ActSort: An active-learning accelerated cell sorting algorithm for large-scale calcium imaging datasets

Hakki Orhun Akengin[*], **Mehmet Anil Aslihak**[*], **Yiqi Jiang**[*]
CNC Program, Stanford University, Stanford, CA

Yang Li, Oscar Hernandez
CNC Program, Stanford University, Stanford, CA

Hakan Inan
Meta, Menlo Park, CA

Christopher Miranda[†], **Marta Blanco-Pozo**[†]
CNC Program, Stanford University, Stanford, CA

Fatih Dinc[†]
Department of Applied Physics, Stanford University, Stanford, CA

Mark J. Schnitzer[†]
Howard Hughes Medical Institute, CNC Program, Stanford University, Stanford, CA

Abstract

Due to rapid progress in optical imaging technologies, contemporary neural calcium imaging studies can monitor the dynamics of 10,000 or more neurons at once in the brains of awake behaving mammals. After automated extraction of the neurons' putative locations, a typical experiment involves extensive human labor to cull false-positive cells from the data, a process called *cell sorting*. Efforts to automate cell sorting via the use of trained models either employ pre-trained, suboptimal classifiers or require reduced but still substantial human labor to train dataset-specific classifiers. In this workshop paper, we introduce an active-learning accelerated cell-sorting paradigm, termed ActSort, which establishes an online feedback loop between the human annotator and the cell classifier. To test this paradigm, we designed a benchmark by curating large-scale calcium imaging datasets from 5 mice, with approximately 40,000 cell candidates in total. Each movie was annotated by 4 (out of 6 total) human annotators, yielding about 160,000 total annotations. With this approach, we tested two active learning strategies, discriminative active learning (DAL) and confidence-based active learning (CAL). To create a baseline representing the traditional strategy, we performed random and first-to-last annotations, in which cells are annotated in either a random order or the order they are received from the cell-extraction algorithm. Our analysis revealed that, even when using the active learning-derived results of $< 5\%$ of the human-annotated cells, CAL surpassed human performance levels in both precision and recall. In comparison, the first-to-last strategy required 80% of the cells to be annotated to achieve the same mark. By decreasing the human labor needed from hours to minutes while also enabling more accurate predictions than a typical human annotator, ActSort overcomes a bottleneck in neuroscience research and enables rapid pre-processing of large-scale brain-imaging datasets.

Keywords: Active learning, calcium imaging, cell sorting, human annotation

∗. These authors contributed equally to this work.
†. This group of authors supervised this work.

1. Introduction

Over the last two decades, systems neuroscience research has been revolutionized by the large-scale fluorescence calcium imaging techniques for tracking the dynamics of large populations of individual neurons in awake behaving animals (Grienberger and Konnerth (2012); Packer et al. (2015); Sofroniew et al. (2016); Kim and Schnitzer (2022)). With the rapid, roughly an order of magnitude per decade, increase in the number of recorded neurons, identification of neural footprints and their activity traces from brain imaging movies required automation, giving birth to so-called cell extraction algorithms (Pachitariu et al. (2016); Giovannucci et al. (2019); Inan et al. (2021)). Though they opened the doors for probing brain-wide neural computations by extracting activities of thousands of cells, cell extraction algorithms can occasionally output non-neuronal sources of perceived Ca^{2+} activity. Hence, to prevent spurious, non-neuronal, signals from leading to incorrect biological conclusions (See Gauthier et al. (2022); Inan et al. (2021) for examples of altered biological results due to incorrect pre-processing of calcium imaging movies), neuroscience experiments regularly require extra steps to refine and prune the cell extraction results (Pachitariu et al. (2016)).

To prune non-neuronal candidates from the cell extraction output, a process called *cell sorting*, experimentalists either spend hours of research time annotating all the cell candidates via custom software (See Corder et al. (2019)) or utilize "cell classifiers" with occasional fine-tuning through partial annotation, *i.e.*, offline human feedback (Giovannucci et al. (2019); Pachitariu et al. (2016), also see CLEAN in Ciatah pipeline: Corder et al. (2019)). While the former approach is no longer feasible on the large datasets of today, which will eventually become mainstream of systems neuroscience research in the future (Kim and Schnitzer (2022)), the latter still requires a substantial amount of human labor to provide accurate results as we show in this work. That being said, utilizing expert neuroscientists to annotate cells for several hours, only to obtain one accurately annotated imaging session, has become unrealistic in this ever-improving experimental landscape. Hence, there is an urgent need for a principled strategy that leads to minimal human labor from the experimentalist's side without sacrificing cell sorting accuracy. Instead of developing heuristic solutions that are impervious to human errors and aim to resolve challenges associated with narrow sets of datasets (Giovannucci et al. (2019); Pachitariu et al. (2016)), it may be past time that we address this challenge in a scalable, generalizable, and robust manner, *e.g.*, by promoting online human feedback.

In this work, we propose a well-studied approach from machine learning, called "active learning" (see Appendix B), as a solution to address broad challenges surrounding the cell sorting problem. By leveraging online human feedback, active learning provides a path forward for developing scalable solutions *and* testing their accuracy online, limiting the arbitrariness of the human annotation and decreasing the required labor. We introduce ActSort: an active-learning accelerated cell sorting algorithm for large-scale calcium imaging datasets. Our novel contributions include the introduction of the active learning paradigm to the calcium imaging literature via ActSort (Fig. 1), the engineering of unique features to aid classifier training (Fig. S1), creating a large-scale annotation benchmark for training and testing ActSort on large-scale calcium imaging datasets (Figs. S1, S3, S4, and Table S1), and decreasing the required human labor to annotating 5% of the dataset, which enables ActSort to outperform human annotators (Fig. 2).

2. Results

2.1 A new calcium imaging benchmark for cell sorting

Though there are publicly available calcium imaging movies with annotated cells, see for example the Neurofinder benchmark (Berens et al., 2017), our primary motivation has been to develop a cell sorting algorithm that is benchmarked on the contemporary large datasets that contain up to 10,000 simultaneously recorded neurons across multiple brain regions in a single imaging session. To the best of our knowledge, no such dataset is publicly available.

To bridge the gap, we performed cell extraction on three one-photon calcium imaging movies (from three distinct mice), spanning the half-hemisphere through a 7mm window (Figs. 1, 2, S1, S2, and Table S1), one movie from Ebrahimi et al. (2022) spanning eight neocortical regions (Fig. S3), and one two-photon imaging movie capturing layer 2/3 cortical pyramidal neurons (Fig. S4). The cell extraction algorithm (EXTRACT, Inan et al. (2021)) identified a total of approximately 40,000 cells across five imaging datasets. Next, we asked a total of 6 annotators to independently perform cell sorting on these movies, utilizing 4 annotators per movie (Table S1, Figs. S3 and S4). The resulting dataset contained approximately 160,000 annotations, which we used to benchmark ActSort.

2.2 A new feature set for cell classification

When working with images, deep-learning based approaches have proven to be valuable within experimental neuroscience (Stringer et al. (2021)). However, in this work, we took the opposite approach: feature engineering (See Fig. S1), followed by logistic regression. There were several motivations behind this design choice: i) the convexity of the logistic regression proved to be desirable for online training of cell classifiers during cell sorting, ii) pre-training a deep-network to learn features and then performing logistic regression for fine-tuning required substantial amount of data and standardization across different sizes and styles of calcium imaging videos, which was not feasible due to the size and heterogeneous nature of the different datasets, and iii) logistic regression on expert-engineered features was sufficient to outperform human annotators. Hence, our work can be considered as a strong baseline driven from domain knowledge for any potential future work that aims to utilize deep-learning approaches.

2.3 ActSort: Active-learning accelerated cell sorting

We illustrate the inner-workings of the ActSort algorithm in Figs. 1A, B and Algorithm 1. The cell candidates, *i.e.*, the cell extraction outputs, are fed into the query algorithm together with the engineered features (Figs. S1 and 1A). The query algorithm, which is the heart of the active learning framework here, creates a closed-loop online feedback between human annotator and the cell classifier training. Specifically, it decides the next cell to be sorted by the human annotator. Then, the new labeled dataset is used to train the cell classifier and the process continues iteratively. The goal of the query algorithm is to minimize the number of annotated samples without compromising on the cell classification accuracy, which is achieved by selecting the most "informative" cells for annotation that best capture the uncertainty in the cell classification.

