Reproducibility of in-vivo electrophysiological measurements in mice

International Brain Laboratory*, Kush Banga[7], Julius Benson[11], Jai Bhagat[7], Dan Biderman[3], Daniel Birman[4], Niccolò Bonacchi[2], Sebastian A Bruijns[1], Robert AA Campbell[13], Matteo Carandini[7], Gaëlle A Chapuis[5], Anne K Churchland[6], M Felicia Davatolhagh[6], Hyun Dong Lee[3], Mayo Faulkner[7], Berk Gerçek[5], Fei Hu[9], Julia Huntenburg[2], Cole Hurwitz[3], Anup Khanal[6], Christopher Krasniak[10], Guido T Meijer[2], Nathaniel J Miska[7], Zeinab Mohammadi[12], Jean-Paul Noel[11], Liam Paninski[3], Alejandro Pan-Vazquez[12], Noam Roth[4], Michael Schartner[2], Karolina Socha[7], Nicholas A Steinmetz[4], Karel Svoboda[14], Marsa Taheri[6], Anne E Urai[8], Miles Wells[7], Steven J West[7], Matthew R Whiteway[3], Olivier Winter[2], Ilana B Witten[12]

[1]Max-Planck-Institute, Tübingen, Germany; [2]Champalimaud Foundation, Lisbon, Portugal; [3]Columbia University, NY, USA; [4]University of Washington, WA, USA; [5]University of Geneva, Switzerland; [6]University of California Los Angeles, USA; [7]University College London, UK; [8]Leiden University, The Netherlands; [9]University of California, Berkeley, USA; [10]Cold Spring Harbor Laboratory, NY, USA; [11]New York University, NY, USA; [12]Princeton University, NJ, USA; [13]Sainsbury Wellcome Center, London, UK; [14]Allen Institute for Neural Dynamics WA, USA

Abstract

Understanding whole-brain-scale electrophysiological recordings relies on the collective work of many labs. Because two labs recording from the same region can reach different conclusions, it is critical to quantify and control features that hinder reproducibility. To address this, we formed a multi-lab collaboration using a shared, open-source behavioral task and experimental apparatus. Experimenters in ten laboratories repeatedly targeted Neuropixels probes to the same location (spanning secondary visual areas, hippocampus, and thalamus) in mice making decisions. After applying quality-control criteria, we found that neuronal yield, firing rates, spike amplitudes, and task-modulated neuronal activity were largely reproducible across laboratories. To quantify variance in neural activity explained by task variables, we developed a multi-task neural network model, and found that within- and between-lab random effects captured by this model were comparable. Our results demonstrate that across-lab standardization can produce reproducible results from large-scale Neuropixels recordings. Our dataset, code, and protocols are openly accessible.

Introduction

Reproducibility is a cornerstone of the scientific method: a given sequence of experimental methods should lead to comparable results if applied in different laboratories. In some areas of biological and psychological science, however, the reliable generation of reproducible results is a well-known challenge (*Baker, 2016*; *Voelkl et al., 2020*; *Li et al., 2021*; *Errington et al., 2021*). In systems neuroscience at the level of single-cell-resolution recordings, evaluating reproducibility is difficult: experimental methods are sufficiently complex that replicating experiments is technically challenging, and many experimenters feel little incentive to do such experiments since negative results can be difficult to publish. Unfortunately, variability in experimental outcomes has been well-documented on a number of occasions. These include the existence and nature of "preplay" (*Dragoi and Tonegawa, 2011*; *Silva et al., 2015*; *Ólafsdóttir et al., 2015*; *Grosmark and Buzsáki, 2016*; *Liu et al., 2019*), the persistence of place fields in the absence of visual inputs (*Hafting et al., 2005*; *Barry et al., 2012*; *Chen et al., 2016*; *Waaga et al., 2022*), and the existence of spike-timing dependent plasticity (STDP) in nematodes (*Zhang et al., 1998*; *Tsui et al., 2010*). In the latter example, variability in experimental results arose from whether the nematode being studied was pigmented or albino, an experimental feature that was not originally known to be relevant to STDP. This highlights that understanding the source of experimental variability can facilitate efforts to improve reproducibility.

For electrophysiological recordings, several efforts are currently underway to document this variability and reduce it through standardization of methods (*de Vries et al., 2020*; *Siegle et al., 2021*). These efforts are promising, in that they suggest that when

approaches are standardized and results undergo quality control, observations conducted within a single organization can be reassuringly reproducible. However, this leaves unanswered whether observations made in separate, individual laboratories are reproducible when they likewise use standardization and quality control. Answering this question is critical since most neuroscience data is collected within small, individual laboratories rather than large-scale organizations. A high level of reproducibility of results across laboratories when procedures are carefully matched is a prerequisite to reproducibility in the more common scenario in which two investigators approach the same high-level question with slightly different experimental protocols. Therefore, establishing the extent to which observations are replicable even under carefully controlled conditions is critical to provide an upper bound on the expected level of reproducibility of findings in the literature more generally.

We have previously addressed the issue of reproducibility in the context of mouse psychophysical behavior, by training 140 mice in 7 laboratories and comparing their learning rates, speed, and accuracy in a simple binary visually-driven decision task. We demonstrated that standardized protocols can lead to highly reproducible behavior (*The International Brain Laboratory et al., 2021*). Here, we build on those results by measuring within- and across-lab variability in the context of intra-cerebral electrophysiological recordings. We repeatedly inserted Neuropixels multi-electrode probes (*Jun et al., 2017*) targeting the same brain regions (including posterior parietal cortex, hippocampus, and thalamus) in mice performing an established decision-making task (*The International Brain Laboratory et al., 2021*). We gathered data across ten different labs and developed a common histological and data processing pipeline to analyze the resulting large datasets. This pipeline included stringent new histological, and electrophysiological quality-control criteria (the "Recording Inclusion metrics and Guidelines for Optimal Reproducibility", or RIGOR) that are applicable to datasets beyond our own.

We define reproducibility as a lack of systematic across-lab differences: that is, the distribution of within-lab observations is comparable to the distribution of across-lab observations, and thus a data analyst would be unable to determine in which lab a particular observation was measured. This definition takes into account the natural variability in electrophysiological results. After applying the RIGOR quality control measures, we found that features such as neuronal yield, firing rate, and normalized LFP power were reproducible across laboratories according to this definition. Similarly, the proportions of cells modulated by decision-making variables (such as the sensory stimulus or the choice) was largely reproducible across labs. Finally, to quantify variance in neural activity explained by task variables (e.g., stimulus onset time), behavioral variables (timing of licks/paw movements), and other variables (e.g., spatial location in the brain or the lab ID), we developed a multi-task neural network encoding model that extends common, simpler regression approaches by allowing nonlinear interactions between variables. Again, we found that within-lab random effects captured by this model were comparable to between-lab random effects. Taken together, these results suggest that across-lab standardization of electrophysiological procedures can lead to reproducible results across laboratories.

Results

Neuropixels recordings during decision-making target the same brain location

To quantify reproducibility across electrophysiological recordings, we set out to establish standardized procedures across the International Brain Laboratory (IBL) and to test whether this standardization led to reproducible results. Ten IBL labs collected Neuropixels recordings from one repeated site, targeting the same stereotaxic coordinates, during a standardized decision-making task in which head-fixed mice reported the perceived position of a visual grating (*The International Brain Laboratory et al., 2021*). The experimental pipeline was standardized across labs, including surgical methods, behavioral training, recording procedures, histology, and data processing (Figure 1a, b); see Methods for full details. Neuropixels probes were selected as the recording device for this study due to their standardized industrial production, and their ability to sample many neurons in each of multiple brain regions simultaneously. Further, the commercial availability and popularity of Neuropixels probes made them an attractive alternative to probes that must be made in-house (*Shin et al., 2019*) or currently have limited availability (*Zhao et al., 2022*; *Chung et al., 2019*). In each experiment, Neuropixels 1.0 probes were inserted, targeted at -2.0 mm AP, -2.24 mm ML, 4.0 mm DV relative to bregma; 15° angle (Figure 1c). This site was selected because it encompasses brain regions implicated in visual decision-making, including visual area A (*Najafi et al., 2020*; *Harvey et al., 2012*), dentate gyrus, CA1, (*Turk-Browne, 2019*),

and thalamic nuclei LP and PO (*Saalmann and Kastner, 2011*; *Roth et al., 2016*).

