to assist neuroanatomists with the task of labeling marked cells. The standard, fully manual approach –referred to here as "unaided"– provides anatomists with a subset of the sections (one in four, here), requiring them to detect as many marked cells as possible.

In our approach, a computer detector first identifies confident vs. unconfident detections. The anatomist performs Quality Control (QC) on the confident detections, which requires only a small sample. The process for the unconfident subset is similar to that of the unaided approach.

The reduction in anatomist labor therefore depends on the ratio between the confident and unconfident detections (as well as the accuracy of the confident detections).

We show that our method achieves significant reduction in labor and that its accuracy is similar to the level of agreement between different anatomists.

1.1 Significance for Neuroscience

How do we study the computational power of brains? Brains are composed of circuits at multiple scales of organization: from the microscopic-scale of connections between neurons (ranging from one to ten micrometers), to the mesoscopic-scale of collections of neurons with similar functionality (spanning hundreds of micrometers to millimeters). Molecular tagging of neuronal cell types by the expression of genetically encoded reporters and light-level imaging of the cells and tags, using both transmission and fluorescent microscopy, plays an essential role in this process. The resulting raw data files are extremely large, from 1 to 100 TB per rodent brain. The current means for quantification and spatial mapping of tagged neuronal populations in brain sections require labor and time intensive manual annotation by expert neuroanatomists. How can machine learning assist with this process and minimize the amount of manual work involved while maintaining accuracy and consistency?

We use mouse brains that are serially sectioned and counterstained with one fluorescent tag to reveal all brain cells. In addition, specific cell types are labeled with a second, different fluorescent

tag. This results in a dual tagging approach where all cells are labeled by an inexpensive counterstain and a subset of interest is labeled by a second experimentally incorporated fluorophore. This has the advantage of narrowing the population of false detections as labeled neurons are identified by combining two independent tags.

1.2 Significance to machine learning

The standard goal of a machine learning task is to reach a performance level close to or better than a human, thereby replacing the human. However, in practice, it is hard to compare the performance of a learned model to that of a human. This is because different humans often disagree with each other on a significant fraction of the labels. This phenomenon, termed the inter-rater disagreement rate [**?**], has been well studied in clinical neuroscience [**?**] but remains less explored in general neuroscience research. One contribution of this paper is an evaluation of inter-rater agreements in the context of marked cell detection.

Inter-rater disagreement rates quantify the average level of disagreement. However, the level of disagreement on each particular example can differ. We say that an example is "easy positive" if most graders label it positive and "easy negative" similarly for negative. We say that an example is "hard" if significant disagreement exists between labelers.

We allow our learned classifiers to output "sure positive", "sure negative" for confident detections and "unsure" for unconfident detections. The goal is to achieve high accuracy on the confident classifications while minimizing the fraction of unsure examples.

1.3 Comparison to other work

Accurate identification of neuronal cells is crucial for understanding the functional characteristics of distinct regions within the nervous system. However, selecting appropriate strategies for marking labeled cells can be challenging, especially for non-expert users, due to the diversity of microscopy modalities, staining methods, scales, and experimental conditions. In particular, fluorescent labeling comes with undesirable side effects, including photo-bleaching and photo-toxicity [18].

Current user-friendly bio-image analysis tools [13, 9, 19] typically require manual configuration and parameter adjustment, leading to time-consuming and expertise-dependent processes. They often rely on classical segmentation algorithms such as thresholding, which may struggle with the heterogeneity of biological samples and technical artifacts [5, 10, 17]. These limitations impede research progress and hinder the widespread adoption of imaging technologies in biological laboratories[2].

While existing user-friendly software, such as Ilastik [15] and ImageJ [6], offers machine learning-based solutions for biological image analysis, they require users to create experiment-specific models. This is still time-consuming, which involves preparing annotations, training models and configuring algorithms. Moreover, existing bio-image analysis packages [13, 4, 9, 1], whether commercial or open-source, are often designed for 2D images or smaller 3D volumes [16], rendering them inadequate for analyzing large-scale mouse brain data.

Machine learning methods, mostly based on Neural Network (NN), have been used to automate brain section analysis. However, as NN models are black boxes that define input-output relationships, the models provide no insight as to how decision are made, nor do NNs give a measure of confidence in their predictions. It is thus very hard for a neuroanatomist to understand why the NN made particular decisions. In this paper, we present an alternative approach with models that operate in a way more analogous to that of a neuroanatomist. These models generate outputs which the neuroanatomist can interpret and correct.

An important observation regarding locating labeled cells is that, while a typical section will contain some hard to identify features and locations, most of the identifications are relatively easy. This observation bolsters our approach, which uses the computer to assist rather than replace the human

anatomist. We use confidence rated detectors, which associate a confidence score with each detection. High confidence detections are labeled by the computer while low confidence detections are passed on to the human anatomist.

1.3.1 Quantifying work reduction

We compare the amount of work done by the anatomist when unaided to the amount of work when aided by confidence rated detections. When using an accurate and confident detection system, the work of the anatomist reduces to the following steps:

1. **Performing quality assurance on confident detections:** the anatomist receives a small sample of the confident detections and verifies that they are correct. The sample size depends on the desired accuracy.
2. **Searching for misses:** the anatomist looks for locations that were completely missed by the detector.
3. **Classifying the unconfident detections:** the anatomist labels all of the low confident predictions.

Steps 1 and 2 are based on samples and are therefore relatively light. Most of the work of the anatomist is in step 3. We call the ratio between the unconfident detections and the confident detections the "effective work ratio". When the effective work ratio is small, the savings in manual work is large.

2 The Problem

To demonstrate our methodology, we focus on a representative challenging cell detection problem. The input consists of two 3D images of a brain, using two florescent markers: a GFP marker that labels cells of interest and a Neurotrace marker which labels all neurons. While the GFP is the main identifier of the cells of interest, the Neurotrace channel is used to eliminate false detections. A typical false detection occurs when there is GFP signal, but no indication from Neurotrace that a neuron exists in the location.

Detection of marked cells is a demanding mental process which requires identifying neural shapes which vary greatly both within a brain and between brains, and integrating cues from GFP and Neurotrace.

Our estimate is that it takes a trained anatomist about 20 seconds to detect a single cell. The total number of labeled cells varies between 1000 and 20,000 cells, or 5-50 hours. As marking cells is tiring, anatomists typically assign no more than 2 hours per day, which means that manual detection can take weeks.

While a fraction of neurons are difficult to detect, a larger fraction are easy for both humans and machine. We develop algorithms that distinguish between easy and hard examples, label the easy detections and identify the hard examples for further human analysis.