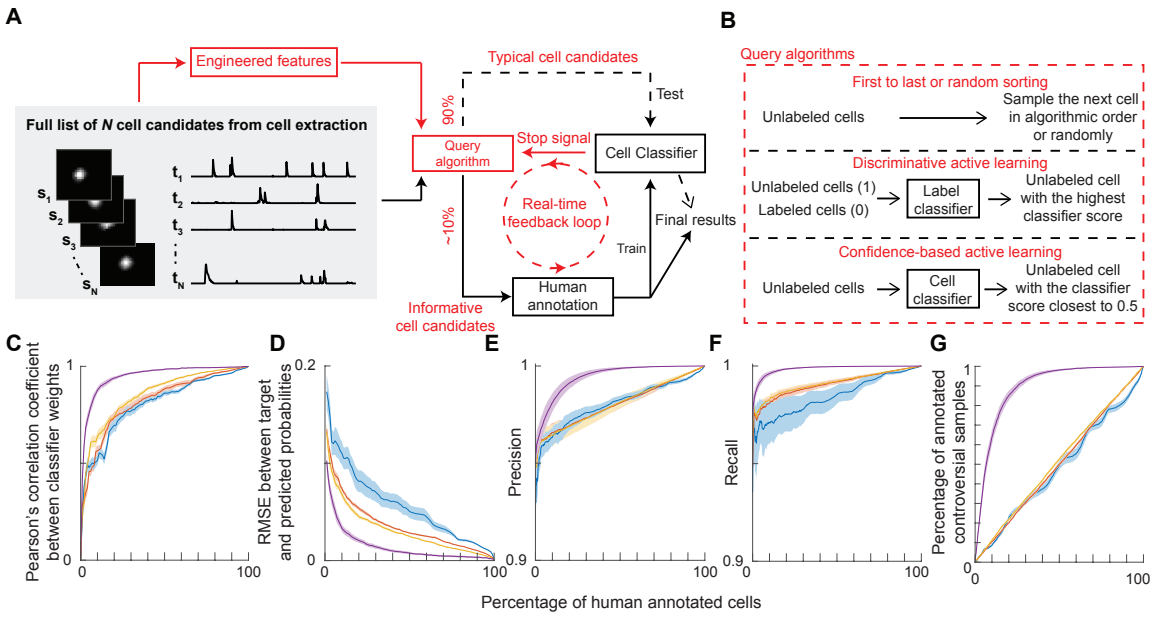

Figure 1: **ActSort convergences rapidly to the annotator's preference. A** ActSort workflow, which contains an online closed-loop feedback between the human annotator and the cell classifier. Engineered features and classifier predictions are fed back into the query algorithm, which provides the next cell or group of cells for the human annotator to sort. The (default) query algorithm picks the most informative cells *i.e.*, those that are on the decision boundary, to present to the annotator. This results in faster convergence of the classifier to the asymptotic performance with fewer number of samples than using a randomized approach that picks up samples that are easy to be classified with the features alone, and therefore, are not needed to be shown to the annotator. Red text denotes innovations introduced in ActSort to the traditional pipeline in Fig. S1. **B** Query algorithms tested: First-to-last, random sorting, discriminative active learning (DAL), and confidence-based active learning (CAL, the default for ActSort). DAL trains a second classifier, termed the label classifier (**Methods**), whereas CAL makes use of the existing cell classifier.**C-G** Comparison between query strategies as a function of the percentage of human-annotated cells. **C** Pearson's correlation between the current and the final (*i.e.*, the one trained on the full dataset) classifier weights. **D** Root mean-squared error (RMSE) between target and predicted probabilities. **E** Precision and **F** recall of ActSort where the annotator labels were allowed to supersede the classifier labels. **G** Fraction of annotated controversial samples, defined as the data points where the final classifier and the human annotator disagreed. CAL, but not others, preferentially picked the data points that the cell classifier would eventually disagree with the annotator, allowing the final result to match the annotator's preference. Solid line, mean; shaded area, standard errors of the mean across 15 annotations.

In this work, we tested two active-learning approaches that aim to mitigate two distinct types of uncertainties: i) uncertainty regarding whether currently labeled samples faithfully represent the full dataset, and ii) uncertainty in the fine location of the decision boundary of the current model. The active learning approach that addresses the former is called "discriminative active learning (DAL, Gissin and Shalev-Shwartz (2019))." In this paradigm,

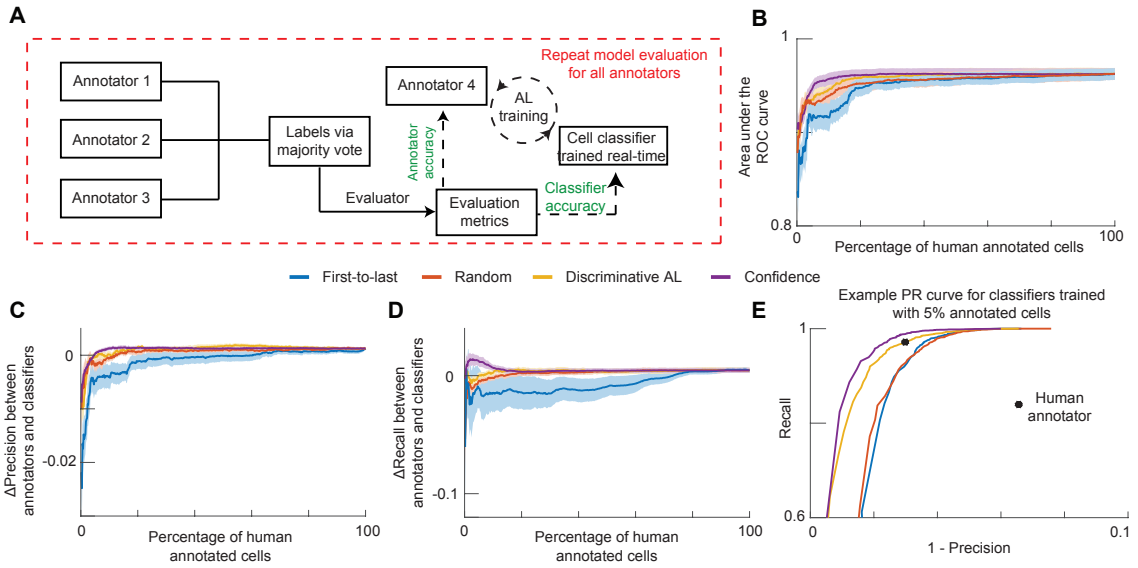

Figure 2: **ActSort with CAL outperforms human annotators with 5% sorted cells.** **A** Workflow to test online classifiers' accuracy with respect to individual annotators. **B-E** Benchmark of the different sampling approaches: First-to-last, random, DAL, and CAL. **B** Area under the receiver operating characteristic (ROC) curve as a function of the percentage of human-annotated cells, summarizing the accuracy of the classifiers without specifying any classification threshold. **(C)** Precision and **(D)** Recall the difference between annotators and the cell classifiers, using 0.65 as the classification threshold (based on the calibration curves in Fig. S2). Positive values reflect that the cell classifier more accurately predicted the ground truth labels than the human annotator. Solid lines, mean; shaded areas, standard error of means over 12 annotation instances. **E** Example precision-recall (PR) curve for a single annotation instance. Solid lines, PR curves of each classifier's predictions on the full dataset; black dot, human annotator accuracy. Only with 5% annotated cells, ActSort outperforms the human annotator in both recall and precision.

a second label classifier is trained within the query algorithm to pick the least represented unlabeled sample (Fig. 1B, Methods). In contrast, confidence-based active learning (CAL, Settles and Craven (2008)) picks the unlabeled sample closest to the decision boundary, *i.e.*, where the cell classifier is least certain (Fig. 1B). In simpler terms, DAL assumes the labeled set does not faithfully represent the full dataset and explores far away from the cell classifier boundary; whereas CAL exploits the boundary of the current cell classifier for fine-tuning, accelerating convergence substantially if the current estimate of the boundary is approximately correct.

To test ActSort with both DAL and CAL, we used the three movies from Fig. S1. To provide baselines for the traditional approaches, we also allowed the query algorithm to pick the next sample randomly or in the order the cells are received from the cell extraction algorithm (Fig. 1B). Our results indicated that CAL, but not others, leads to fast convergence (Fig. 1C-G). Notably, CAL preferentially picked the samples with which the annotator and the fully trained cell classifier would eventually disagree on (Fig. 1G). In other words, ActSort with CAL converged rapidly to the annotator's preference.

2.4 ActSort beats human level performance with 5% annotation

So far, our tests focused on ActSort's ability to match the annotator's preference in sorting cells. But, what if the human annotator is not always correct and has some inherent biases? Can ActSort provide more accurate cell classification results? To test the accuracy of the online classifiers with respect to the individual annotators, we designed an evaluation process (Fig. 2A). To validate one annotator, we first created a ground truth evaluator by taking the majority vote of the other annotators. The classifier was trained by the left-out annotator in an online manner following the active learning paradigm. Both the left-out annotator and the classifier accuracies were then computed from the ground truth evaluator labels.

In this scenario, once again, CAL outperformed other approaches (Fig. 2B). Notably, ActSort with CAL outperformed individual human annotators *even though it was trained on their output* (Fig. 2C-E). Specifically, when the online classifiers were trained with only 5% of the data, CAL outperformed the human annotator in the precision-recall curve (Fig. 2E). In comparison, the approach of sorting from first-to-last achieved the same mark only when approximately 80% of the cells were sorted (Fig. 2C, D). We benchmarked ActSort further on diverse conditions, results are discussed in Appendix C.

3. Discussion

In this work, we introduced ActSort, an active-learning accelerated cell sorting algorithm for large-scale calcium imaging datasets. We prepared a large-scale benchmark to evaluate the cell sorting performance of different approaches. On this benchmark, we showed that ActSort can rapidly match the annotator performance in comparison to random or first-to-last sampling, and outperforms a typical human annotator when approximately 5% of the full dataset was annotated.

In order to train a classifier to achieve accelerated cell sorting that surpasses human performance without requiring a large amount of pre-training data, we engineered 76 features that quantify the aspects of the data that human annotators pay attention to. These included temporal features (e.g., number of spikes in a trace, the signal-to-noise ratio of the spikes, etc.), spatial features (e.g., mean pixel intensity, circumference, etc.), and spatiotemporal features (e.g. severity of non-Gaussian noise contamination in movie snapshots). The newly engineered features were able to more accurately reject false-positive cells while having a minimal impact on the true-positive accuracy.

The potential impact of this work is two-fold: i) vastly decrease the amount of hours humans spent annotating modern calcium-imaging data sets and ii) surpass the accuracy of cell classification by experts. A paradigm requiring that merely 5% of cells be annotated by a human, can reduce cell sorting from several hours per calcium-imaging data set, to merely tens of minutes. Through ActSort, researchers can reclaim tens of hours that would have otherwise been spent labeling modern massive data sets. Importantly, increasing cell detection accuracy can be largely beneficial to modern systems neuroscience, which relies on minimizing false positive cell identification to avoid spurious results. Thus, removing the human element from repetitive tasks that can be prone to errors can ultimately improve the reliability, robustness, and reproducibility of systems neuroscience research.