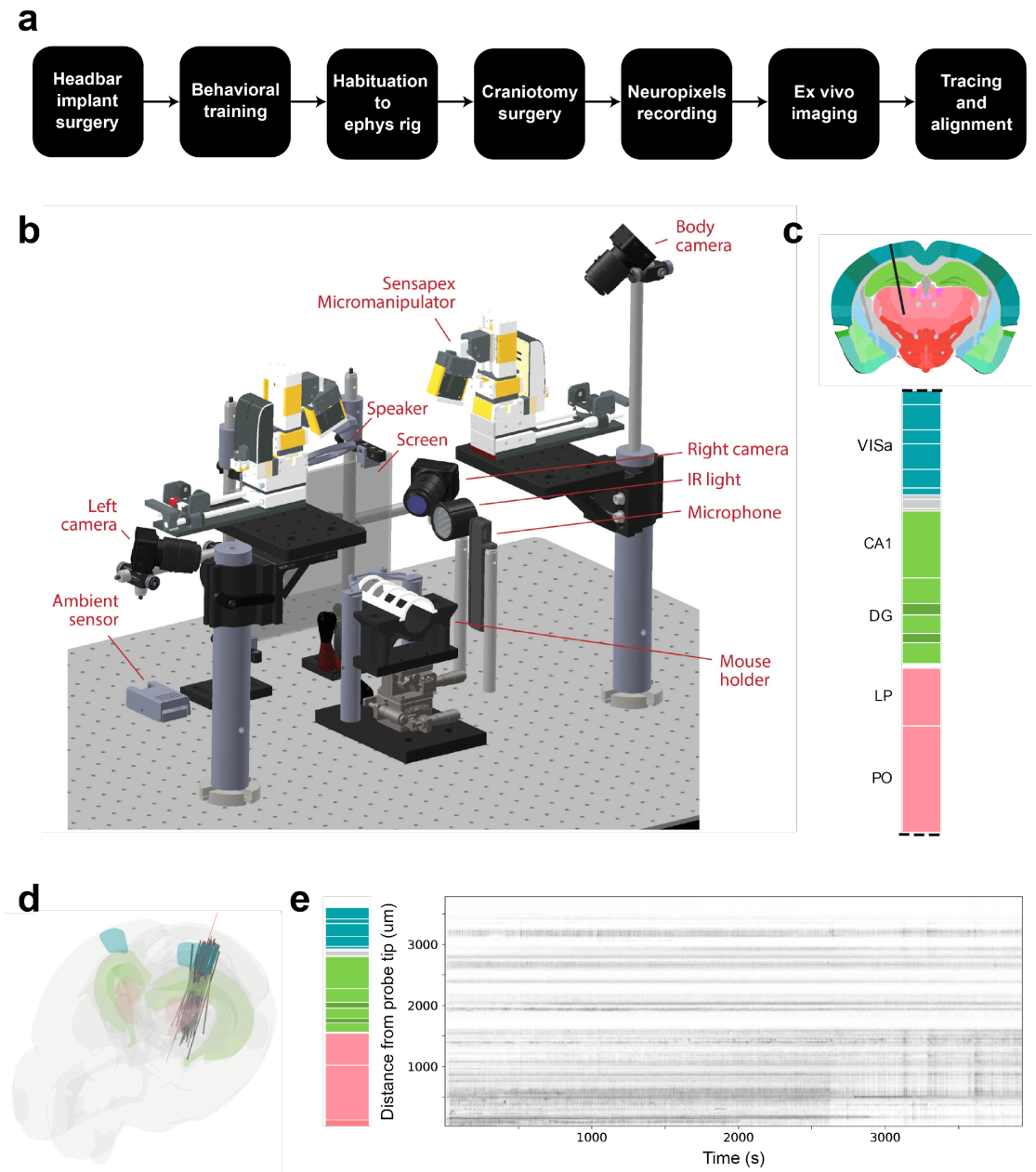

Figure 1. Standardized experimental pipeline and apparatus; location of the repeated site. a, The pipeline for electrophysiology experiments. **b**, Drawing of the experimental apparatus. **c**, Location and brain regions of the repeated site. VISa: Visual Area A; CA1: Hippocampal Field CA1; DG: Dentate Gyrus; LP: Lateral Posterior nucleus of the thalamus; PO: Posterior Nucleus of the Thalamus. **d**, Acquired repeated site trajectories shown within a 3D brain schematic. Target trajectory shown in red. **e**, Raster plot from one example session.

Figure 1–Figure supplement 1. Detailed experimental pipeline for the Neuropixels experiment.
Figure 1–Figure supplement 2. Spiking activity qualitatively appears heterogeneous across recordings.
Figure 1–Figure supplement 3. Electrophysiology data quality examples.

Stereotaxic probe placement limits resolution of probe targeting

As a first test of experimental reproducibility, we assessed variability in Neuropixels probe placement around the planned repeated site location. Brains were perfusion-fixed, dissected, and imaged using serial section 2-photon microscopy for 3D reconstruction of probes (Figure 2a). Whole brain auto-fluorescence data was aligned to the Allen Common Coordinate Framework (CCF) (*Wang et al., 2020*) using an elastix-based pipeline (*Klein et al., 2010*) adapted for mouse brain registration (*West, 2021*). cm-DiI labelled probe tracks were manually traced in the 3D volume (Figure 2b; supp. 1). Trajectories obtained from our stereotaxic system and traced histology were then compared to the planned trajectory. To measure probe track variability, each traced probe track was linearly interpolated to produce a straight line insertion in CCF space (Figure 2c).

We first compared the micro-manipulator brain surface coordinate to the planned trajectory to assess variance due to targeting strategies only (targeting variability, Figure 2d). Targeting variability occurs when experimenters must move probes slightly from the planned location to avoid blood vessels or irregularities. These slight movements of the probes led to a degree of variance from the planned insertion site (Figure 2d, total mean displacement = 103.7 µm).

We tested for systematic differences in targeting variability across labs via permutation testing. As a metric to quantify across lab variability, we consider the differences between the cumulative distribution function (CDF) of individual lab displacements to the CDF of all other labs. We took the largest of these deviations as the overall measure for variability (see Methods for details). To generate a null distribution, we computed this maximum deviation for data in which the lab labels were shuffled over mice. From this, we generated a p-value from the value in the null distribution corresponding to the observed deviation (p=0.10, Figure 2g). This provides reassurance that there were not systematic, lab-to-lab differences in targeting variability.

Geometrical variability, obtained by calculating the difference between planned and final identified probe position acquired from the reconstructed histology, encompasses targeting variance, anatomical differences, and errors in defining the stereotaxic coordinate system. Geometrical variability was more extensive (Figure 2e and h, total mean displacement = 336.0 µm). Assessing geometrical variability for all probes with permutation testing revealed a p-value of 0.23 across laboratories (Figure 2h), arguing, again, that systematic lab-to-lab differences don't account for the observed variability.

To determine the extent that anatomical differences drive this geometrical variability, we regressed histology-to-planned probe insertion distance at the brain surface against estimates of animal brain size. Regression against both animal body weight and estimated brain volume from histological reconstructions showed no correlation to probe displacement ($R^2 < 0.03$), suggesting differences between CCF and mouse brain sizes are not the major cause of variance. An alternative explanation is that geometrical variance in probe placement at the brain surface is driven by inaccuracies in defining the stereotaxic coordinate system, including discrepancies between skull landmarks and the underlying brain structures.