We demonstrate that the examples identified as hard by the algorithm are also hard for humans. We show this by assigning two human labelers to label a sample of hard examples and counting the disagreements. Our results show a human disagreement rate of about 20%.

The level of agreement between human labelers is studied in inter-rater and intra-rater experiments [**?**]. Inter-rater agreement measures the rate of agreement between different human labelers. Intra-rater agreement measures the rate of agreement between labels chosen by the same labeler at different times. A common way to quantify the of the agreement between two raters is the Cohen's κ (kappa) coefficient [8] .[1]

[1] κ is computed from two more basic quantities: $0 \leq a \leq 1$ is the fraction of cases on which the two raters agree, and

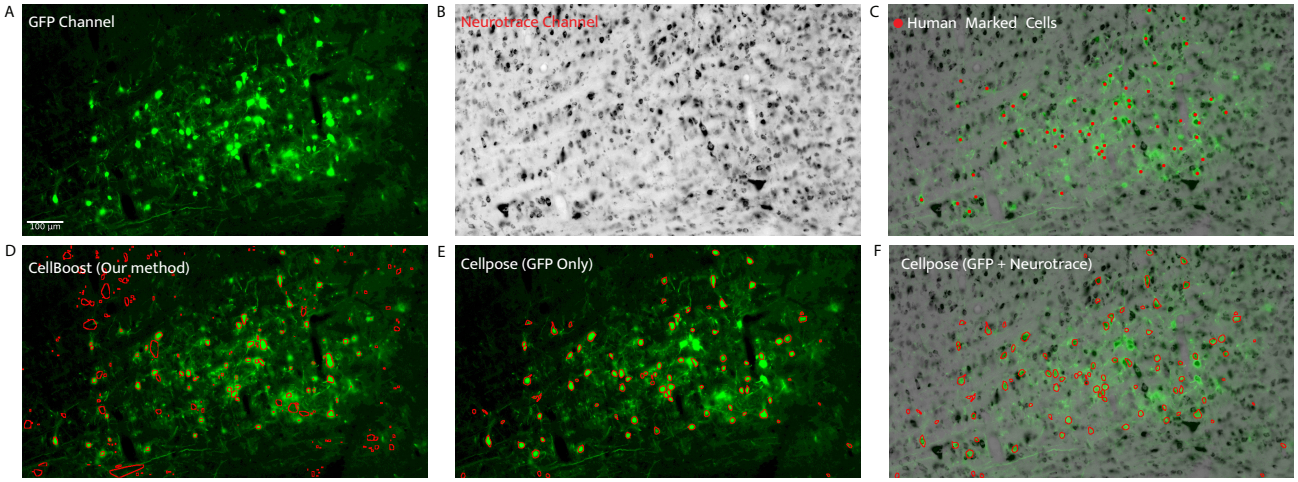

Figure 1: **An example of cell detection, as performed by a neuroanatomist and using different segmentation methods.** (A) A brain section image marked by GFP. (B) An image of the same region stained with Neurotrace. (C) A two channel image merging the GFP (green) and the Neurotrace (gray) images. The red dots correspond to annotations by the anatomist. Note that some of the large green blobs are not judged to be cells. (D) Results of the simple segmentation method used by CellBoost, applied only to the GFP channel. Note that all of the human-marked cells are identified, although there are also a large number of false detections. (E) and (F) Results of a deep learning segmentation method called Cellpose, using just the GFP channel and both channels, respectively. Cellpose similarly detects most of the marked cells, but also has high false detection rate, as do most segmentation methods. To remove the false detections reliably, we use a filter created using boosting.

3 Methods

Our method consists of three steps: a segmentation, feature calculation and a classification step.

3.1 Segmentation

We use adaptive thresholding [**?**], which offers the advantages of simplicity and speed. More complex segmentation methods have been devised; however, our experiments show that the error rates are not significantly lower, see Figure 1. Our approach is to use a simple and efficient segmentation step followed by a machine learning classification step for detecting the true positives.

Cellpose [**?**] is a popular cell segmentation method based on NN. While sophisticated, it yields an error rate similar to our simple method and is much more computationally complex (Figure 1). In addition, deep learning methods require annotating cell boundaries to retrain models, a process that demands extensive manual labor.

Our adaptive thresholding is conducted with the following steps on the GFP channel.

1. Convolve the image with a Gaussian filter. Since most of our cells are smaller than 200 pixels in length, we set the sigma value of the filter to be 100 and the kernel size to be 401×401 (according to the discrete Bessel approximation). This can enhance robustness against artifacts such as stained lines.

$0 \leq c \leq 1$ is the fraction of agreements that would occur by chance if the two raters are statistically independent. The definition of kappa is $\kappa = \frac{a-c}{1-c}$.

If $\kappa = 1$, the raters always agree; if $\kappa = 0$, the rate of agreement corresponds to chance; and if $\kappa < 0$, then the rate of agreement is lower than chance, i.e. the two raters tend to have different opinions. An interpretation of κ recommended by Cohen [8] is: $\kappa \leq 0$: no agreement, $0 < \kappa \leq 0.20$: none to slight agreement, $0.2 < \kappa \leq 0.40$: fair agreement, $0.4 < \kappa \leq 0.60$: moderate agreement, $0.6 < \kappa \leq 0.80$: substantial agreement, and $0.8 < \kappa \leq 1.00$: perfect agreement.

2. Subtract the original from the convolved image.
3. Threshold the difference image with a global constant C. C is set to be 2000 empirically for our datasets, whose data type is 16-bit unsigned integer.
4. Use cv2.connectedComponentsWithStats to find connected components as cell candidates in the thresholded image.

3.2 Feature calculation

The result of the segmentation step is a list of candidates, each defined by matching regions in the GFP and Neurotrace images. Each candidate is then mapped to 40 features to characterize its shape, both in the GFP and the Neurotrace images (see Table 1). These features have been handcrafted to be an over-complete representation of the candidates.

3.3 Classification

Our classification is based on boosted trees [**?**, 3] combined with bagging-type averaging [**?**] to partition examples into positive, negative for confident detections and unsure for unconfident detections.

Rather than generating a single boosted decision trees, we generate an ensemble of boosted trees by using random seeds to generate a "composite detector" (Figure 2). Two quantities are computed based on the composite detector: the mean of the scores and the standard deviation of the scores. Both are indicative of the prediction confidence. The mean is indicative of the prediction margin [**?**], while the standard deviation is inversely proportional to the stability across the ensemble.