Acknowledgements

HOA acknowledges funding support from Ozyegin University. JZ thanks Matlab's Toolbox Trainee program for fostering a connecting to the EXTRACT team at Schnitzerlab. FD and YL received funding from Stanford University's Mind, Brain, Computation and Technology program, which is supported by the Stanford Wu Tsai Neuroscience Institute. CM acknowledges funding support from the Stanford University Wu Tsai Neurosciences Institute Interdisciplinary Scholar Award.

References

Hillel Adesnik and Lamiae Abdeladim. Probing neural codes with two-photon holographic optogenetics. *Nature neuroscience*, 24(10):1356–1366, 2021.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

P Berens, L Theis, J Stone, N Sofroniew, A Tolias, M Bethge, and J Freeman. Standardizing and benchmarking data analysis for calcium imaging. In *Computational and Systems Neuroscience Meeting (COSYNE 2017)*, pages 66–67, 2017.

Jessica A Cardin, Michael C Crair, and Michael J Higley. Mesoscopic imaging: shining a wide light on large-scale neural dynamics. *Neuron*, 108(1):33–43, 2020.

Luis Carrillo-Reid, Shuting Han, Weijian Yang, Alejandro Akrouh, and Rafael Yuste. Controlling visually guided behavior by holographic recalling of cortical ensembles. *Cell*, 178 (2):447–457, 2019.

Lynne Chantranupong, Celia C Beron, Joshua A Zimmer, Michelle J Wen, Wengang Wang, and Bernardo L Sabatini. Dopamine and glutamate regulate striatal acetylcholine in decision-making. *Nature*, 621(7979):577–585, 2023.

Tsai-Wen Chen, Trevor J Wardill, Yi Sun, Stefan R Pulver, Sabine L Renninger, Amy Baohan, Eric R Schreiter, Rex A Kerr, Michael B Orger, Vivek Jayaraman, et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300, 2013.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

Gregory Corder, Biafra Ahanonu, Benjamin F Grewe, Dong Wang, Mark J Schnitzer, and Grégory Scherrer. An amygdalar neural ensemble that encodes the unpleasantness of pain. *Science*, 363(6424):276–281, 2019.

Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI*, volume 5, pages 746–751, 2005.

Aniruddha Das, Sarah Holden, Julie Borovicka, Jacob Icardi, Abigail O'Niel, Ariel Chaklai, Davina Patel, Rushik Patel, Stefanie Kaech Petrie, Jacob Raber, et al. Large-scale recording of neuronal activity in freely-moving mice at cellular resolution. *Nature Communications*, 14(1):6399, 2023.

Jeffrey Demas, Jason Manley, Frank Tejera, Kevin Barber, Hyewon Kim, Francisca Martínez Traub, Brandon Chen, and Alipasha Vaziri. High-speed, cortex-wide volumetric recording of neuroactivity at cellular resolution using light beads microscopy. *Nature Methods*, 18 (9):1103–1111, 2021.

Daniel A Dombeck, Anton N Khabbaz, Forrest Collman, Thomas L Adelman, and David W Tank. Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron*, 56(1):43–57, 2007.

Sadegh Ebrahimi, Jérôme Lecoq, Oleg Rumyantsev, Tugce Tasci, Yanping Zhang, Cristina Irimia, Jane Li, Surya Ganguli, and Mark J Schnitzer. Emergent reliability in sensory cortical coding and inter-area communication. *Nature*, 605(7911):713–721, 2022.

Jonathan Folmsbee, Xulei Liu, Margaret Brandwein-Weber, and Scott Doyle. Active deep learning: Improved training efficiency of convolutional neural networks for tissue classification in oral cavity cancer. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 770–773, 2018. doi: 10.1109/ISBI.2018.8363686.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

Jeffrey L Gauthier, Sue Ann Koay, Edward H Nieh, David W Tank, Jonathan W Pillow, and Adam S Charles. Detecting and correcting false transients in calcium imaging. *Nature Methods*, pages 1–9, 2022.

Andrea Giovannucci, Johannes Friedrich, Pat Gunn, Jérémie Kalfon, Brandon L Brown, Sue Ann Koay, Jiannis Taxidis, Farzaneh Najafi, Jeffrey L Gauthier, Pengcheng Zhou, et al. Caiman an open source tool for scalable calcium imaging data analysis. *elife*, 8: e38173, 2019.

Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. *arXiv preprint arXiv:1907.06347*, 2019.

Francesco Gobbo, Rufus Mitchell-Heggs, Dorothy Tse, Meera Al Omrani, Patrick A. Spooner, Simon R. Schultz, and Richard G. M. Morris. Neuronal signature of spatial decision-making during navigation by freely moving rats by using calcium imaging. *Proceedings of the National Academy of Sciences*, 119(44):e2212152119, 2022. doi: 10.1073/pnas.2212152119. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2212152119`.

Christine Grienberger and Arthur Konnerth. Imaging calcium in neurons. *Neuron*, 73(5): 862–885, 2012.

Yuhong Guo. Active instance sampling via matrix partition. *Advances in Neural Information Processing Systems*, 23, 2010.

Christopher D Harvey, Philip Coen, and David W Tank. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature*, 484(7392):62–68, 2012.

Alexander G Hauptmann, Wei-Hao Lin, Rong Yan, Jun Yang, and Ming-Yu Chen. Extreme video retrieval: joint maximization of human and computer performance. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 385–394, 2006.

Steven CH Hoi, Rong Jin, and Michael R Lyu. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web*, pages 633–642, 2006.

Hakan Inan, Murat A Erdogdu, and Mark Schnitzer. Robust estimation of neural signals in calcium imaging. *Advances in neural information processing systems*, 30, 2017.

Hakan Inan, Claudia Schmuckermair, Tugce Tasci, Biafra O Ahanonu, Oscar Hernandez, Jérôme Lecoq, Fatih Dinç, Mark J Wagner, Murat A Erdogdu, and Mark J Schnitzer. Fast and statistically robust cell extraction from large-scale neural calcium imaging datasets. *bioRxiv*, 2021.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*, pages 2372–2379. IEEE, 2009.

Tony Hyun Kim and Mark J Schnitzer. Fluorescence imaging of large-scale neural ensemble dynamics. *Cell*, 185(1):9–41, 2022.

David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994.

Xin Li and Yuhong Guo. Adaptive active learning for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 859–866, 2013.

Michael Z Lin and Mark J Schnitzer. Genetically encoded indicators of neuronal activity. *Nature neuroscience*, 19(9):1142–1153, 2016.

Peng Liu, Hui Zhang, and Kie B. Eom. Active deep learning for classification of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(2):712–724, 2017. doi: 10.1109/JSTARS.2016.2598859.

Ying Liu. Active learning with support vector machine applied to gene expression data for cancer classification. *Journal of chemical information and computer sciences*, 44(6): 1936–1941, 2004.

Zimo Liu, Jingya Wang, Shaogang Gong, Huchuan Lu, and Dacheng Tao. Deep reinforcement active learning for human-in-the-loop person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6122–6131, 2019.

Andrew McCallum, Kamal Nigam, et al. Employing em and pool-based active learning for text classification. In *ICML*, volume 98, pages 350–358. Citeseer, 1998.

Malavika Murugan, Hee Jae Jang, Michelle Park, Ellia M Miller, Julia Cox, Joshua P Taliaferro, Nathan F Parker, Varun Bhave, Hong Hur, Yupu Liang, et al. Combined social and spatial coding in a descending projection from the prefrontal cortex. *Cell*, 171 (7):1663–1677, 2017.

Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuroscience*, 22(10):1677–1686, 2019.

Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79, 2004.

Marius Pachitariu, Carsen Stringer, Sylvia Schröder, Mario Dipoppa, L Federico Rossi, Matteo Carandini, and Kenneth D Harris. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *BioRxiv*, page 061507, 2016.

Adam M Packer, Lloyd E Russell, Henry WP Dalgleish, and Michael Häusser. Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo. *Nature methods*, 12(2):140–146, 2015.

Liam Paninski and John P Cunningham. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology*, 50: 232–241, 2018.

Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 89(2):285–299, 2016.

John Peter Rickgauer, Karl Deisseroth, and David W Tank. Simultaneous cellular-resolution optical perturbation and imaging of place cell firing fields. *Nature neuroscience*, 17(12): 1816–1824, 2014.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2:441–448, 2001.

Oleg I Rumyantsev, Jérôme A Lecoq, Oscar Hernandez, Yanping Zhang, Joan Savall, Radosław Chrapkiewicz, Jane Li, Hongkui Zeng, Surya Ganguli, and Mark J Schnitzer. Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, 580(7801):100–105, 2020.

G Sayantan, PT Kien, and KV Kadambari. Classification of ecg beats using deep belief network and active learning. *Medical and Biological Engineering and Computing*, 56(10): 1887–1898, 2018.

Burr Settles. Active learning literature survey. 2009.

Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008.

Krishna V Shenoy and Jonathan C Kao. Measurement, manipulation and modeling of brain-wide neural population dynamics. *Nature communications*, 12(1):633, 2021.

Asim Smailagic, Pedro Costa, Hae Young Noh, Devesh Walawalkar, Kartik Khandelwal, Adrian Galdran, Mostafa Mirshekari, Jonathon Fagert, Susu Xu, Pei Zhang, et al. Medal: Accurate and robust deep active learning for medical image analysis. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 481–488. IEEE, 2018.

Nicholas James Sofroniew, Daniel Flickinger, Jonathan King, and Karel Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *elife*, 5:e14472, 2016.

Ian H Stevenson and Konrad P Kording. How advances in neural recording affect data analysis. *Nature neuroscience*, 14(2):139–142, 2011.

Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021.