Accurate placement of probes in deeper brain regions is critically dependent on probe angle. We assessed probe angle variability by comparing the final histologically-reconstructed probe angle to the planned trajectory. We observed a consistent mean displacement from the planned angle in both medio-lateral (ML) and anterior-posterior (AP) angles (Figure 2f and i, total mean difference in angle from planned: 7.3 degrees). Angle differences can be explained by the different orientations and geometries of the CCF and the stereotaxic coordinate systems. The difference in histology angle to planned probe placement was assessed with permutation testing across labs, and shows a p-value of 0.56 (Figure 2i). This suggests there are no systematic differences in insertion angle across labs.

In conclusion, insertion variance, in particular geometrical variability, is sizeable enough to impact probe targeting to desired brain regions and could serve as a hindrance to reproducibility. We were unable to identify a prescriptive analysis to predict probe placement accuracy, which may reflect that the major driver of probe placement variance derives from differences in skull landmarks used for establishing the coordinate system and the underlying brain structures. Our data suggest the resolution of probe insertion targeting on the brain surface to be approximately 370 µm, which must be taken into account when planning probe insertions.

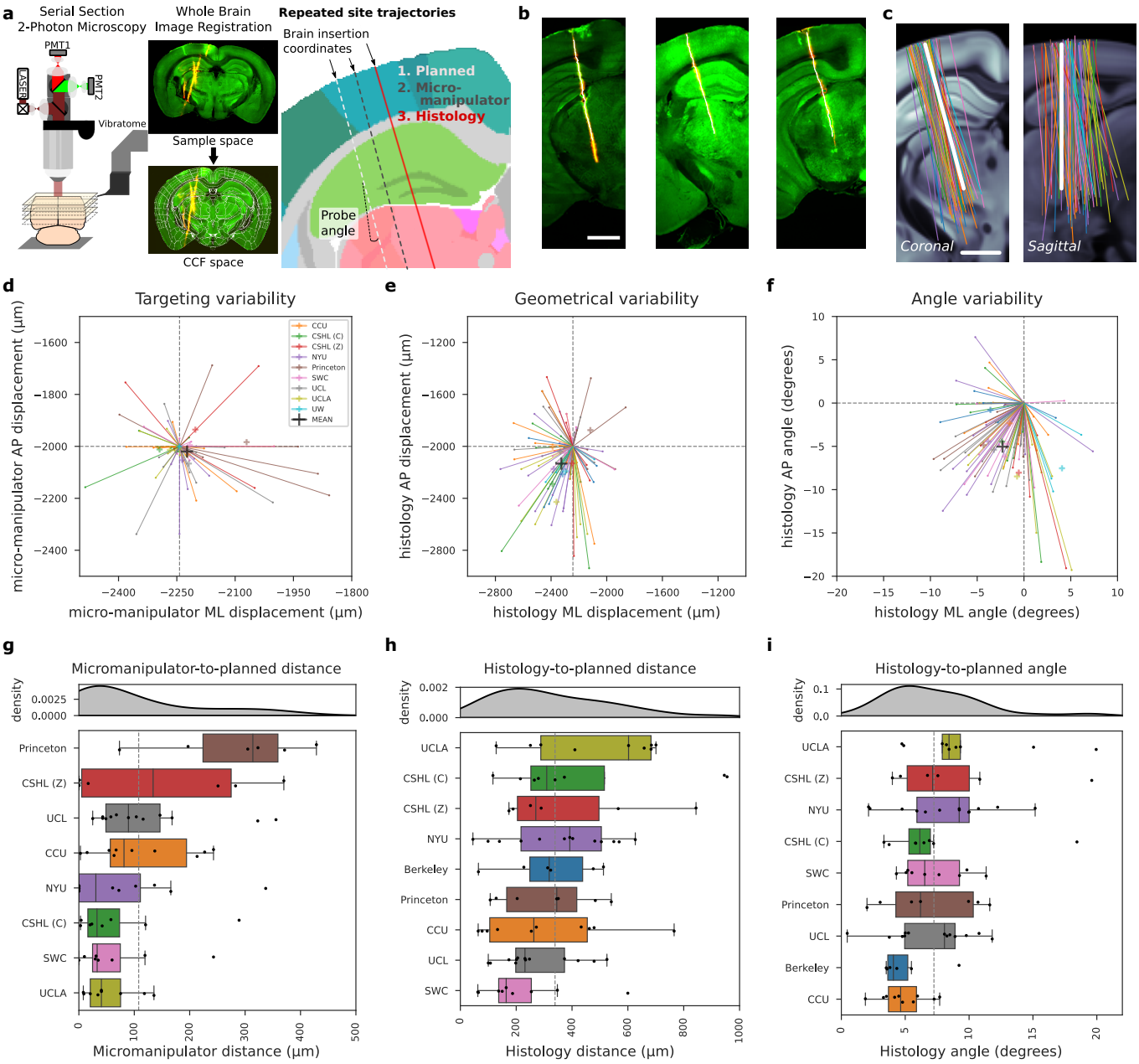

Figure 2. Histological reconstruction reveals resolution limit of probe targeting. a, Histology pipeline for electrode probe track reconstruction and assessment. Three separate trajectories are defined per probe: planned, micro-manipulator (based on the experimenter's stereotaxic coordinates) and histology (interpolated from tracks traced in the histology data). **b**, Example tilted slices through the histology reconstructions showing the repeated site probe track. Plots show the green auto-fluorescence data used for CCF registration and red cm-DiI signal used to mark the probe track. White dots show the projections of channel positions onto each tilted slice. Scale bar: 1mm. **c**, Histology probe trajectories, interpolated from traced probe tracks, plotted as 2D projections in coronal and sagittal planes, tilted along the repeated site trajectory over the allen CCF, colors: laboratory. Scale bar: 1mm. **d, e, f**, Scatterplots showing variability of probe placement from planned to: micro-manipulator brain surface insertion coordinate (d, targeting variability, N=83), histology brain surface insertion coordinate (e, geometrical variability, N=89), and histology probe angle (f, angle variability, N=89). Each line and point indicates the displacement from the planned geometry for each insertion in anterior-posterior (AP) and mediolateral (ML) planes, color-coded by institution. **g, h, i**, Assessment of probe displacement by institution from planned to: micro-manipulator brain surface insertion coordinate (g), histology brain surface insertion coordinate (h), and histology probe angle (i). Kernel density estimate plots (top) are followed by boxplots (bottom) for probe displacement, ordered by descending median value. *Legend continues on next page.*

Figure 2 *(previous page)*. A minimum of four data points per institution led to their inclusion in the plot and subsequent analysis. Dashed vertical lines display the mean displacement across institutions, indicated in the respective scatterplot in (d, e, f).

Figure 2–Figure supplement 1. Tilted slices along the histology insertion for all insertions used in assessing probe placement.

Electrophysiological features are reproducible across laboratories in cortex and thalamus

Having established that targeting of Neuropixels probes to the desired target location was a source of substantial variability, we next measured the variability and reproducibility of electrophysiological features recorded by the probes. We implemented eleven exclusion criteria. Two of these were specific to our experiments: a behavior criterion where the mouse completed at least 400 trials of the behavioral task, and a session number criterion for analyses that directly compared across labs (permutation tests; Fig 3d-f, 4e, 5). The remaining nine criteria, which we collectively refer to as the "Recording Inclusion metrics and Guidelines for Optimal Reproducibility" (RIGOR; Table 1), were more general and could be applied widely to electrophysiology experiments: a yield criterion, a noise criterion, qualitative criteria for visual assessment (lack of drift, epileptiform activity, noisy channels and artifacts, see Figure 1-supplement 3 for examples), and single unit metrics (refractory period violation, amplitude cutoff, and median of the amplitudes).

We recorded a total of 96 sessions targeted at our planned repeated site (Figure 3a). Of these, 12 sessions were excluded due to incomplete data acquisition caused by a hardware failure during the experiment (7) or because we were unable to complete histology on the subject (5). Next, we applied exclusion criteria to the remaining complete datasets. We first applied the RIGOR standards described in Table 1. Upon manual inspection, we observed 1 recording fail due to drift, 8 recordings fail due to noisy channels, 2 recordings fail due to artefacts, and 1 recording fail due to epileptiform activity. 13 recordings failed our criterion for low yield, and 3 recordings failed our criterion for noise level. Next, we applied criteria specific to our behavioral experiments, and found that 2 recordings failed our behavior criterion by not meeting the minimum of 400 trials completed). Some of our analysis also required further (stricter) inclusion criteria (see Methods).