We use the average to partition easy from hard examples and the standard deviation to verify that low average candidates are unstable.

3.4 Human machine cooperation

Our goal in this work is to devise a machine learning algorithm and a work-flow that will minimize the amount of human work, maximize accuracy, and have a comparable level of "unsure" predictions to the level of disagreement between humans.

The system we designed uses an adaptive-threshold segmentation method, which extracts on the order of 100,000 "cell-candidates". We train the composite detector to divide these into three groups: *positive/unsure/negative*. These are designed so that the "positive" and the "negative" have accuracy comparable to that of humans and therefore can be labeled automatically. The candidates labeled "unsure" are too hard to call and are left for humans to label. Human labeling is thereby reduced to the following two tasks, which are done by two independent humans:

1. **Quality control (QC):** Labeling a sample of the 250 "positive" and 250 "unsure" to verify the accuracy of the "positive" and "unsure".
2. **Unaided labeling:** Labeling five sections in an **unaided** mode to estimate the false negatives of the detector.

After completing these tasks, the user has a good estimate of the false positive and false negative rates. If these are sufficiently low, the task is done and the positive detections are used for the neuroanatomical analysis.

If the QC performance of the composite detector on a brain is insufficient, then the composite detector is retrained by using the QC examples together with examples that are high confidence positive and high confidence negative. It is not surprising that this improves performance on the new brain. More significant is that the performance on older brains shows an improvement as well. After retraining, a second round QC

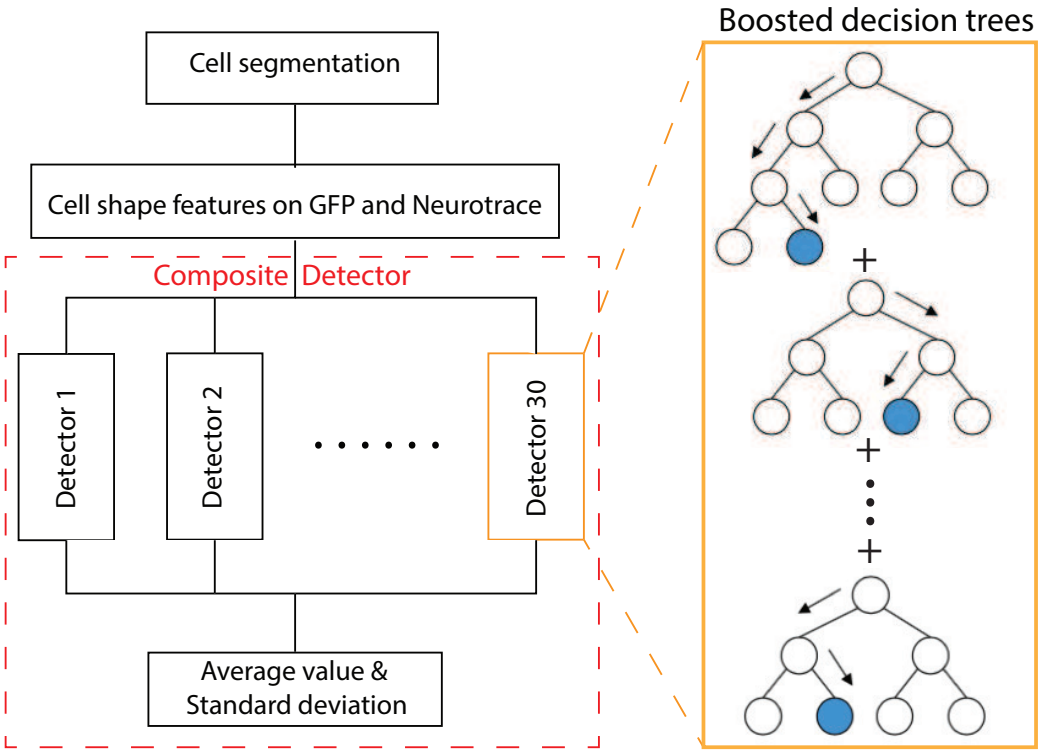

Figure 2: **Machine Learning architecture** The system consists of four stages. The first stage is cell segmentation, which takes as input the GFP image and outputs a set of detection candidates. In the second stage, cell shape features are computed for each candidate. The third step consists of 30 scoring functions, each of which is a boosted decision trees. The structure of a boosted decision tree is shown on the right. The fourth stage, called the composite detector, computes the mean and standard deviation of the 30 scores generated in stage 3. This mean and standard deviation are used to classify candidates into positive, negative and unsure.

4 Results

4.1 Segmentation

As can be seen in Figure 1, the simple adaptive threshold we use in CellBoost performs similarly to established methods such as Cellpose. When the parameters are set to ensure low false negatives, the number of false positives is similar across methods. Notably, the deep learning method overlooks some true positive cases when two or three cells overlap or are closely situated.

4.2 Machine Learning Performance

Figure 3 summarizes the performance of the classification stage.

We make the following general observations:

- Most of the segmented candidates receive negative scores. This reflects the fact that most cell candidates correspond to small regions that are not cells.
- The QC labeling shows a low error rate on the confident positive and the confident negative.
- The unaided labeling shows that most of the unaided labels are classified as positive. A small fraction are unsure and about 5% are missed.
- The relationship between the average and standard deviation (STD) of the detector scores shows that, in general, the margin and the standard deviation are closely related. But when many

examples are labeled, as in brain1, there are many unstable negative examples that are still recognized as confidently negative.

The goal of retraining is to adapt a detector to the properties of a specific dataset which was not used in the original training. We make the following observations regarding retraining:

- **Brain2:** This brain was analyzed using detector data from other brains resulting in a well trained detector (G4). To improve the performance, a new detector (G5) was trained on the existing dataset with the addition of 500 cell QC from Brain2. Comparing **Panel A and B**, we see that a significant improvement in performance was achieved using a small number of training examples.
- **Brain1:** This brain was analyzed using an early detector (G3) and the performance was compared to that of a later detector (G6) which was trained on other brains. In this case, we observe that the performance on Brain1 did not degrade by adding training data from other brains. This gives evidence that retraining a detector on new brains does not degrade the performance on old brains.