Atika Syeda, Lin Zhong, Renee Tung, Will Long, Marius Pachitariu, and Carsen Stringer. Facemap: a framework for modeling neural activity based on orofacial tracking. *Nature Neuroscience*, pages 1–9, 2023.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Roger Y Tsien. New calcium indicators and buffers with high selectivity against magnesium and protons: design, synthesis, and properties of prototype structures. *Biochemistry*, 19 (11):2396–2404, 1980.

Anne E Urai, Brent Doiron, Andrew M Leifer, and Anne K Churchland. Large-scale neural recordings call for new insights to link brain and behavior. *Nature neuroscience*, 25(1): 11–19, 2022.

Jie Yang et al. Automatically labeling video data using multi-class active learning. In *Proceedings Ninth IEEE international conference on computer vision*, pages 516–523. IEEE, 2003.

Pengcheng Zhou, Shanna L Resendez, Jose Rodriguez-Romaguera, Jessica C Jimenez, Shay Q Neufeld, Andrea Giovannucci, Johannes Friedrich, Eftychios A Pnevmatikakis, Garret D Stuber, Rene Hen, et al. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *Elife*, 7:e28728, 2018.

Appendix A. Methods

A.1 Cell sorting with binary classifiers

The accelerated automated cell sorting is built with a traditional cell classifier and an active learning framework that prompts the human annotator to label the sample that will result in the highest improvement in the prediction accuracy, *i.e.*, creating an online feedback loop between the human annotator and the cell classifier. Within this paradigm, the query algorithm controls the iterative process of training and labeling between the cell classifier and the human annotator, which continues until convergence.

To start with, we defined a binary classification problem between *cell* and *not cell*, with input data $\mathcal{X} \in \mathbb{R}^{N \times d}$, where N is the total number of extracted putative neurons and d is the dimension of features, and ground truth labels $\mathcal{Y} \in \mathcal{R}^N$. We labeled *cell* as 1 and *not cell* as 0. We further defined two sets, the labeled set $\mathcal{L}^{(t)} = \{(x_i, y_i)\}_{i=1}^{n_1^{(t)}}$ and the unlabeled set $\mathcal{U}^{(t)} = \{x_i\}_{i=1}^{n_2^{(t)}}$, where $n_1^{(t)} + n_2^{(t)} = N$.

To allow real-time training and owing to its already sufficient accuracy thanks to the new feature set, we used logistic regression as our cell classifier with the following model:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^\top x}}, x \in \mathcal{X} \tag{S1}$$

We trained the cell classifier h_θ using the labeled dataset $\mathcal{L}$

$$\hat{\theta} = \arg\max_\theta \ell(\theta) = \arg\max_\theta \sum_{i=1}^{n_1} c_i [y_i \log h_\theta(x_i) + (1 - y_i) \log(1 - h_\theta(x_i))] + \lambda ||\theta||_1, \tag{S2}$$

where c_i are empirical class prior probabilities and λ is the L1 regularization strength. Though we used a fixed value of $\lambda = 1/n_{\text{samples}}$ for our experiments, to allow near real-time training that would be slowed down in a cross-validation scenario, we tested in Figure S5 a range of regularization parameters. We concluded that undershooting, but not overshooting, the optimal regularization parameter leads to comparable accuracies, a condition that the default value we chose satisfied.

A.2 Benchmarking

In this work, we tested four strategies as query algorithms, first to last, random sampling, confidence-based active learning (CAL, see for example Culotta and McCallum (2005); Settles and Craven (2008)), and discriminative active learning (DAL) Gissin and Shalev-Shwartz (2019). The overall training structure is depicted in Algorithm 1.

Traditional baselines

The first-to-last algorithm simply selects $x^{(t)} = u_1 \in \mathcal{U}$ for human labeling, where u_1 is the first unlabeled sample in the order of cell extraction output. The random sampling algorithm randomly select an unlabeled sample $x^{(t)}$ from $\mathcal{U}$. After human annotation, the sample point is moved to the labeled set $\mathcal{L}^{(t+1)}$ and the cell classifier is updated accordingly:

$$\mathcal{L}^{(t+1)} = \mathcal{L}^{(t)} \cup \{(x^{(t)}, y^{(t)})\}, \tag{S3a}$$

$$\mathcal{U}^{(t+1)} = \mathcal{U}^{(t)} \backslash x^{(t)} \tag{S3b}$$

Confidence-based active learning

The CAL algorithm selects sample according to the uncertainty sampling decision rule, which identifies unlabeled items that are near a decision boundary in the current cell classifier model. In a multi-class setting, one can use entropy to define the uncertainty of the sample points in the unlabeled dataset, which is how we shall motivate the CAL algorithm. Let $\mathcal{H}(\cdot)$ represent the entropy of a set of samples. In round $t+1$, we want to choose the sample that produces the maximum reduction in entropy given the trained cell classifier $h_\theta^{(t)}$ in round t:

$$\begin{aligned} x^{(t+1)} &= \arg\max_x \mathcal{H}\left(\mathcal{U}^{(t)}|\mathcal{L}^{(t)}\right) - \mathcal{H}\left(\mathcal{U}^{(t+1)}|\mathcal{L}^{(t+1)}\right) \\ &= \arg\min_x \mathcal{H}\left(\mathcal{U}^{(t+1)}|\mathcal{L}^{(t+1)}\right) \\ &= \arg\min_{x\in\mathcal{U}^{(t)}} h_\theta^{(t)}(x) \log\left(h_\theta^{(t)}(x)\right). \end{aligned} \tag{S4}$$

In this work, since we are using only two classes, the probability distribution that maximizes the entropy is the one where h_θ is as close to 0.5 as possible. However, extension of ActSort to multi-class problems, for example one that includes the identification of dendrites, could benefit from the more general entropy minimization framework.

Discriminative active learning

The DAL algorithm considers a different type of uncertainty, *i.e.*, how well a data point is represented in the labeled pool Gissin and Shalev-Shwartz (2019). Here, one constructs a new label space, $\tilde{\mathcal{Y}} = \{l, u\}$, that specifies whether the data sample is from the labeled set or the unlabeled set. For example, in round t, the training dataset becomes $\tilde{\mathcal{D}} = \{(x_i, u), (x_j, l) \mid \forall x_i \in \mathcal{U}^{(t)}, \forall x_j \in \mathcal{L}^{(t)}\}$. For each round of the active learning process, we train a binary classifier, $p_\phi^{(t)}(\tilde{y}|x)$, to predict whether the samples in $\tilde{\mathcal{D}}$ are labeled or not. Then, one selects the sample from the unlabeled set that satisfies

$$x^{(t+1)} = \arg\max_{x\in\mathcal{U}^{(t)}} \hat{p}_\phi^{(t)}(y = u|\tilde{\mathcal{D}}). \tag{S5}$$

A.3 Calcium imaging datasets

In the calcium imaging studies depicted in Figures 1, 2, S1, S2, and Table S1, we used triple transgenic GCaMP6f-tTA-dCre mice from Allen Institute (Rasgrf2-2A-dCre/CaMK2a-tTA/Ai93). To prepare mice for in vivo imaging sessions, we performed surgeries while mice were mounted in a stereotaxic frame under anaesthesia. We created a cranial window by removing a $7mm$ diameter skull flap over the right cortical area S1 and surrounding cortical tissue. We covered the exposed cortical surface with a $7mm$ diameter glass coverslip. A custom

Algorithm 1 ActSort

Data: $\mathcal{L}^{(0)} = \{(x_i, y_i)\}_{i=1}^{n_1^{(0)}}$, $\mathcal{U}^{(0)} = \{x_i\}_{i=1}^{n_2^{(0)}}$ ▷ Initial Labeling
for $t = 1, \ldots, T$ **do**
  Train the cell classifier $h_\theta^{(t)}$ using $\mathcal{L}^{(t)}$
  Select $x^{(t+1)}$ based on the `Query-Algorithm` and $h_\theta^{(t)}$
  Human annotator label $x^{(t+1)}$ with $y \in \{0, 1\}$
  $\mathcal{L}^{(t+1)} = \mathcal{L}^{(t)} \cup (x^{(t+1)}, y)$
  $\mathcal{U}^{(t+1)} = \mathcal{U}^t \backslash x^{(t+1)}$
end for

wide-field fluorescence macroscope with a field of view covering the full cranial window was used for neural activities imaging. For epifluorescence illumination, we used a LED with spectrum centered 475nm. We acquired Ca2+ videos of neural activity (50 Hz frame rate, $1,708 \times 1,708$ pixels) on the fluorescence macroscope.

In the calcium imaging studies depicted in Figure S3, we used previously published imaging datasets imaging eight neocortical brain regions Ebrahimi et al. (2022). To test ActSort under the scenario of multiple false-positives, we performed cell extraction with weak quality checking Inan et al. (2017, 2021) on one of the sessions of roughly 20 min duration, which led to 2345 false-positive regions of interest.

In the calcium imaging studies depicted in Figure S4, we employed a custom two-photon mesoscope with a multi-foci illumination technique Rumyantsev et al. (2020). This system facilitated recordings over an extensive field-of-view ($2 \times 2\,mm^2$) at a frame rate of 30 Hz and a resolution of 2 μm pixels (1024×1024 pixels). Utilizing this equipment, we captured a movie of layer 2/3 cortical pyramidal neurons in live triple transgenic mice from the Allen Institute.

A.4 Feature engineering

Below is the list of features we used for this work. We have 76 features in total and we categorize the features as follows: Spatial Filter-based features, Trace-based features, Snapshot-based (from the movie) features, and Spatiotemporal (Combination of all of the previous ones) features. The traditional ones are colored in blue.

Features 1 to 36 are based on the Ca^{2+} trace activities (activity signals of the neurons detected by the cell extraction algorithm) of the given cell dataset.

Feature 1 calculates the signal-to-noise ratio for the given trace activity. A lower signal-to-noise ratio indicates a noisy data, which is more likely to contain false positives. A higher ratio indicates a higher quality data. This feature is taken from EXTRACT Inan et al. (2021).