When plotting all recordings, including those that failed to meet quality control criteria, one can observe that discarded sessions were clear outliers (Figure 3-supplemental 1). In subsequent figures, only recordings that passed these quality control criteria were included. Overall, we analyzed data from the 54 remaining sessions recorded in ten labs to determine the reproducibility of our electrophysiological recordings. The responses of 3013 single neurons (all passing the metrics defined in Table 1) are analyzed below; this total reflects an average of 61 ± 44 [mean $\pm$ std] per insertion.

We then evaluated whether electrophysiological features of these neurons, such as firing rates and LFP power, were reproducible across laboratories. In other words, is there consistent variation across laboratories in these features that is larger than expected by chance? We first visualized LFP power, a feature used by experimenters to guide the alignment of the probe position to brain regions, for all the repeated site recordings (Figure 3b). The dentate gyrus (DG) is characterized by high power spectral density of the LFP (*Penttonen et al., 1997*; *Bragin et al., 1995*; *Senzai and Buzsáki, 2017*) and this feature was used to guide physiology-to-histology alignment of probe positions (Figure 3-supplementary 2). By plotting the LFP power of all recordings along the length of the probe side-by-side, aligned to the boundary between the DG and thalamus, we confirmed that this band of elevated LFP power was visible in all recordings at the same depth. The variability in the extent of the band of elevated LFP power in DG was due to the fact that the DG has variable thickness and due to targeting variability not every insertion passed through the DG at the same point in the sagittal plane (Figure 3-supplemental 2).

The probe alignment allowed us to attribute the channels of each probe to their corresponding brain regions to investigate the reproducibility of electrophysiological features for each of the target regions of the repeated site. To visualize all the neuronal data, each neuron was plotted at the depth it was recorded overlaid with the position of the target brain region locations (Figure 3c). From these visualizations it is apparent that there is recording-to-recording variability. This raises two questions: (1) Is the recording-to-recording variability within an animal the same or different compared to across animals? (2) Is the recording-to-recording variability lab dependent?

To answer the first question we performed several bilateral recordings in which the same insertion was targeted in both hemispheres, as mirror images of one another. This allowed us to quantify the variability between insertions within the same

		Criterion	Definition
Whole recording	Computed	Yield	At least 0.1 neurons (that pass single unit criteria) per electrode channel in each region.
		Noise level	AP band: Median action-potential band RMS (AP RMS) less than 40 uV (computed post-destriping) LF band: Median LFP power less than -140 dB (may differ for other electrodes) (computed pre-destriping).
	Visually assessed	Drift	Absence of pronounced instability of recording ("drift"), as observed on the raster plot.
		Epileptiform activity	Absence of epileptiform activity, which is characterized by sharp discontinuities on the raster plot (not driven by movement or noise artifacts) or strong periodic spiking spanning many channels.
		Noisy channels	Absence of noisy or poor impedance channels groups (e.g., lack of visible action potential on the raw data plot).
		Artefacts	Absence of artefacts, which are characterised by a sudden discontinuity in the raw signal, spanning nearly all channels at once.
Single unit	Computed	Refractory period violation	Each neuron must pass a sliding refractory period metric, a false positive estimate which computes the confidence that a neuron has below 10% contamination for all possible refractory period lengths (from 0.5 to 10 ms). A neuron passes if the confidence metric is greater than 90% for any possible refractory period length.
		Amplitude cutoff	Each neuron must pass a metric that estimates the number of spikes missing (false negative rate) and ensures that the distribution of spike amplitudes is not cut off, without a Gaussian assumption.
		Median amplitude	Each neuron must have a median amplitude greater than 50 uV.

Table 1. Recording Inclusion metrics and Guidelines for Optimal Reproducibility (RIGOR). These criteria could be applied to other electrophysiology experiments to enhance reproducibility. Note that whole recording metrics are most relevant to large scale (>20 channel) recordings. For those recordings, manual inspection of datasets passing these criteria will further enhance data quality.

animal and compare this variability to the across-animal variability (Figure 3-supplemental 3). We found that whether within- or across-animal variance was larger depended on the electrophysiological feature in question (Figure 3-supplemental 3f). For example, for neuronal yield across variance was larger compared to within variance whereas for action-potential band root mean square (AP band RMS), the within variance was often the largest.

Is the recording-to-recording variability lab dependent? To answer this question, the reproducibility of electrophysiological features over laboratories was investigated using a permutation testing approach. The tested features were neuronal yield, firing rate, spike amplitude, LFP power, and AP band RMS and were calculated per brain region (Figure 3-supplemental 4). As was the case when analysing probe placement variability, the permutation test assesses whether the distribution over features varies significantly across labs. When correcting for multiple testing, systematic corrections (like a Bonferroni correction) proved too strong in light of our power analysis, instead we opted to use a slightly more stringent alpha of 0.01 when establishing significance. The permutation test revealed a significant effect for the AP band RMS values and neuron yield in CA1, the former presumably reflects 50/60 Hz line noise which is likely to be rig-dependent. Otherwise, we found that all electrophysiological features were reproducible across laboratories for all regions studied (Figure 3d).

The permutation testing approach tested the reproducibility of each electrophysiological feature separately. It could also be the case, however, that some combination of these features varied systematically across laboratories. To test whether this was the case we trained a Random Forest classifier to try to predict in which lab a recording was made, based on the electrophysiological markers. The decoding was performed separately for each brain region because of their distinct physiological signatures. A null distribution was generated by shuffling the lab labels and decoding again using the shuffled labels (500 iterations). The significance of the classifier was defined as the fraction of times the accuracy of the decoding of the shuffled labels was higher compared to the original labels. To validate the decoder we first decoded brain region instead of lab identity from the electrophysiological features; the decoder was able to decode brain region with very high accuracy (Figure 3e, left). The classifier could only decode lab identity from the dentate gyrus and not from any of the other regions above chance, indicating that the electrophysiological features were reproducible across laboratories for these regions (Figure 3e, right). Importantly, when including all recordings, regardless of QC status, the classifier was able to decode lab identity from 4/5 regions (Figure 3f). This indicates that our QC criteria were successful in reducing lab-to-lab variability.