4.2.1 QC results

Our composite detector begins its training with **Brain0**, which uses masseter injection to trace premotor neurons. Subsequently, it undergoes testing and retraining on **Brain1** (whisker pad injection), **Brain2** (whisker pad injection), and **Brain3** (tongue injection) in that order. The quantitative results of QC and unaided tasks are:

- **QC:** The accuracy rates of positive detections are 95.60%, 87.20% and 76.80% respectively. The human disagreement rates of unsure detections are 9.20%, 16.80%, and 24.40% respectively.
- **Unaided:** The false negative rates are 7.80%, 8.30%, 9.02% respectively. The false positive rates are close to 0 due to the large amount of negative detections.

Our detector achieves an excellent performance on **Brain1**. The error rates of positive detection appear proportional to the human disagreement rates of unsure detections.

4.3 Feature selection

Figure 4 depicts the relative importance of the features that are used in the composite detectors. The importance is defined as the total gain across all splits the feature is used in a decision tree model. The main observation is that the the most important features are some image moments of the GFP channel and the cross-correlation coefficient (defined in Table 1) calculated from the GFP channel as well as from the Neurotrace channel. As anticipated, features from the Neurotrace channel play a pivotal role. It is out of expectation that central moments and other invariant moments appear to be less important than raw moments. This indicates that scale and rotation might significantly influence the detection process.

Panel B and C give detailed scatter plots for a sample of cell shapes comparing GFP across-correlation coefficient to Neurotrace across-correlation coefficient (Panel B) and GFP across-correlation coefficient to m_{11} (Panel C). These scatterplots justify the large weights of the corresponding features in Panel A.

4.4 Human performance

The QC process is depicted in Figure 5. Most premotor neurons are situated in the brainstem. Panel A identifies two regions in a brainstem section with clear and robust fluorescent signals. Correspondingly,

Panel B shows that our positive detections are primarily localized in these two regions. Conversely, negative detections are observed throughout the entire brain, consistent with the fact that most cell candidates are not real cells. An area is magnified to show typical premotor neurons, characterized by apparent green-stained shapes accompanied by distinct black cell bodies. As shown in Panel C, our detector identifies these neurons and categorizes other minor stained objects as negative. Three of these detections were selected as samples for QC and there was a consensus among our labelers regarding the detection outcomes.

Figure 6 describes the disagreements between two human labelers. For one QC sample, a label could either be positive or negative, yielding four potential outcomes between two human labelers. We separate GFP and Neurotrace images to explore characteristics of these outcomes.

- **Panel A** presents an "easy positive" example labeled as positive by both labelers. This is a typical premotor neuron, evidenced by a green-stained shape with clear boundaries and a distinct gray cell body. Notably, our detector awards it a confident positive score.
- **Panel B** shows an "easy negative" example labeled as negative by both labelers. Despite the presence of a green-stained shape in the GFP channel, the corresponding location in the Neurotrace channel offers no indication that a neuron exists. In response, our detector assigns this cell candidate a confident negative score.
- **Panel C and D** display two hard cases, with different labels from two human annotators. Both cell candidates exhibit green-stained shapes with blurry boundaries, while vague signals can be found in the Neurotrace channel. These traits lead to disagreement between labelers and also an unsure detection by our detector.

Neurotrace Nissl staining is a very useful histological method to facilitate the identification of cell populations. Our anatomists predominantly utilize the Neurotrace channel to verify the existence of premotor neurons. Therefore, vague signals in this channel can pose challenges in decision-making. If labelers establish individualized criteria, their discrepancies might stem from systematic errors. Examining Panel E, F and G, we observe that nearly 15% to 20% of QC samples for each brain display divergent labels from the two labelers. Strikingly, of the total disagreement cases, 83 out of 86 in Panel F and 91 out of 97 in Panel G come from the same category, labeled positive by Human1 while negative by Human2. This indicates that some disagreements may be attributed to a systematic bias, with image quality playing a pivotal role in the labelers' judgment.

5 Discussion

A challenge for extant methods is the identification of brain neurons. Classically, this relies on the use of a stain, i.e., a molecule that labels a restricted part of brain cells and allows both individual cells and clusters of similar cells to be identified. Currently, human scanning is used to find the best match to the expectation of a neuron. This match-to-sample approach does not readily support assessment of observational differences between laboratories nor provide a means for assessing and compensating for common systematic errors. An extended machine learning (ML) approach would support a mechanism for ML "self-annotation" of cells in sections. The output tag of a cell would be readily proofread by human observation. Importantly, the curated set of output tags could serve to define the coordinates of all labeled cells. This annotation allows tagged cells to be placed into a common reference space, such as the Allen Atlas, so that the spatial distributions of functionally identified cells can be compared.

The central motivation of the work presented here is to develop tools to leverage human expertise rather than attempt to replace it; doing so demands the use of ML methods that support meaningful confidence levels and can provide explanations for these confident predictions. By training an ensemble

of boosted decision tree models, we simulate the cell-labeling process undertaken by multiple human annotators. The standard deviation among these models indicates the level of their discrepancies. Consequently, we establish a prediction margin based on the ensemble's mean and standard deviation, partitioning confident from unconfident detections. The unconfident detections are demonstrated to align with instances that humans find challenging and frequently disagree upon.

Many existing methods for cell detection in whole brain images demonstrate their accuracy by matching their detection counts with manual counts, without examining the precision of individual cell detection. However, a matching count does not necessarily imply accuracy of individual cell detection. We show that the results of confident positive detections compare well to human annotations. Moreover, the error rate of positive detections scales proportional to the human disagreement rate of unconfident detections. This suggests potential systematic biases such as image artifacts, especially in the Neurotrace channel. These biases contribute to detection errors, unconfident detections and human discrepancies. On the flip side, we can evaluate the image quality by checking the number of unconfident detections without any QC. In addition, our results also show the importance of retraining the models for new datasets. We used three types of injections for brains in the experiment: masseter injection, tongue injection and whisker pad injection. Each type traces premotor neurons in brain regions that specifically control the associated muscles. By incorporating a limited number of QC examples for retraining, our system effectively achieves accurate detection of premotor neurons across various brain areas.

Another motivation of our work is to significantly reduce labor work. Traditional neural network approaches often demand labor-intensive generation of training data. In contrast, our method overcomes this by requiring only a binary label for each cell candidate, rather than an outline of the cell shape. When a new brain is introduced, the total time for QC and unaided labeling is less than three hours. Thus, processing an entire brain (300 GB of images) within a single day becomes a feasible goal.

In the future, we intend to integrate our models with a brain atlas to enumerate cells within specific brain structures. This can be achieved either by mapping cells into an atlas space or by overlaying brain structure outlines from the atlas onto whole-brain images. Such integration would enable a detailed analysis of the spatial distribution of functionally identified cells. Additionally, our approach can serve as a foundational step for segmentation. In our pipeline, a subsequent classification step after segmentation acts as a filter to enhance performance. In a reciprocal manner, the generated detections can facilitate the creation of masks for training segmentation models. By improving segmentation, our detector's performance can be further optimized.