Feature 2 calculates how much a cell's trace activity signal differs from its smoothed version. If the difference is high after smoothing, then it is likely for the cell trace activity to contain a good amount of noise. This feature is taken from EXTRACT Inan et al. (2021).

Feature 3 calculates the number of bad (noisy) spikes within the trace. Noisy trace activities tend to have more 'bad' spikes than clear trace activities. The presence of more bad spikes can effectively signal a bad candidate cell (false positive).

Feature 4 calculates the standard deviation of the upper decile (top 10%) of each trace activity, revealing the variability within the higher range of the activity.

Feature 5 measures the average time interval between 'good' spikes. Non-noisy trace activity tends to exhibit less frequent spike occurrences. This characteristic helps the identification of clear cell activity.

Feature 6 measures the average time interval between 'bad' (noisy) spikes. Noisy trace activity tends to exhibit more frequent spike occurrences. This characteristic helps the identification of 'bad' cell activity.

Feature 7 computes the average time interval between all spikes within the trace activity, providing insights of the regularity of cell activities.

Feature 8 measures the average width of all spikes in a trace activity. Spikes with narrower widths generally indicate more distinct spike events, which can be useful for spotting a higher proportion of good spikes within a given trace.

Feature 9 measures the average width of 'good' spikes. A smaller average width suggests more precise and ideal spike events.

Feature 10 calculates the standard deviation of the spike width within a trace of cell activity. A more widely distributed spike width can be a good indicator of noisy and less precise activity.

Feature 11 calculates the standard deviation of the spike width among the 'good' spikes. This can serve as an indicator of their quality. A narrower distribution of spike widths suggests a higher level of consistency and precision in spike events.

Feature 12 measures the proportion of total activity in a cell trace is made up of significant events rather than noise, referred as relative power. This metric helps the distinction between actual neural events (consistent signal energy) and a likelihood of noise.

Feature 13 calculates the proportion of cell trace activity that surpasses the noise threshold. This tells the frequency of cell activities that is distinguished from the background noise.

Feature 14 counts the instances when the cell's activity surpasses the noise threshold significantly. This provides insights into the frequently of firing or notable cell activities.

Features 15 - 29 are derived from different scales and threshold values: 3, 10, 30, and 100, respectively— for the calculation of relative power (Feature 12), the proportion of the cell trace activity that surpasses the noise level (Feature 13), and the number of significant events (Feature 14).

Features 30 - 34 calculate a quality score that indicates how much of a neuron's activity occurs within certain frequency ranges of the spectrum. A higher score indicates that the majority of the neuron's activity is happening within the specified frequency range, suggesting that the data quality is good and that the neuron is reliably detected. The frequency ranges are determined by lower and higher threshold parameters. In our work, the thresholds are set to define four specific ranges within the frequency spectrum: the lowest 5% (0-0.05), the range from the lowest 5% up to the highest 95% (0.05-0.95), the lower half (0-0.5), and the upper half (0.5-1).

Feature 35 calculates for each cell the proportion of other cells in the dataset with which it has a high correlation, using a similarity threshold of 70%. This essentially measures how similar a cell's activity is to the activity of the other cells in the dataset.

Feature 36 finds the total fluorescence that was caused by the highest 10% activity in a given cell activity trace. This metric can be useful in detecting unusual neuron activity that doesn't reflect a common neuron behavior in the dataset.

Features 37 to 46 are based on the spatial filters (the overall pictures of the cells according to the cell extraction algorithm) in the given cell dataset.

Feature 37 calculates the total surface area of each cell in the movie. This can be useful in detecting cells that are too large or too small in size.

Feature 38 finds the number of distinct bodies that a cell candidate has. Cells that are detected to have more than one body tend to be false positives.

Feature 39 calculates the total sum of pixel values within each cell's spatial filter, which is a measure of the total activity or brightness. 'Bad' cells tend to have unusually high or low activity.

Feature 40 measures the circumference of the cell. Typically, cells with too large or too small circumference tend to be false positives.

Feature 41 measures the distance from the cell center to the nearest edge of the movie's field of view. Cells that are located closer to the edges tend to be noise or artifacts due to disruptions in the movie, such as boundary effects or uneven illumination.

Feature 42 computes the circularity of a given cell, which measures how closely it's shape aligns with a perfect circle. Cells with a higher degree of circularity are generally preferable candidates. This feature is taken from EXTRACT Inan et al. (2021).

Feature 43 computes the eccentricity of a given cell, which indicates how stretched out a cell's shape is. Candidates with higher eccentricity values tend to be 'bad' cells. This feature is taken from EXTRACT Inan et al. (2021).

Feature 44 computes the average pixel value of the given cell, which is a measure of the average activity. 'Bad' cells tend to have unusually high or low activity.

Feature 45 calculates the spatial corruption, which is known as errors or inconsistencies within the spatial filter of the cell by evaluating the local variance relative to the overall variance of the cell image. This feature is taken from EXTRACT Inan et al. (2021).

Feature 46 calculates the maximum spatial correlation of each cell with all other cells in the movie. This is helpful to identify cells that have similar activity patterns. This feature is taken from EXTRACT Inan et al. (2021).

Features 47 to 76 are based on spatial filters and trace activities of the cells that are captured by the cell extraction algorithm, along with key movie frames that correspond to the highest points of cell activity. We call these spatiotemporal features.

Features 47 - 50 evaluate the alignment between the cell's spatial filter, obtained from the cell detection algorithm, and critical movie frames extracted from the actual movie during the cell's highest activity time intervals. Feature 47 averages the errors to identify cells with pronounced deviations. Feature 48 seeks the minimum error to check the accuracy of the closest match. Feature 49 uses the 10th percentile to detect rare but significant errors. Feature 50 uses the 20th percentile for a broader to avoid missing detections in the alignment of spatial filters and key movie frames.

Features 51 - 66 calculate the correlations of activity within the neuropil area -a dense network of neurons and their components that can cause false positives when activated- by analyzing the interactions between neighboring pixels in and around the cells. This feature is taken from EXTRACT Inan et al. (2021).

Features 67 - 76 calculates correlation between the cell activity traces and movie frames. This feature is taken from EXTRACT Inan et al. (2021).

Appendix B. Background

B.1 Processing calcium imaging movies

Enabled by the development of the first fluorescent calcium indicators (Tsien (1980)), calcium signals represent an accurate proxy for simultaneous measurement of the neural activities (Gobbo et al. (2022); Ebrahimi et al. (2022); Musall et al. (2019); Harvey et al. (2012); Chantranupong et al. (2023); Murugan et al. (2017)) and can be combined with optogenetics techniques, allowing both recording and manipulation of neural circuits with unprecedented precision (Rickgauer et al. (2014); Packer et al. (2015); Adesnik and Abdeladim (2021); Carrillo-Reid et al. (2019)). This, in turn, allows the study of neural computation and discovery of the neural code at an unprecedented level across multiple brain regions and months of imaging sessions at single-cell level resolutions (Lin and Schnitzer (2016); Kim and Schnitzer (2022); Cardin et al. (2020); Shenoy and Kao (2021)). The advent of the experimental techniques, however, calls for an equally urgent push at the computational frontier to analyze these large-scale neural datasets (Urai et al. (2022); Stevenson and Kording (2011); Paninski and Cunningham (2018)).

In the early days of calcium imaging, owing to the low number of cells that could be recorded and tracked, identifying the sources of neural signals was as simple as drawing

regions of interest (ROIs) within the neural videos, and subsequently averaging their pixel activities (Dombeck et al. (2007); Chen et al. (2013)). As neuroscience experiments increased in complexity, the computational burden to process resulting raw datasets became an experimental bottleneck. This urgent need gave rise to the development of cell extraction algorithms that extract neural signals from large-scale recordings in an automated manner (Pnevmatikakis et al. (2016); Zhou et al. (2018); Pachitariu et al. (2016)), though primarily developed for datasets of $\sim O(10)$ gigabyte sizes. Hence, even though cell extraction regularly led to false-positive neuronal sources, which could disrupt the biological conclusions (Inan et al. (2021); Gauthier et al. (2022)), these datasets could in principle be fully annotated by humans within reasonable time frames (Corder et al. (2019)).

In this decade, on the contrary, large-scale neural recordings regularly produce video datasets of terabytes in size, with tens of thousands of neurons imaged across weeks (Kim and Schnitzer (2022); Syeda et al. (2023); Demas et al. (2021); Das et al. (2023)). Therefore, processing high volumes of large datasets efficiently is crucial. Recently, Inan et al. (2017, 2021) introduced a fast, scalable, and robust convex optimization algorithm, called EXTRACT, to extract neural dynamics from calcium imaging movies, with the ability to scale to large neural datasets. However, despite the high precision of EXTRACT, cell sorting may still be desired to achieve a perfect accuracy. This could, however, take days or even weeks of manual curation. As a first attempt, widely used CIATAH (Corder et al. (2019)), Suite2p (Pachitariu et al. (2016)), and CAIMAN (Giovannucci et al. (2019)) pipelines train cell classifiers using basic spatial and temporal features, and a small subset of randomly annotated cells. However, as we show in this work, blind randomization of training data is suboptimal and requires a substantial amount of human labor to be accurate, whereas existing features are insufficient for classifier training. Instead, we show below that a more principled solution not only reduces human labor in this post-processing bottleneck, but also improves the reliability, robustness, and reproducibility of the results.