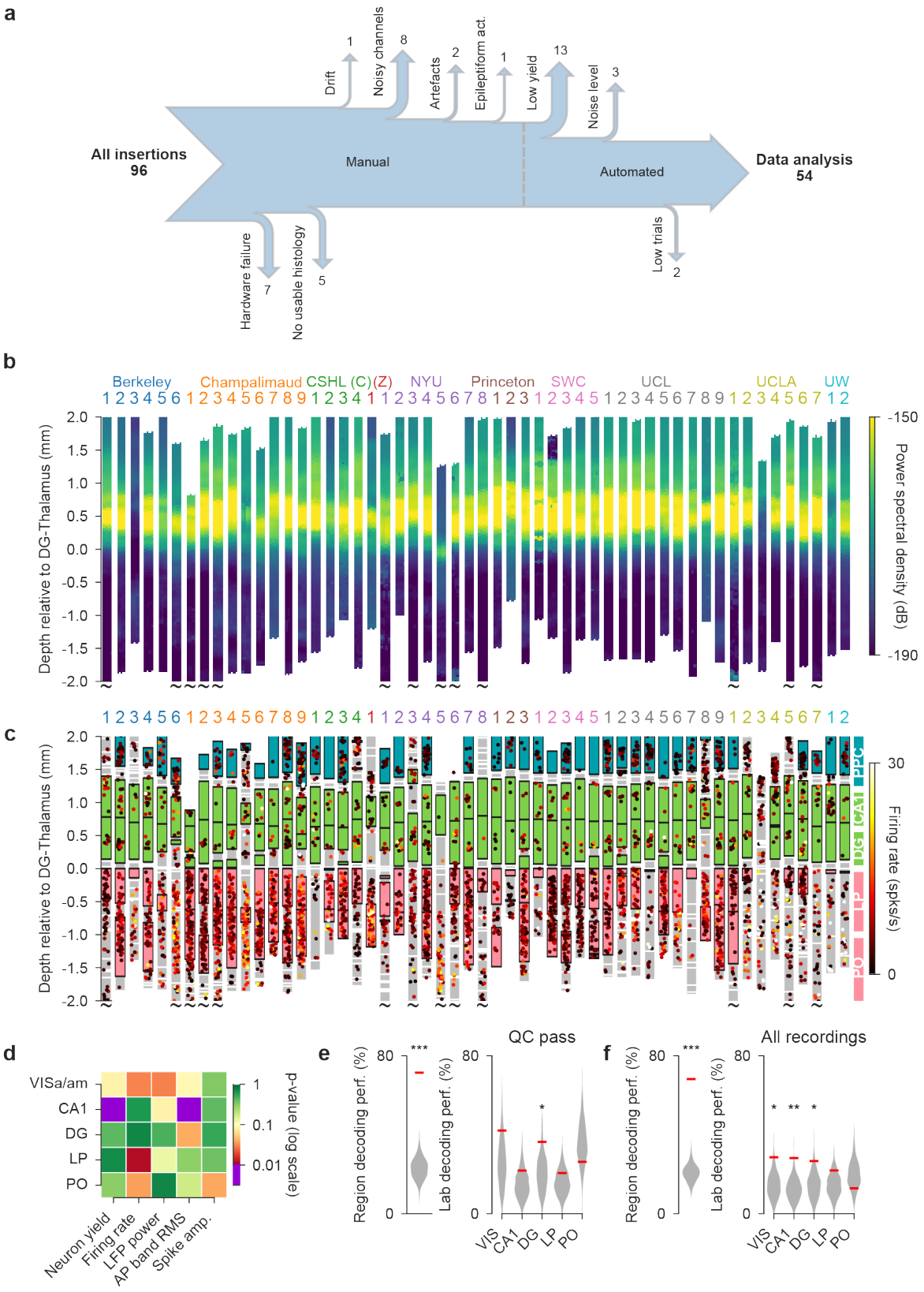

Figure 3. Electrophysiological features are reproducible across laboratories in cortex and thalamus. (a) Number of experimental sessions recorded; number of sessions used in analysis due to exclusion criteria. Up arrows indicate general-use RIGOR criteria on the quality of the electrophysiology, as presented in Table 1); down arrows indicate IBL-specific criteria applied. **(b)** Power spectral density between 20 and 80 Hz of each channel of each probe insertion (vertical columns) shows reproducible alignment of electrophysiological features to histology. Insertions are aligned to the boundary between the dentate gyrus and the thalamus. *Legend continues on next page.*

Figure 3 *(previous page)*. CSHL: Cold Spring Harbor Laboratory [(C): Churchland lab, (Z): Zador lab], NYU: New York University, SWC: Sainsbury Wellcome Centre, UCL: University College London, UCLA: University of California, Los Angeles, UW: University of Washington. **(c)** Firing rates of individual neurons according to the depth at which they were recorded. Colored blocks indicate the target brain regions of the repeated site, grey blocks indicate a brain region that was not one of the target regions. If no block is plotted, that part of the brain was not recorded by the probe because it was inserted too deep or too shallow. Each dot is a neuron, colors indicate firing rate. **(d)** P-values for five electrophysiological metrics, computed separately for all target regions, assessing the reproducibility of the distributions over these features across labs. P-values are plotted on a log-scale to visually emphasize values close to significance. **(e)** Trained on sessions that passed QC, a Random Forest classifier could successfully decode the brain region from five electrophysiological features (neuron yield, firing rate, LFP power, AP band RMS and spike amplitude), but could only decode lab identity from the dentate gyrus. The red line indicates the decoding accuracy and the grey violin plots indicate a null distribution obtained by shuffling the labels 500 times. The decoding of lab identity was performed per brain region. **(f)** Trained on all recording, regardless of QC level, the classifier could successfully decode lab identity from 4/5 brain regions. (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$)

Figure 3–Figure supplement 1. Recordings that didn't pass QC were visual outliers

Figure 3–Figure supplement 2. High LFP power in dentate gyrus was used to align probe locations in the brain.

Figure 3–Figure supplement 3. Bilateral recordings assess within- vs across-animal variance.

Figure 3–Figure supplement 4. Values used in the decoding analysis, per metric and per brain region.

Functional activity is broadly reproducible across laboratories

Concerns about reproducibility extend not only to electrophysiological properties, but also functional properties. To address this, we analyzed the reproducibility of the relationship between neural activity and task variables from the decision-making task. In particular, we were interested in whether the targeted brain regions have comparable neural responses to task events, such as stimulus onset, movement onset, and reward delivery. An inspection of individual neurons revealed clear modulation by, for instance, the onset of movement in a specific direction (Figure 4a). Neurons were variable in the extent to which they were modulated by task events (Figure 4b) (*Urai et al., 2022*). Plotting the session-averaged response for each experiment in a given area revealed that despite this variability, many key features were reproduced, such as the general response time course (Figure 4c and d; also Figure 5d).

Having observed that many individual neurons are modulated by task variables during decision-making, we examined the reproducibility of the proportion of modulated neurons. Within each brain region, we compared the proportion of the neural population that was sensitive to specific elements of the task (Figure 4e). We used six tests (Wilcoxon sign-rank tests and Wilcoxon rank-sum tests (*Steinmetz et al., 2019*)) to identify neurons with significantly modulated firing rates during candidate time windows (Figure 4e and Figure 4-supplemental 1). For most tests, the proportions of modulated neurons across sessions and across brain regions were quite variable (Figure 4e and Figure 4-supplemental 1).

To evaluate the reproducibility of task modulation quantitatively, we applied a permutation test to the proportion of modulated neurons by each task event, and the distribution of firing rate differences between the test-specific time-periods across labs (Figure 4g). We did not observe systematic lab-to-lab differences in the proportion of neurons modulated by task events (Figure 4h, top). This reproducibility seems at first reassuring, and indeed many papers use the proportion of responsive neurons as evidence that a particular area subserves a particular function. However, when we instead compared not the proportion of modulated neurons but the full distribution of firing rate modulations, reproducibility was more tenuous, and failed in some areas/tests (Figure 4h, bottom). Failures were driven by, for instance, outlier labs that had fewer low firing rate modulations. We propose that future studies report not only the proportion of modulated neurons, but also the distribution of firing rates observed so that functional activity can be more comprehensively compared.

To ensure that our measurements afforded sufficient power to detect differences across labs, we conducted a power analysis (Figure 4f,g). For this analysis we considered a particular modulation, such as the distribution of firing rate modulations by the stimulus onset. For each individual lab, we shifted the entirety of its datapoints either upwards or downwards, searching for the shift at which the modulation would register as significantly different from the other labs. This provides insight into how sensitive each test is to deviations within individual labs. For the majority of tests and regions, our tests were sufficiently

well-powered to detect shifts within the expected range, given the type and amount of data examined (an exception is explained below) (Figure 4-supplemental 2).

To further investigate how neural activity is modulated by decision-making, we measured the Fano Factor of single units. The Fano Factor, defined as the spike count variance over trials divided by spike count mean, enables the comparison of the fidelity of signals across neurons and regions, despite differences in firing rates (*Tolhurst et al., 1983*). We selected the period between 40-200 ms after movement onset (for correct trials with full-contrast stimuli on the right side) to calculate an average Fano Factor per neuron and quantify differences in Fano Factor across labs. This window was selected since the Fano Factor tends to be consistently low around movement time (*Churchland et al., 2010*, *2011*). We found no reliable difference across labs after applying a permutation test (Figure 4h). However, because the Fano Factor was more variable across neurons/sessions than other measures, our power analysis suggested limits in our ability to detect systematic differences among labs (Fig4-Figure Supplement 2). Therefore we report reproducibility of the Fano Factor across labs with some caution.