This software is fully open-source, along with our custom visualization and annotation tools built upon Neuroglancer, a WebGL-based software.

Cell detector: https://github.com/ActiveBrainAtlas2/cell_extractor

Neuroglancer: https://github.com/ActiveBrainAtlas2/neuroglancer

References

[1] Berg, S., Kutra, D., Kroeger, T., Straehle, C.N., Kausler, B.X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M., et al.: Ilastik: interactive machine learning for (bio) image analysis. Nature methods **16**(12), 1226–1232 (2019)

[2] Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., et al.: Nucleus segmentation across imaging experiments: the 2018 data science bowl. Nature methods **16**(12), 1247–1253 (2019)

[3] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)

[4] De Chaumont, F., Dallongeville, S., Chenouard, N., Hervé, N., Pop, S., Provoost, T., Meas-Yedid, V., Pankajakshan, P., Lecomte, T., Le Montagner, Y., et al.: Icy: an open bioimage informatics platform for extended reproducible research. Nature methods **9**(7), 690–696 (2012)

[5] Dima, A.A., Elliott, J.T., Filliben, J.J., Halter, M., Peskin, A., Bernal, J., Kociolek, M., Brady, M.C., Tang, H.C., Plant, A.L.: Comparison of segmentation algorithms for fluorescence microscopy images of cells. Cytometry Part A **79**(7), 545–559 (2011)

[6] Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al.: U-net: deep learning for cell counting, detection, and morphometry. Nature methods **16**(1), 67–70 (2019)

[7] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55**(1), 119–139 (Aug 1997)

[8] McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica **22**(3), 276–282 (2012)

[9] McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B.A., Karhohs, K.W., Doan, M., Ding, L., Rafelski, S.M., Thirstrup, D., et al.: Cellprofiler 3.0: Next-generation image processing for biology. PLoS biology **16**(7), e2005970 (2018)

[10] Meijering, E.: Cell segmentation: 50 years down the road [life sciences]. IEEE signal processing magazine **29**(5), 140–145 (2012)

[11] Schapire, R.E., Freund, Y.: Boosting: Foundations and algorithms. Emerald Group Publishing Limited (2013)

[12] Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S.: Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics **26**(5), 1651–1686 (October 1998)

[13] Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al.: Fiji: an open-source platform for biological-image analysis. Nature methods **9**(7), 676–682 (2012)

[14] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al.: Mastering the game of go without human knowledge. nature **550**(7676), 354–359 (2017)

[15] Sommer, C., Straehle, C., Koethe, U., Hamprecht, F.A.: Ilastik: Interactive learning and segmentation toolkit. In: 2011 IEEE international symposium on biomedical imaging: From nano to macro. pp. 230–233. IEEE (2011)

[16] Tyson, A.L., Rousseau, C.V., Niedworok, C.J., Keshavarzi, S., Tsitoura, C., Cossell, L., Strom, M., Margrie, T.W.: A deep learning algorithm for 3d cell detection in whole mouse brain image datasets. PLoS computational biology **17**(5), e1009074 (2021)

[17] Ulman, V., Maška, M., Magnusson, K.E., Ronneberger, O., Haubold, C., Harder, N., Matula, P., Matula, P., Svoboda, D., Radojevic, M., et al.: An objective comparison of cell-tracking algorithms. Nature methods **14**(12), 1141–1152 (2017)

[18] Wang, Z., Millet, L., Chan, V., Ding, H., Gillette, M.U., Bashir, R., Popescu, G.: Label-free intracellular transport measured by spatial light interference microscopy. Journal of Biomedical Optics **16**(2), 026019–026019 (2011)

[19] Wiesmann, V., Franz, D., Held, C., Münzenmayer, C., Palmisano, R., Wittenberg, T.: Review of free software tools for image analysis of fluorescence cell micrographs. Journal of microscopy **257**(1), 39–53 (2015)

A Methods details

A.1 Boosting and sparse representations

An important part of the design of any learning algorithm is finding a representation of input feature vectors that captures the aspects that are most relevant for the classification task. In some situations deep neural networks can find internal representations autonomously, without human intervention. However, a close look at the design of alpha-Go [14] reveals the high level of human expertise was used to design the features used by the neural network.

Boosting [7, 11] is another popular learning algorithm which combines a large number of so-called "weak" rules to construct a single "strong" rule. Here we follow an approach to feature detection for boosting that can be described as the kitchen sink approach. This approach starts by the human constructing a very large number of candidate rules. The boosting algorithm performs both feature selection i.e. finding the rules that provide significant information about the label, as well as feature weighting and combination, i.e. finding how to combine the informative features to predict the label.[2]

In general, increasing the number of features or rules increases the danger of over-fitting. However, as shown in [12], the number of features has only a small influence on overfitting. Rather, it was shown that the distribution of the large normalized margin guarantees low overfitting even if the number of features goes to infinity.

[2]Popular boosting software, such as XGBoost and LightGBM, use decision trees to combine the features.