B.2 Active learning for systems neuroscience

When training machine learning models for classification, some samples are more informative than others (Joshi et al. (2009); McCallum et al. (1998)). When labeling of training samples comes at a cost, it is not uncommon to design a closed-loop data annotation process that preferentially picks up the informative samples for the human annotators to label (Folmsbee et al. (2018); Liu et al. (2017); Sayantan et al. (2018)). The branch of machine learning that addresses the data annotation optimization is called *active learning,* in which the model's performance is optimized while requiring as few annotated samples as possible. To achieve this, active learning paradigms utilize a query algorithm to select the most informative samples from the unlabeled dataset and hand them over to human annotators for labeling.

Our application of interest falls into the *pool-based sampling* subcategory of active learning (Lewis and Catlett (1994)), where the models have access to a full unlabeled scientific dataset that needs to be fully classified with as little human labor as possible. This approach has already supported scientific research in several disciplines, such as text classification (Lewis and Catlett (1994); Tong and Koller (2001); Hoi et al. (2006)), image classification(Tong and Koller (2001); Joshi et al. (2009); Li and Guo (2013)), video classification

(Yang et al. (2003); Hauptmann et al. (2006)), and cancer diagnosis (Liu (2004); Smailagic et al. (2018)), to name a few. On the other side, several benchmarks such as MNIST (Gal et al. (2017)), CIFAR-10/100 (Gissin and Shalev-Shwartz (2019)), and Image-net (Citovsky et al. (2021)) are broadly used to test active learning approaches in the field of machine learning. Given its success across several disciplines, the active learning paradigm may be the missing piece for speeding up and scaling the cell sorting process for the large-scale calcium imaging datasets.

Our work starts with this premise but primarily aims to introduce active learning as a framework. In fact, there are several query strategies already developed in the field of active learning (Lewis and Catlett (1994); Gal et al. (2017); Settles (2009); Nguyen and Smeulders (2004); Guo (2010); Roy and McCallum (2001); Ash et al. (2019); Liu et al. (2019)). Yet, as we show here, even the two most simplest of approaches, confidence-based active learning (CAL, see for example Culotta and McCallum (2005); Settles and Craven (2008)), and discriminative active learning (DAL, Gissin and Shalev-Shwartz (2019)), provide massive improvements compared to the trivial strategies of picking labels randomly, or going from the first to last identified cell candidates. Inspired by this observation, this work will not develop new methods for calcium imaging datasets, which is left as future work. Instead, we will discuss the incorporation of the active learning framework into the calcium imaging preprocessing pipelines and introduce strong baselines via traditional approaches.

Appendix C. Additional results

C.1 ActSort becomes well calibrated with sparse samples

Having shown that CAL leads to rapid convergence compared to traditional and DAL baselines, we next asked the question: How well is ActSort with CAL calibrated compared to annotator agreements? For example, when ActSort made a mistake in annotation, was the correct label obvious for all human annotators?

To test whether the probabilities of the cell classifiers were well aligned with the uncertainty among the annotators, we divided the full dataset into five subsets, based on how many annotators accepted a sample as a cell (Fig. S2). Notably, the probabilities predicted by CAL, but not others, reached a steady equilibrium quickly (Fig. S2A,B), in which the distinction was apparent even with 1% annotated samples (Fig. S2C-G). Moreover, both the probabilities and the fraction of candidates classified as cells were distinct between five subgroups, with CAL making little-to-no mistakes ($< \%1$) when all annotators agreed. Hence, the predictions made by ActSort with CAL matched the annotator agreement scores, hinting that ActSort becomes well calibrated after only a few rounds of human annotation.

C.2 ActSort predictions are robust to experimental non-idealities

The half-hemisphere-wide movies we processed so far can be considered in the upper quality bound of imaging and cell extraction conditions, with false-positive rates already as low as 5-10% per movie and fairly high-quality cells. Next, we asked whether ActSort would still be able to accurately process a dataset with many false-positives. To test this, we processed one of the cortex-wide movies from (Ebrahimi et al. (2022)) with less stringent quality parameters in EXTRACT such that the resulting cell extraction output had many

false-positives, around 35%, by design (Fig. S3). CAL outperformed all other approaches in this dataset as well. Finally, we performed the same tests on a two-photon movie with non-zero residual motion (Fig. S4) and observed once again that CAL outperformed all other approaches. Hence, the low false-positive rate in the cell candidates or the optimal imaging/processing conditions cannot explain the success of ActSort in converging rapidly with few annotated samples.

Acceptance Accuracy

Predict / Annotate	Ann. 1	Ann. 2	Ann. 3	Ann. 4	Ann. 5	Ann. 6	Ann. Con.	Class. trad.	Class. new
Ann. 1	NA	0.967 0.990 NA	0.874 NA 0.980	0.986 0.993 0.998	NA NA 0.992	NA 0.953 Na	0.981 0.995 0.999	0.98(6) 0.99(3) 0.99(7)	0.98(4) 0.99(2) 0.99(7)
Ann. 2	0.971 0.902 NA	NA	0.880 NA NA	0.988 0.986 NA	NA NA NA	NA 0.923 Na	0.986 0.974 NA	0.98(9) 0.98(1) NA	0.98(7) 0.98(0) NA
Ann. 3	0.983 NA 0.891	0.985 NA NA	NA	0.998 NA 0.993	NA NA 0.979	NA NA NA	0.999 NA 0.989	0.99(5) NA 0.99(0)	0.99(5) NA 0.98(9)
Ann. 4	0.959 0.895 0.869	0.958 0.975 NA	0.863 NA 0.951	NA	NA NA 0.971	NA 0.917 Na	0.970 0.967 0.973	0.98(4) 0.97(8) 0.98(2)	0.98(2) 0.97(6) 0.98(0)
Ann. 5	NA NA 0.874	NA NA NA	NA NA 0.949	NA NA 0.983	NA	NA NA NA	NA NA 0.975	NA NA 0.98(0)	NA NA 0.97(8)
Ann. 6	NA 0.930 NA	NA 0.989 NA	NA NA NA	NA 0.993 NA	NA NA NA	NA	NA 0.994 NA	NA 0.99(0) NA	NA 0.99(0) NA
Ann. Consensus	0.979 0.923 0.892	0.980 0.992 NA	0.887 NA 0.971	0.995 0.996 0.997	NA NA 0.987	NA 0.946 NA	NA	0.99(1) 0.98(7) 0.98(9)	0.98(9) 0.98(7) 0.98(8)
Average	0.9(7) 0.9(1) 0.9(7)	0.9(7) 0.9(8) NA	0.8(8) NA 0.9(6)	0.9(9) 0.9(9) 0.9(9)	NA NA 0.9(8)	NA 0.9(4) NA	0.9(8) 0.9(8) 0.9(8)	0.9(9) 0.9(9) 0.9(9)	0.9(9) 0.9(9) 0.9(9)

Movie1 Movie2 Movie3

Rejection Accuracy

Predict / Annotate	Ann. 1	Ann. 2	Ann. 3	Ann. 4	Ann. 5	Ann. 6	Ann. Con.	Class. trad.	Class. new
Ann. 1	NA	0.653 0.439 NA	0.826 NA 0.434	0.508 0.393 0.288	NA NA 0.323	NA 0.626 Na	0.755 0.569 0.425	0.56(2) 0.43(4) 0.30(5)	0.58(6) 0.46(0) 0.33(6)
Ann. 2	0.623 0.896 NA	NA	0.854 NA NA	0.508 0.708 NA	NA NA NA	NA 0.880 Na	0.777 0.914 NA	0.56(4) 0.74(6) NA	0.59(6) 0.78(4) NA
Ann. 3	0.358 NA 0.824	0.387 NA NA	NA	0.280 NA 0.497	NA NA 0.479	NA NA NA	0.419 NA 0.710	0.29(2) NA 0.50(6)	0.31(3) NA 0.55(0)
Ann. 4	0.756 0.920 0.962	0.792 0.812 NA	0.963 NA 0.877	NA	NA NA 0.685	NA 0.914 Na	0.910 0.946 0.950	0.79(5) 0.80(8) 0.72(7)	0.83(3) 0.84(5) 0.79(8)
Ann. 5	NA NA 0.890	NA NA NA	NA NA 0.695	NA NA 0.564	NA	NA NA NA	NA NA 0.812	NA NA 0.57(6)	NA NA 0.62(8)
Ann. 6	NA 0.721 NA	NA 0.496 NA	NA NA NA	NA 0.450 NA	NA NA NA	NA	NA 0.649 NA	NA 0.48(4) NA	NA 0.51(3) NA
Ann. Consensus	0.772 0.953 0.987	0.832 0.751 NA	0.990 NA 0.868	0.625 0.678 0.659	NA NA 0.684	NA 0.946 NA	NA	0.63(4) 0.67(6) 0.60(1)	0.65(9) 0.71(7) 0.65(7)
Average	0.(6) 0.(8) 0.9(1)	0.(7) 0.(7) NA	0.9(1) NA 0.(7)	0.(5) 0.(6) 0.(5)	NA NA 0.(5)	NA 0.(8) NA	0.(7) 0.(8) 0.(7)	0.(6) 0.(6) 0.(6)	0.(6) 0.(7) 0.(6)

Movie1 Movie2 Movie3

Summary

Metric / Annotate	Average accuracy	F1 score
Ann. 1	0.86 ± 0.07	0.84 ± 0.11
Ann. 2	0.81 ± 0.09	0.76 ± 0.14
Ann. 3	0.87 ± 0.08	0.85 ± 0.11
Ann. 4	0.75 ± 0.07	0.67 ± 0.13
Ann. 5	0.76 ± 0.09	0.69 ± 0.15
Ann. 6	0.89 ± 0.07	0.88 ± 0.08
Ann. Consensus	0.86 ± 0.09	0.83 ± 0.13
Cell classifier	0.80 ± 0.08	0.75 ± 0.13

Table S1: **The fully trained cell classifiers scored similarly to human annotators.** The first two tables denote the cross-accuracy of annotators within each other, whereas the third one summarizes the average scores The three values in each table entry (i, j) represent the acceptance/rejection accuracies of annotator j, by assuming annotator i's labels are ground truth, computed among three movies. (Continued in the next page)

Table S1: (Continued from the previous page) The final row denotes the average annotation accuracies. The annotator consensus baseline is defined as the majority labels across the four human annotators, where the ties are broken with a coin flip. The accuracies of the cell classifiers, which were trained with both traditional and new feature sets, are obtained via the test sets from 10 independent 50-50 stratified train-test splits. The parentheses in the entries represent the uncertain digit whenever applicable. For the summary table (third table), the classifier with the new feature set was used. Overall, the accuracy metrics of the cell classifiers were within the range of typical human annotators, and the new feature set allowed the cell classifiers to reject more false-positives.