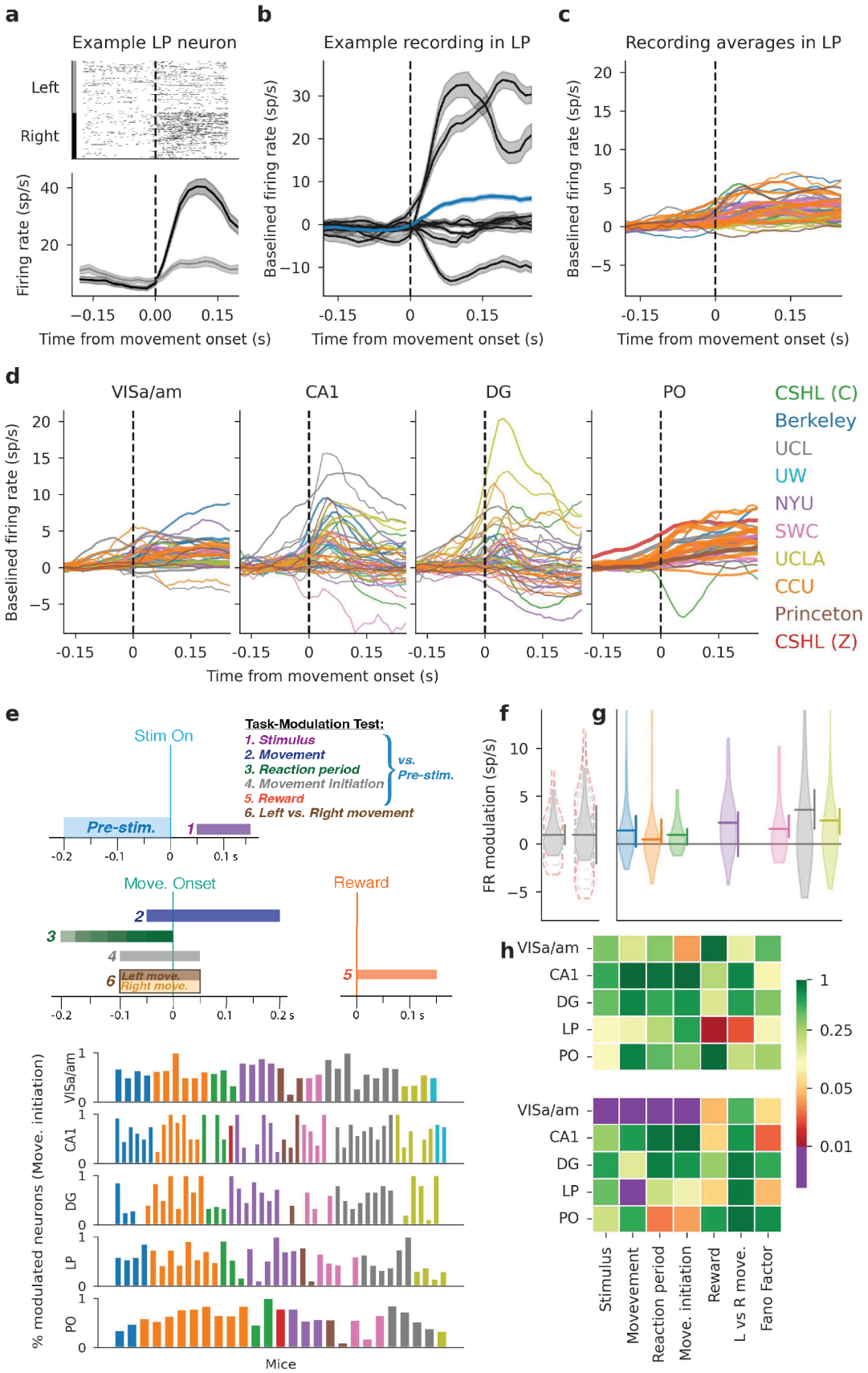

Figure 4. Neural activity is modulated during decision-making in five neural structures, with no significant difference between laboratories. (a) Raster plot (top) and firing rate time course (bottom) of an example neuron in LP, aligned to movement onset, split for correct left and right choices. *Legend continues on next page.*

Figure 4 *(previous page)*. This neuron was task-modulated using the following tests only: movement, movement initiation, and left vs. right movement. The firing rate is calculated using a causal sliding window; each time point includes a 60 ms window prior to the indicated point. **(b)** Peri-event time histograms (PETHs) of all LP neurons from a single mouse, aligned to movement onset (only correct choices in response to right-side stimuli are shown). These PETHs are baseline-subtracted by a pre-stimulus baseline. Shaded areas show standard error of mean (and propagated error for the overall mean). The thicker line shows the average over this entire population, colored by the lab from which the recording originates. **(c,d)** Average PETHs from individual mice across labs (similar to thick line in (b)) for LP (c) and the remaining four repeated site brain regions (d). Line thickness indicates the number of neurons in each recording (ranging from 4 to 86). **(e)** Schematic defining all six task-modulation tests (*top*) and proportion of task-modulated neurons for each mouse in each brain region for an example test (movement initiation) (*bottom*). Each column and color indicate, in order, a different recording session and lab. (Note that there is no correspondence here between columns across different brain regions.) **(f)**Two hypothetical distributions: first, when the test is sensitive, a small shift in the distribution is enough to make the test significant (non-significant shifts shown with broken line in grey, significant shift outlined in red). By contrast, when the test is less sensitive, the vertical line is large and a corresponding large range of possible shifts is present. The possible shifts we find usually cover only a small range. **(g)** Power analysis example for modulation by the stimulus in CA1. Violin plots: distributions of firing rate modulations for each lab; horizontal line: mean across sessions; vertical line at right: how much the distribution can shift up- or downwards before the test becomes significant. **(h)** Permutation test results for task-modulated activity and the Fano Factor. Top: tests based on proportion of modulated neurons; Bottom: tests based on the distribution of firing rate differences. Comparisons performed for correct trials with non-zero contrast stimuli.

Figure 4–Figure supplement 1. Proportion of task-modulated neurons, defined by six tests, across mice, labs, and brain regions.
Figure 4–Figure supplement 2. Power analysis of permutation tests

Neural response dynamics are similar across labs

The results above indicate that task-driven modulations in neural activity are reproducible across labs. However, the tests used thus far leave unanswered whether the dynamics of neural responses within each area are likewise reproducible. To test the reproducibility of dynamics, we first summarized the response for each neuron by computing peri-event time histograms (PETHs, Figure 5a). Because temporal dynamics may depend on reaction time, we generated separate PETHs for fast (< 0.15 s) and slow (> 0.15 s) reaction times. We concatenated the resulting vectors to obtain a more informative summary of each cell's average activity. The results (below) did not depend strongly on the details of the trial-splitting; for example, splitting trials by "left" vs "right" behavioral choice led to similar results.

Next, we projected these high-dimensional summary vectors into a low-dimensional "embedding" space using principal component analysis (PCA). This embedding captures the variability of the population while still allowing for easy visualization and further analysis. Specifically, we stack each cell's summary double-PETH vector (described above) into a matrix (containing the summary vectors for all cells across all sessions) and run PCA to obtain a low-rank approximation of this matrix (see Methods). The accuracy of reconstructions from the top two principal components (PCs) varied across cells (Figure 5a); PETHs for the majority of cells could be well-reconstructed with just 2 PCs (Figure 5b).

This simple embedding is sufficiently powerful to expose differences in brain regions (Figure 5c; e.g., PO and CA1 show displaced clusters, illustrating regional differences in response dynamics). Region-to-region differences are also visible in the region-averaged PETHs and cumulative distributions of the first PCs (Figure 5d, e). By contrast, such clusters are not obvious when coloring the same embedded activity by labs (Figure 5f, g, h). The activity point clouds overlap homogeneously across most labs, indicating similar activity (Figure 5-supplemental 1 for scatter plots, PETHS, cumulative distributions for each region separately, colored by lab).