Table 1: Feature importance in detectors of Composite Detector G6

Rank	Feature name	Definition	Importance in detectors (Median)
1	m_{11}	a image moment: $\sum_x \sum_y xyI(x,y)$	49160.54
2	$Xcorr\_GFP$	mean of cross-correlation to the average shape of positive cells in the GFP channel	6735.25
3	m_{10}	a image moment: $\sum_x \sum_y xI(x,y)$	3352.64
4	m_{12}	a image moment: $\sum_x \sum_y xy^2I(x,y)$	2981.99
5	m_{01}	a image moment: $\sum_x \sum_y yI(x,y)$	2922.80
6	m_{21}	a image moment: $\sum_x \sum_y x^2yI(x,y)$	2087.11
7	$energy\_Ntb$	integral of squared image gradients in the Neurotrace channel	1937.21
8	$Xcorr\_Ntb$	mean of cross-correlation to the average shape of positive cells in the Neurotrace channel	1437.84
9	mu_{02}	a image moment: $\sum_x \sum_y (y-\bar{y})^2I(x,y)$	1279.08
10	m_{20}	a image moment: $\sum_x \sum_y x^2I(x,y)$	1249.88
11	$energy\_GFP$	integral of squared image gradients in the GFP channel	1147.04
12	h_0	the first Hu moment	841.12
13	Contrast_Ntb	$\frac{\bar{I}_{in}(Neurotrace)-\bar{I}_{all}(Neurotrace)}{\bar{I}_{in}(Neurotrace)+\bar{I}_{all}(Neurotrace)}$	746.79
14	Contrast_GFP	$\frac{\bar{I}_{in}(GFP)-\bar{I}_{all}(GFP)}{\bar{I}_{in}(GFP)+\bar{I}_{all}(GFP)}$	731.51
15	h_1	the second Hu moment	610.77
16	nu_{20}	a image moment: mu_{20}/m_{00}^2	355.44
17	mu_{11}	a image moment: $\sum_x \sum_y (x-\bar{x})(y-\bar{y})I(x,y)$	346.37
18	height	height of a candidate	312.76
19	nu_{11}	a image moment: mu_{11}/m_{00}^2	296.03
20	mu_{20}	a image moment: $\sum_x \sum_y (x-\bar{x})^2I(x,y)$	260.27
21	nu_{30}	a image moment: $mu_{30}/m_{00}^{5/2}$	258.57
22	area	area of a candidate	256.55
23	mu_{03}	a image moment: $\sum_x \sum_y (y-\bar{y})^3I(x,y)$	239.08
24	h_2	the third Hu moment	232.17
25	mu_{21}	a image moment: $\sum_x \sum_y (x-\bar{x})^2(y-\bar{y})I(x,y)$	229.82
26	h_3	the forth Hu moment	225.75
27	nu_{03}	a image moment: $mu_{03}/m_{00}^{5/2}$	225.30
28	h_5	the sixth Hu moment	222.02
29	h_6	the seventh Hu moment	214.95
30	m_{03}	a image moment: $\sum_x \sum_y y^3I(x,y)$	213.31
31	mu_{12}	a image moment: $\sum_x \sum_y (x-\bar{x})(y-\bar{y})^2I(x,y)$	208.33
32	h_4	the fifth Hu moment	203.76
33	nu_{02}	a image moment: mu_{02}/m_{00}^2	202.05
34	mu_{30}	a image moment: $\sum_x \sum_y (x-\bar{x})^3I(x,y)$	200.54
35	nu_{21}	a image moment: $mu_{21}/m_{00}^{5/2}$	198.11
36	nu_{12}	a image moment: $mu_{12}/m_{00}^{5/2}$	188.40
37	m_{30}	a image moment: $\sum_x \sum_y x^3I(x,y)$	175.03
38	width	width of a candidate	125.28
39	m_{02}	a image moment: $\sum_x \sum_y y^2I(x,y)$	99.96
40	m_{00}	a image moment: $\sum_x \sum_y I(x,y)$	10.83

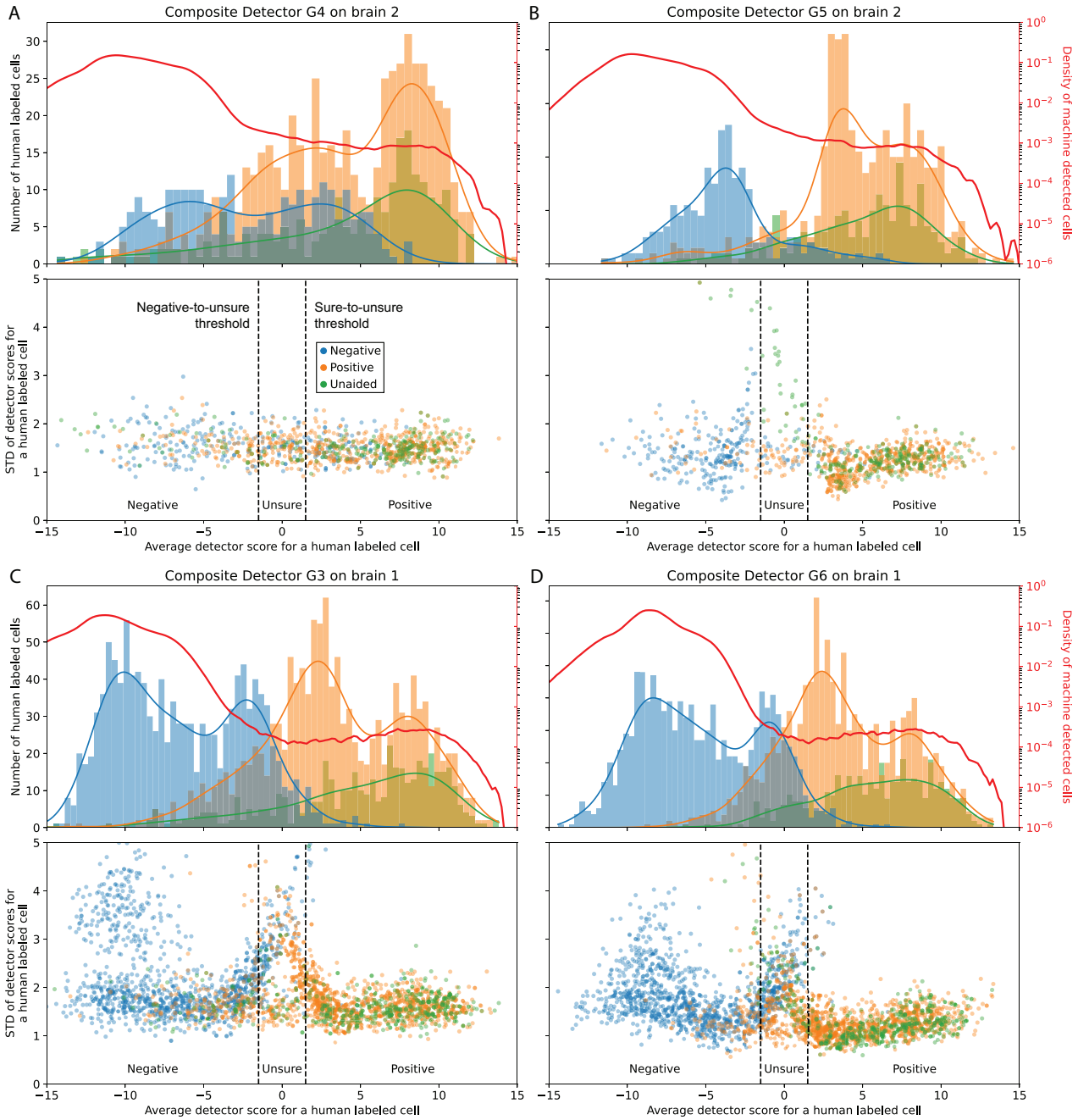

Figure 3: **Performance of composite detector:** Each panel in this figure summarizes the performance of a single composite detector. The horizontal axis corresponds to the average score and is divided into three regions: Positive, Unsure and Negative. The vertical axis of the scatter plot corresponds to the standard deviation of detector scores. The dots correspond to human labeled examples which labeled by three colors: blue and orange dots corresponds to cells that have been labeled positive and negative in QC, while green dots correspond to cells that have been detected by a human unaided. In the histogram plot, we show the quantity distribution of these human labeled examples. Besides, the bold red curve and the corresponding right axis describe the density of the scores of the candidates. Note there are two orders of magnitude difference between the density of the negative and the positive candidates. In (A,B) we observe the improvement in performance on brain 2 before and after training on brain 2. In (C,D) we observe the improvement in performance on brain 1 achieved by training on brains other than brain 1.