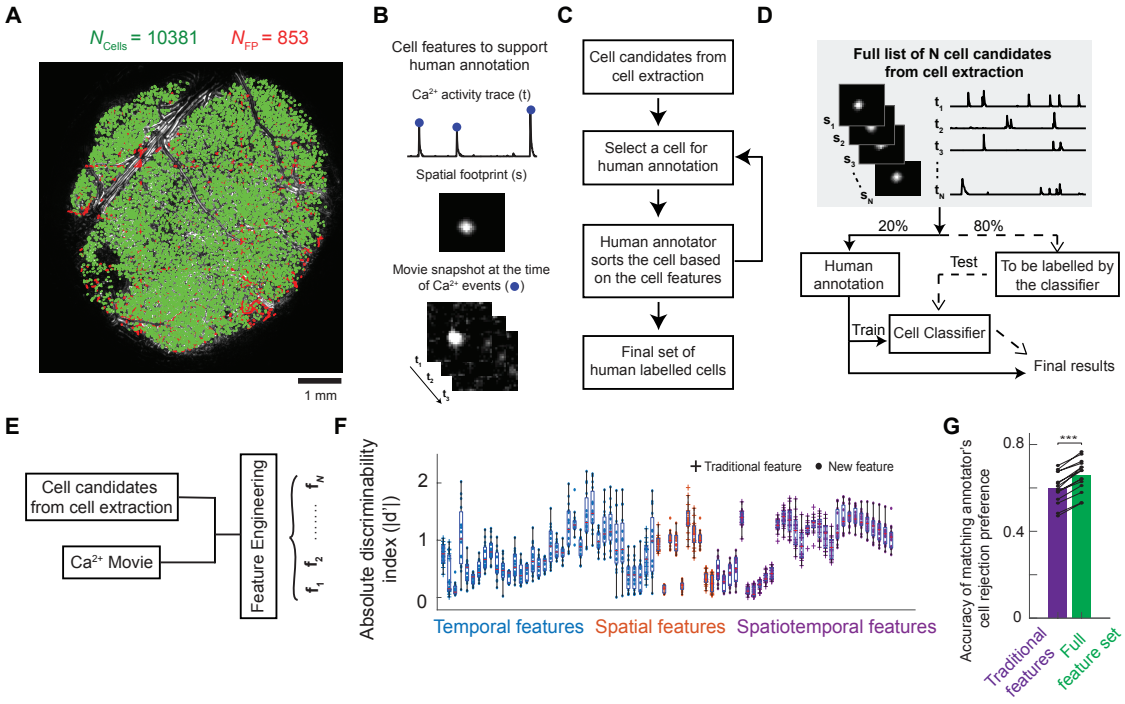

Figure S1: **Expert-assisted feature engineering improves cell classification. A** An example cell map from a 7mm field of view covering half hemisphere. Green lines denote cell boundaries, red ones belong to false-positives. **B** Human annotators use different features for cell sorting: their Ca^{2+} activity time traces, neuronal spatial footprint, and the movie snapshots during Ca^{2+} events. **C-D** The traditional first-to-last cell sorting workflow (**C**), with classifier assistance (**D**). **E** Engineered new features captured spatiotemporal information used by an expert annotator. **F** Absolute discriminability index for traditional and new engineered features. Data points from 15 different annotations (12 human annotators and 3 augmented consensus labels obtained via the majority vote across all annotators, where ties are broken with coin flips). **G** New features increase classifiers' rejection accuracy of false-positive, as quantified by a two-sided Wilcoxon signed-rank test ($^{***}p < 10^{-3}$).

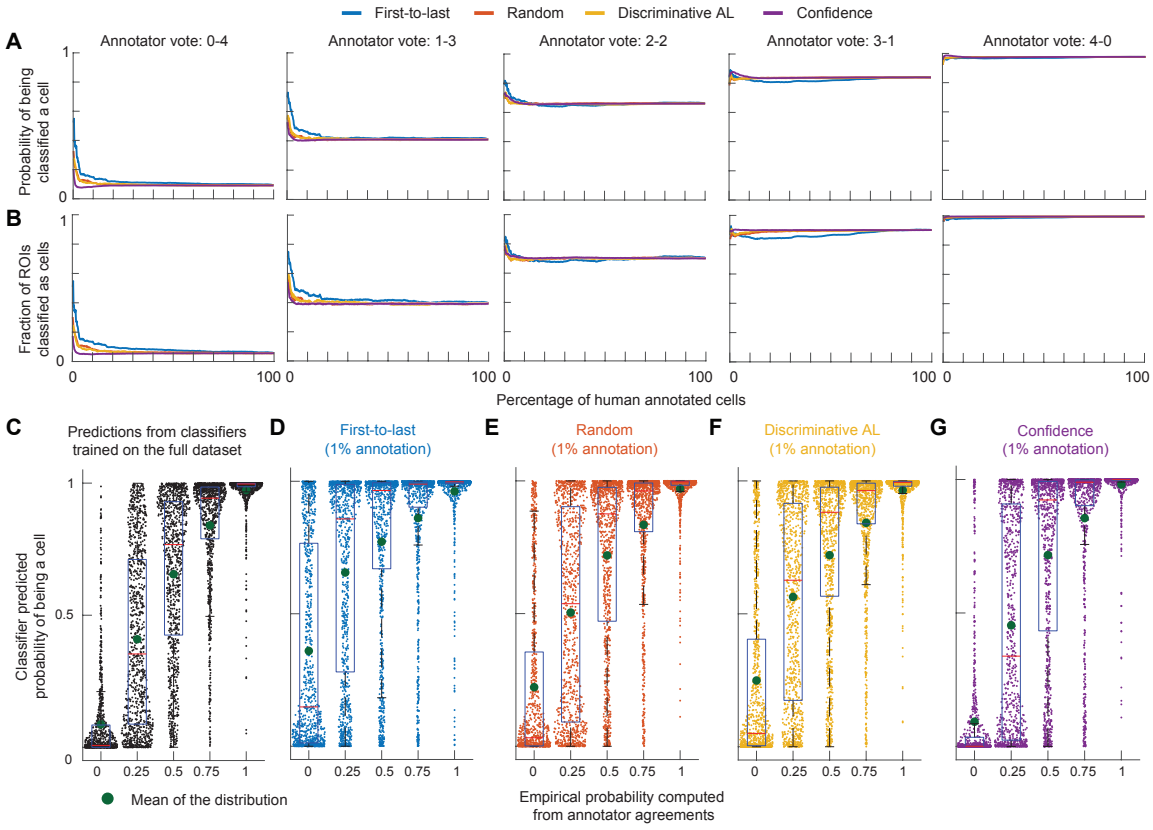

Figure S2: **ActSort predictions rapidly become empirically aligned probabilities.** **A** Probability and **B** fraction of ROIs being a cell across classifiers based on the uncertainty across annotators. Solid lines, mean; shaded area, standard errors of the mean across 15 annotations. **C-G** Empirical vs classifier predicted probabilities of a sample being classified as a cell computed from annotator agreements (as discrete values: $0, 0.25, 0.5, 0.75, 1$). Each dot, single cell. **C** Classifiers trained on the full dataset. **D** First-to-last, **E** random, **F** DAL, and **G** CAL classifiers with 1% annotated cells. CAL, but not others, achieves similar probability distributions as the fully trained classifier, despite being trained on only 1% of data. Red lines, median; green dots, mean of the corresponding distribution.