We quantified this observation via two tests. Firstly, a permutation test using the first 2 PCs of all cells, computing each region's distance between its mean embedded activity and the mean across all remaining regions, then comparing these values to the null distribution of values obtained in an identical manner after shuffling the region labels. Secondly, we directly compared the distributions of the first PCs, applying the Kolmorogov-Smirnov (KS) test to determine whether the distribution of a subset of cells was different from that of all remaining cells, targeting either labs or regions. The KS test results were nearly identical to the distance permutation test results, hence we focus on the KS test results in the following.

When testing regions, we found that all regions differ significantly from the remaining cells, bottom row in Figure 5i. Testing labs, we found that CCU, UCLA, NYU and Berkeley differed significantly from the remaining cells, when using all cells. Region-restricted lab-targeted tests were significant only for Berkeley in VISa/am (Figure 5i and 5-supplemental 1).

These results may be skewed due to sample size difference of the compared distributions. We controlled for this by applying all tests to randomly chosen subsets of all cells (subset sizes randomly sampled across region sizes). Averaging p-values across 100 such sub-sample runs resulted in $p > 0.27$ for all lab-targeting tests while all regions in region-targeting tests have p-values below 0.05, Figure 5j. Taken together, these tests demonstrate that overall, temporal dynamics of firing rates are similar across labs.

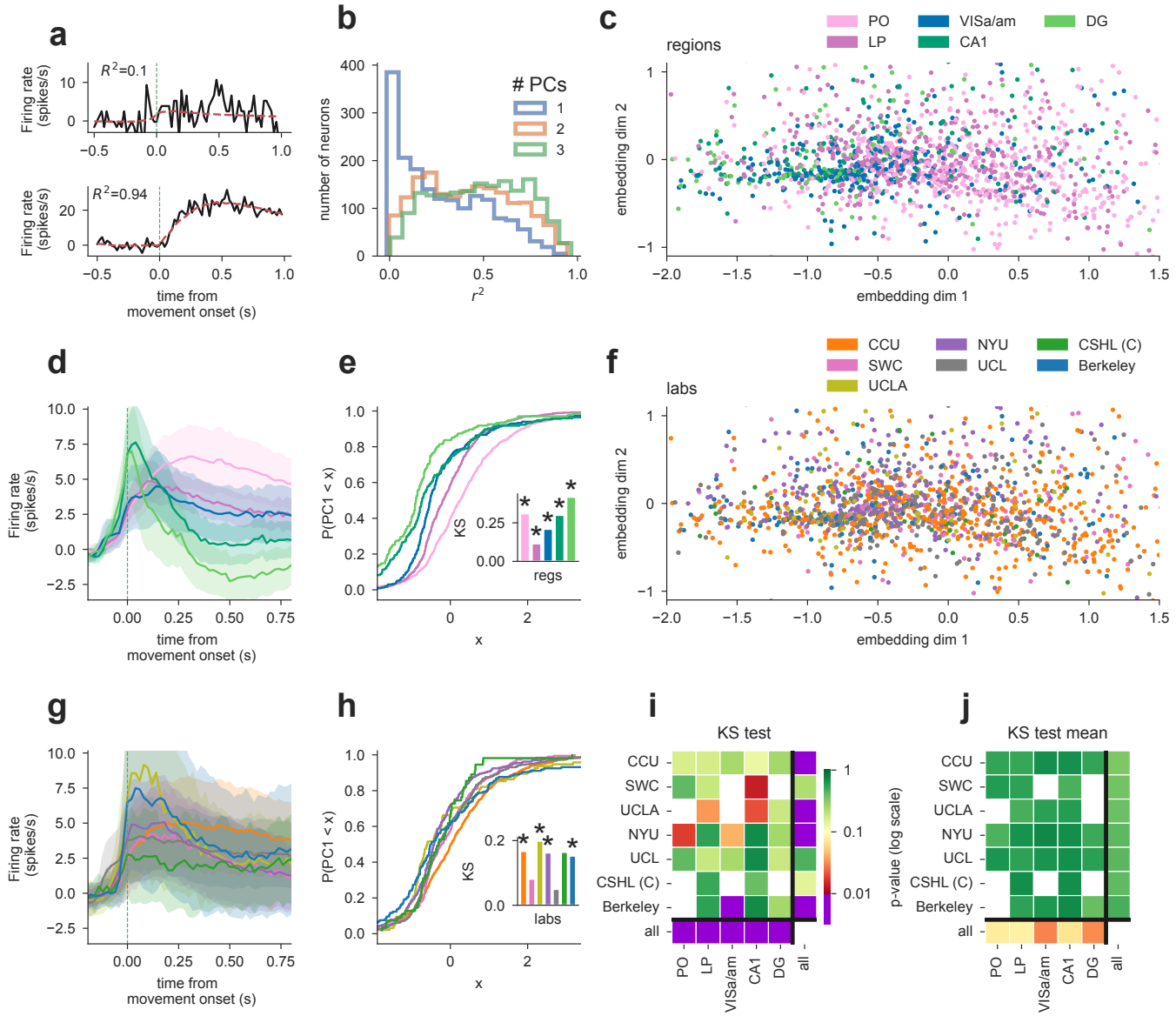

Figure 5. Principal component embedding of peri-event time histograms separates cells from different brain regions but not cells from different labs. (a) PETHs from two example cells (black, fast reaction times only) and 2-PC-based reconstruction (red). Goodness of fit r^2 indicated on top with an example of a poor (top) and good (bottom) fit. **(b)** Histograms of reconstruction goodness of fit across all cells based on reconstruction by 1-3 PCs. Since PETHs are well approximated with only the first 2 PCs, subsequent analyses used the first 2 PCs only. **(c)** Two-dimensional embedding of PETHs of all cells colored by region (each dot corresponds to a single cell). **(d)** Mean firing rates of all cells per region, note visible pink/green divide in line with scatter plot observation. Error bands are standard deviation across cells normalised by the square root of the number of sessions in the region. **(e)** Cumulative distribution of the first embedding dimension (PC1) per region with inset of KS statistic measuring the distance between the distribution of a region's first PC values and that of the remaining cells; asterisks indicate false-discovery-rate (FDR) corrected significance at $p = 0.01$. **(f)** same data as in **(c)** but colored by lab. Visual inspection does not show lab clusters. **(g)** Mean activity for all labs (color conventions the same as in **(f)**). Error bands are standard deviation across cells normalised by square root of number of sessions in lab. **(h)** same as **(e)** but grouping cells per lab. **(i)** FDR-corrected p-values of all KS tests without sub-sampling. The statistic is the KS distance of the distribution of a target subset of cells' first PCs to that of the remaining cells. Columns: the region to which the test was restricted and each row is the target lab of the test. Bottom row "all": p-values reflecting a region's KS distance from all other cells. Right most column "all": p-values of testing a lab's KS distance from all other cells. Small p-values indicate that the target subset of cells can be significantly distinguished from the remaining cells. Note that all region-targeting tests are significant while lab-targeting tests much less so. *Legend continues on next page.*

Figure 5 *(previous page)*. (j) FDR-corrected p-values of all KS tests averaged across 100 random subsets of cells, to account for varying sample sizes, resulting in clearly lower p-values for lab-targeting tests than region-targeting tests. Note that only 7 labs are included in this analysis, as we only include labs that have at least 3 recordings per region (see exclusion criterion Table 1).

Figure 5–Figure supplement 1. Lab-grouped average PETH, CDF of the first PC and 2-PC embedding, separate per brain region.

Differences in spatial position and spike characteristics are only a minor source of variability

While we found little variability between laboratories in terms of electrophysiological features and task variables, we observed considerable variability between recording sessions and mice (Figure 3, Figure 4, and Figure 4-supplemental 1). Due to the variability in the spatial position of the Neuropixels probes (Figure 2), we examined variability in targeting as a potential source of differences in neuronal activity. We also considered single-unit spike waveform characteristics as a source of variability. In the following section, we will examine other potential sources of variability (e.g., mouse movements).