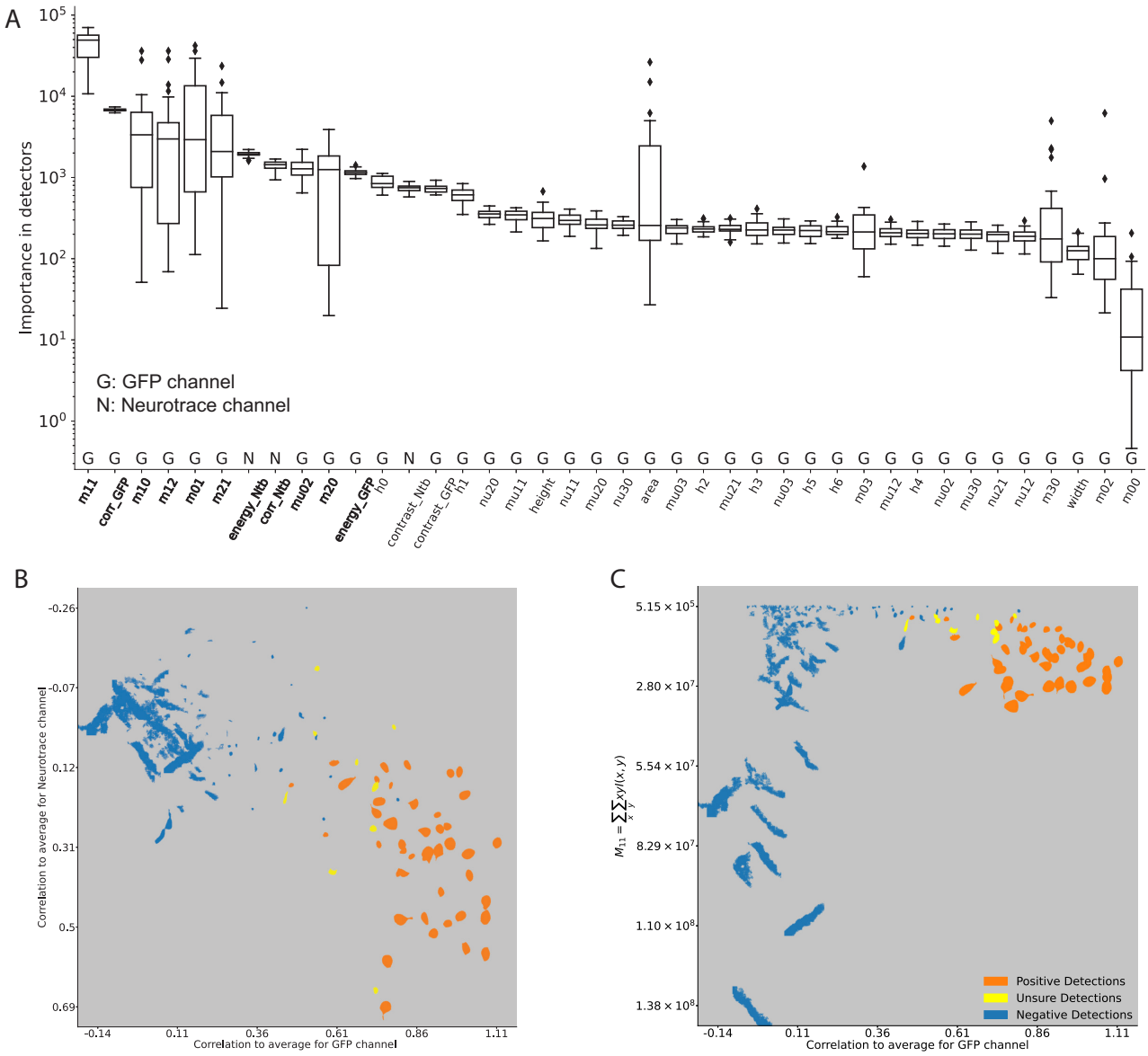

Figure 4: **The important shape features.** The boosting algorithm uses as input all 40 features described in Table 1. However, some of the features carry more weight than others. (A) Display of feature importance, in decreasing order of weight. A box plot is used to show the distribution of the weight across the 30 bagged copies of the detector that are incorporated into the combined detector. (B,C) Scatter plots of the real shapes of the GFP images of the cells distributed according to two high-weight features. In (B) the horizontal distribution is according to the across-correlation to the average shape in the GFP channel, while the vertical is the same for the Neurotrace Channel. In (C) the horizontal is the same as (B) while the vertical corresponds to moment 11 (m_{11}).