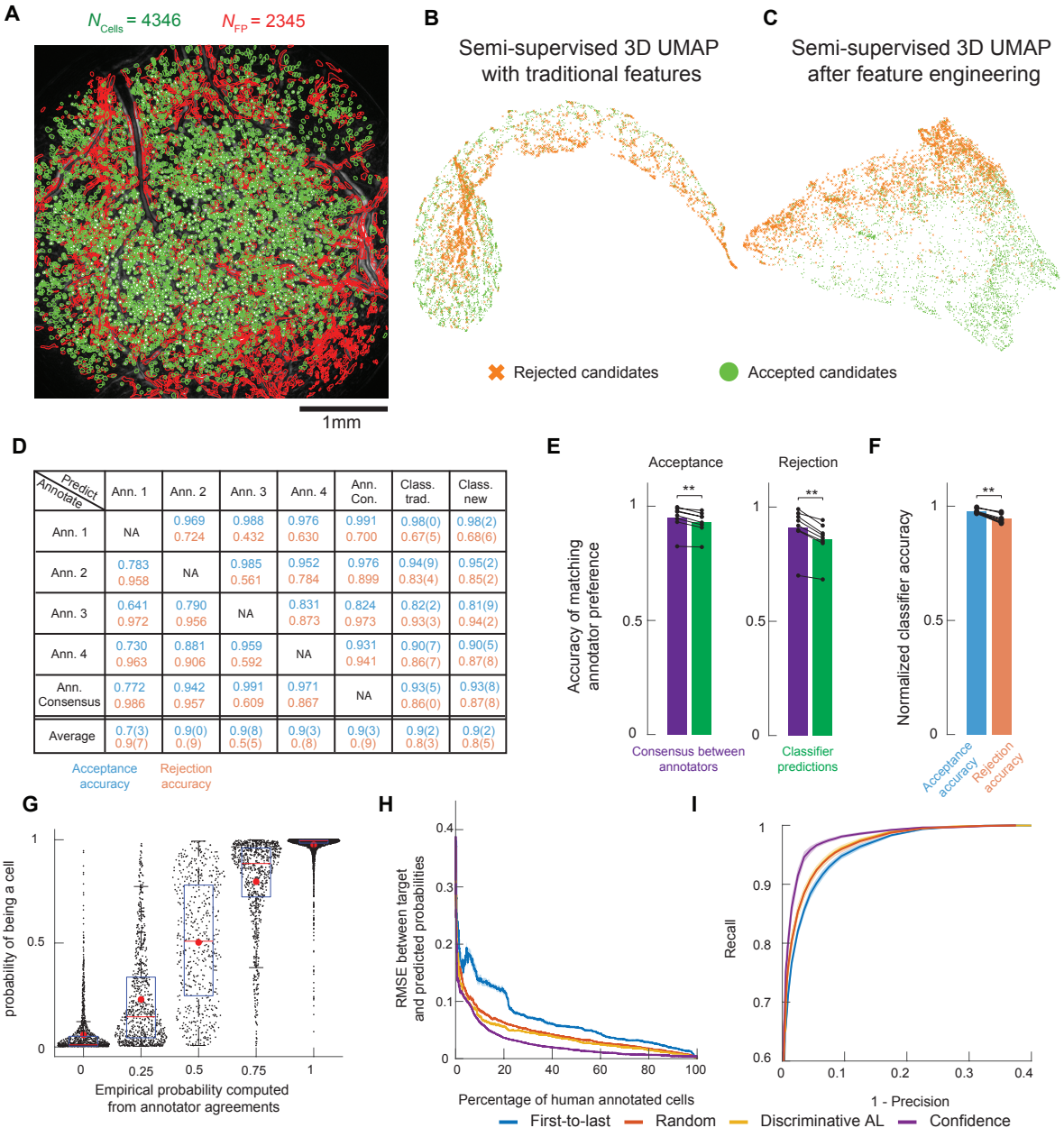

Figure S3: **ActSort improves cell sorting quality in movies with many false-positives. A** Example cell map from 4*mm* field of view including eight neocortical brain regions Ebrahimi et al. (2022). Green lines, cell boundaries; red lines, false-positives.**B, C** Semi-supervised 3D UMAP plots from 600 labeled random samples out of 6,691 cells using the traditional (**B**) or the new (**C**) feature sets, illustrating the separability of true and false positives. (Continued in the next page)

Figure S3: (Continued from the previous page) **D** Table, similar to Fig. S1, quantifying the consistency among the annotators in both accepting and rejecting cells. **E** Accuracy of matching annotator labels for both true positives and true negatives. Purple, consensus between annotators, four new annotators were augmented by taking the majority vote in combination with three annotators; green, classifier predictions to match the preferences of the eight annotators. Classifier predictions were consistent with the human annotators, almost matching the consensus labels. **F** Normalized classifier accuracy. **G** Empirical probabilities vs. predicted test probabilities from classifiers trained on half of the dataset, showing the probability of a sample being classified as a cell. Red lines, median. **H** RMSE between predicted probabilities by the current and the final classifiers as a function of percentage of human-annotated cells. **I** Precision-recall curve obtained from classifiers trained on only 600 out of 6,691 cells. Solid lines, mean; shaded areas, standard errors of the mean across 9 annotations. Statistical comparisons between classifiers were two-sided Wilcoxon signed-rank tests. ($^{*}p < 0.05$, $^{**}p < 10^{-2}$, $^{***}p < 10^{-3}$).

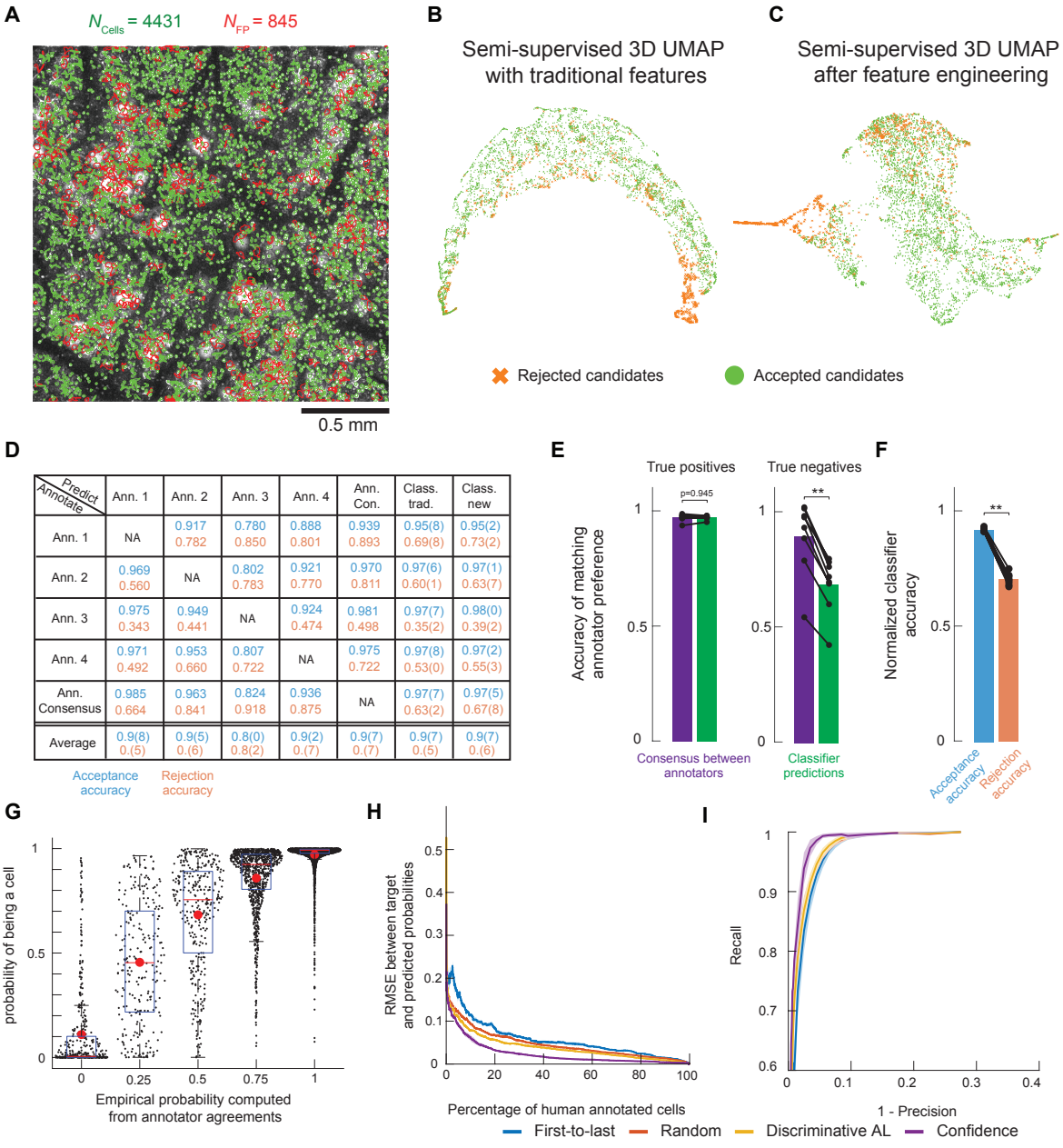

Figure S4: **ActSort improves cell sorting quality when the imaging movie has remaining residual motion.** Same as in Figure S3, but for the dataset acquired with our two-photon microscope (**Methods**). This dataset had residual motion, which led to several false positives in the form of duplicates. Hence, ActSort predicted probabilities were relatively less consistent with the empirical annotation, though ActSort still converged rapidly and outperformed traditional approaches.

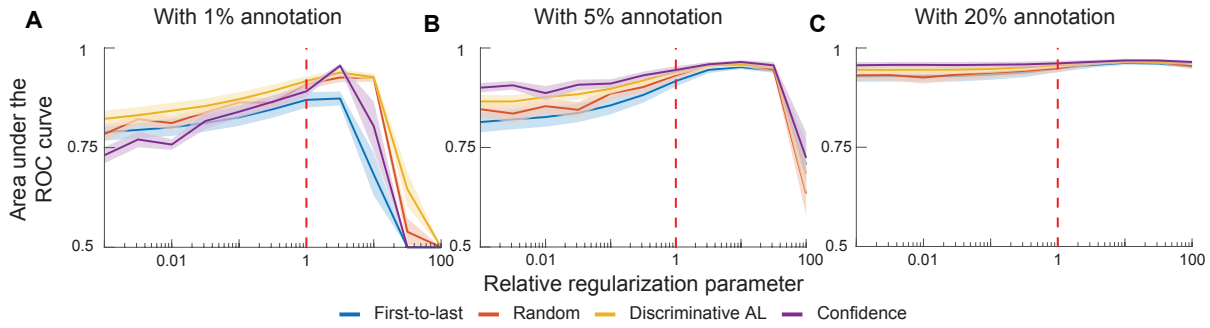

Figure S5: **The insensitivity to the regularization strength allows rapid training without cross-validation.** Our emphasis on real-time training necessitated using a pre-defined regularization strength, which was taken to be $\lambda_{\text{default}} = \#\text{ of samples}^{-1}$. To validate this, we performed the active learning experiments of Figure 2 with varying levels of regularization parameters. $\lambda_{\text{relative}} = 1$ corresponds to the default value for the problem of interest. We observed that undershooting the optimal regularization strength had negligible effect, whereas overshooting had the potential to harm training altogether.