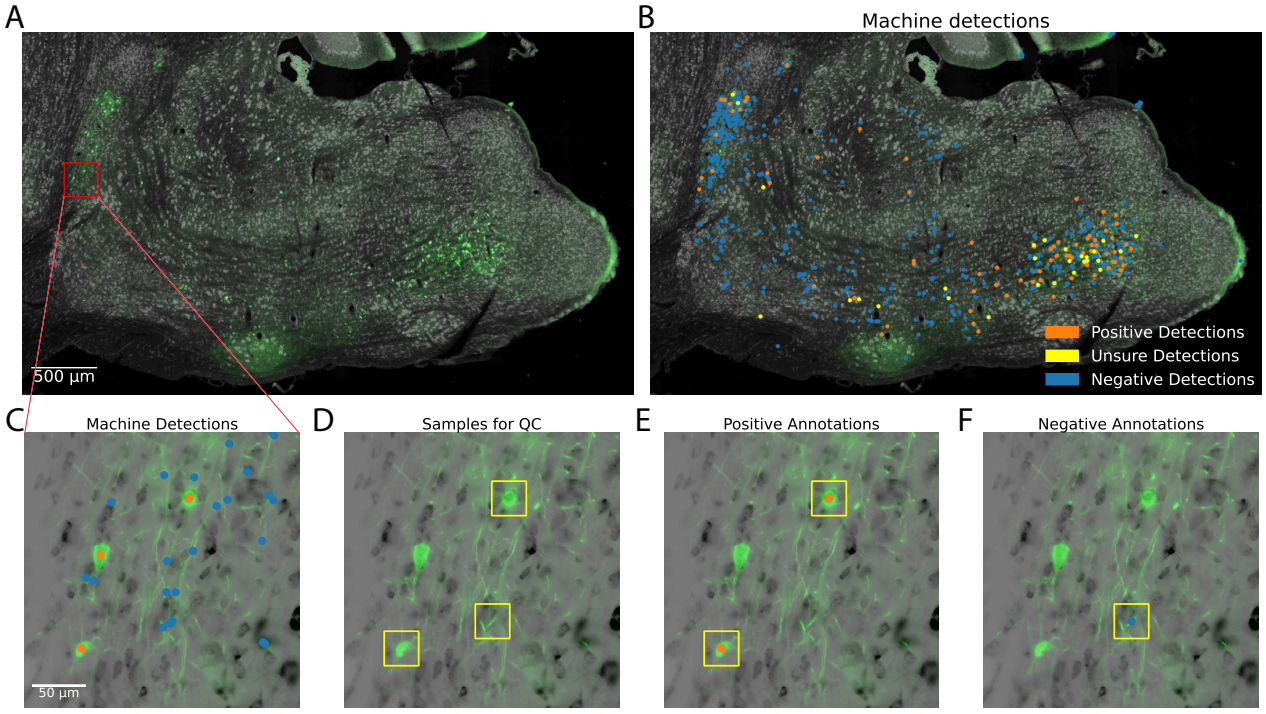

Figure 5: **Overview of the performance.** (A) A section of the brainstem stained with GFP and Neurotrace. (B) Detections classified as positive, unsure, and negative. This illustrates the spatial distribution of machine detections. (C) A magnified area highlighting machine detections of all confidence levels. (D) Randomly selected samples from detections for QC , with positive (E) and negative (F) annotations provided for clarity.

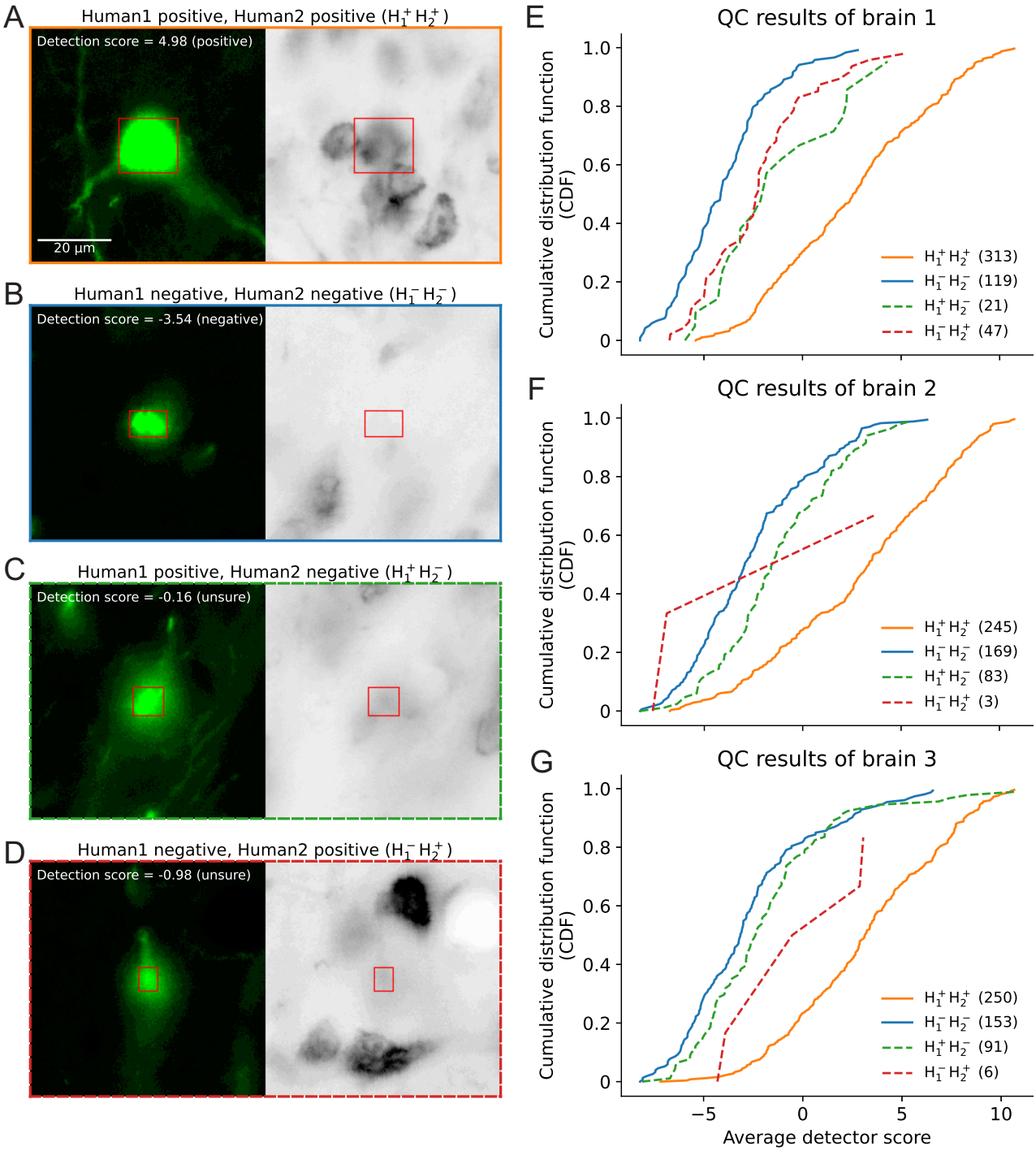

Figure 6: **Human performance** (A,B,C,D) Depiction of the four possible outcomes between two human labelers, with the GFP and Neurotrace channels displayed separately to underscore their roles in recognition. (E,F,G) Performance summary of the two human labelers across QC tests for three distinct brains, with CDFs plotted for the four potential outcomes based on average detector scores. Notably, there's an approximate 15% to 20% disagreement rate between human labelers.