To investigate variability in session-averaged firing rates, we identified neurons that had firing rates different from the majority of neurons within each brain region (absolute deviation from the median firing rate being >15% of the firing rate range). These outlier neurons, which mostly turned out to be high-firing (except in LP), were compared against the general population of neurons in terms of five features: spatial position (x, y, z, computed as the center-of-mass of each unit's spike template on the probe, localized to CCF coordinates in the histology pipeline) and spike waveform characteristics (amplitude, peak-to-trough duration). We observed that recordings in all areas, such as LP (Figure 6a), indeed spanned a wide space within those areas. Interestingly, in areas other than CA1 and DG, the highest firing neurons were not entirely uniformly distributed in space. For instance, in LP, outlier neurons tended to be positioned more laterally and centered on the anterior-posterior axis (Figure 6b). In VISa/am, only the spatial position of neurons, but not differences in spike characteristics, contributed to differences in session-averaged firing rates (Figure 6-supplemental 1b). In contrast, outlier neurons in only LP and PO, but not cortical and hippocampal regions, had different spike characteristics compared to other neurons in their respective regions (Figure 6b and 6-supplemental 3b). It does not appear that high-firing neurons in any brain region belong to a specific neuronal subtype (see Figure 6-supplemental 5).

To quantify the amount of variability in session-averaged firing rates of individual neurons that can be explained by spatial position or spike characteristics, we fit a linear regression model with these five features (x, y, z, spike amplitude, and duration of each neuron) as the inputs. For each brain region, the features that had significant weights were mostly consistent with the results reported above: In VISa/am, z position, or neuron depth, and spike amplitude explained part of the variance; in CA1 and DG, the variance could not be explained by spatial position nor spike characteristics; in LP and PO, x and y positions as well as spike amplitudes explained some of the variance. In LP and PO, where the most amount of variability could be explained by this regression model (having higher R^2 values), these five features accounted for a total of $\sim$13% of the firing rate variability. In VISa/am, CA1, and DG, they accounted for approximately 8%, 1%, and 4% of the variability, respectively.

Next, we examined whether neuronal spatial position and spike features contributed to variability in task-modulated activity. We found that brain regions other than CA1 and DG had minor, yet significant, differences in spatial positions of task-modulated and non-modulated neurons (using the definition of at least of one of the six tests in Figure 4e and Figure 4-supplemental 1). For instance, LP neurons modulated according to the movement initiation test, were positioned more ventrally and centered along the anterior-posterior axis (Figure 6c), while LP neurons modulated according to the left versus right movement test, tended to be more ventral (Figure 6d). Other brain regions had weaker spatial differences than LP (Figure 6-supplemental 1, 2, 3). Spike amplitudes were significantly different between task-modulated and non-modulated neurons only for some tests and only in LP and PO (Figure 6-supplemental 1c-d and 3b-d). On the other hand, the task-aligned Fano Factors of neurons did not have any differences in spatial position except for in VISa/am, where lower Fano Factors (<1) tended to be located ventrally, and in PO, where lower Fano Factors were positioned more laterally (Figure 6-supplemental 4). Spike characteristics of neurons with lower vs. higher Fano Factors were only different in VISa/am (possibly related to differences in cell type; Figure 6-supplemental 4). Lastly, we trained a linear regression model to predict the 2D embedding of PETHs of each cell shown in Figure 5c from the x, y, z coordinates and found that spatial position contains little information ($R^2 \sim 5\%$) about the embedded PETHs of cells.

In summary, our results suggest that spatial position within an area usually contributes only a small amount variability to session-averaged firing rates and task-modulated neuronal activity. Spike characteristics also have a minor contribution to the observed variability. Because the contributions of spatial position and spike features were small, we examine other sources of variability in the next section.

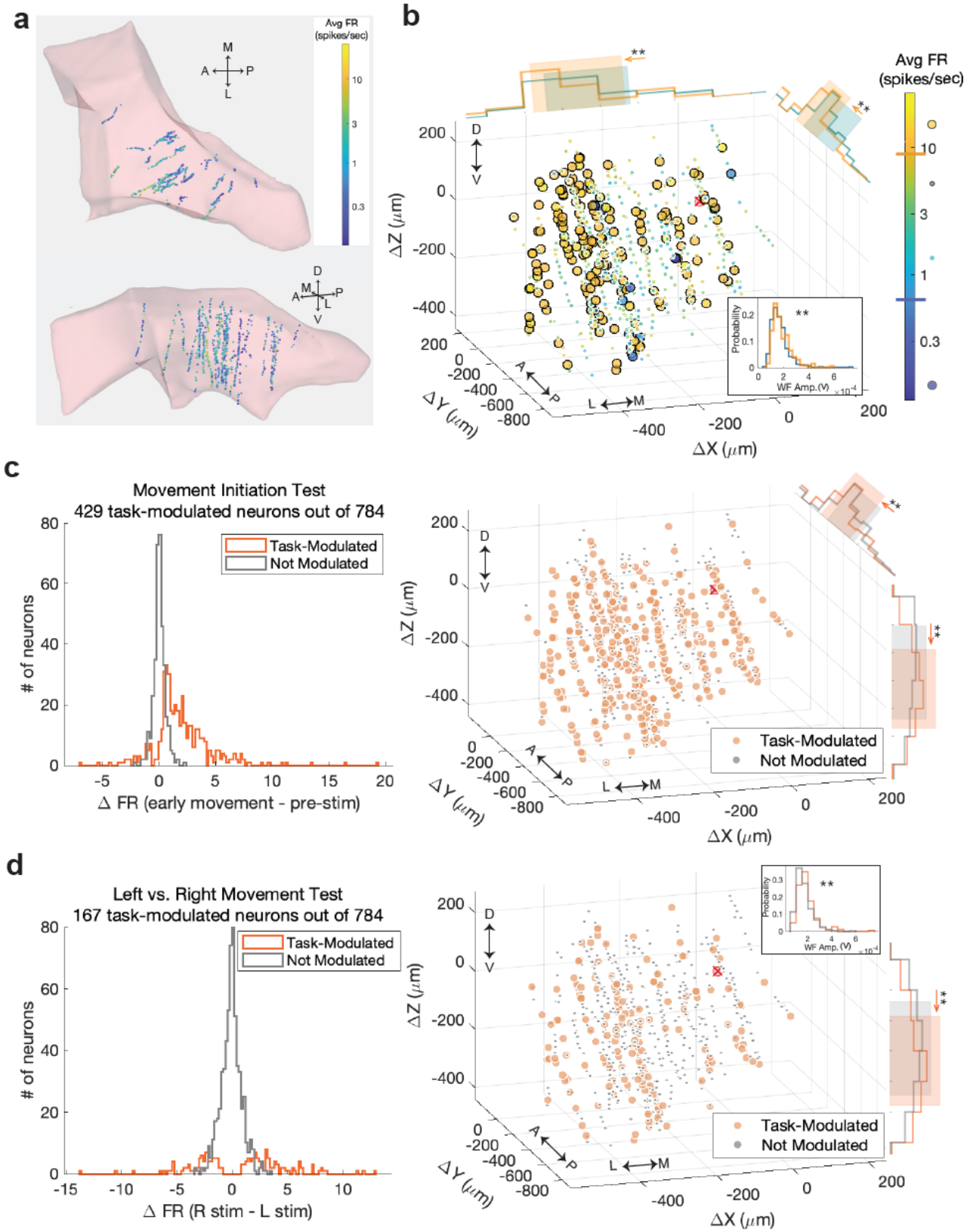

Figure 6. High-firing and task-modulated LP neurons have slightly different spatial positions than other LP neurons, potentially contributing to variability between sessions. (a) Spatial positions of recorded neurons in LP. Colors: session-averaged firing rates.

Legend continues on next page.

Figure 6 *(previous page)*. (b) Spatial positions of LP neurons plotted as distance from the planned target center of mass, indicated with red x. To enable visualization of overlapping data points, jitter was added to the unit locations. Larger circles: outlier neurons, including both high-firing and low-firing neurons (firing rate thresholds shown on the colorbar). In LP, 182 out of 784 neurons were outliers (only 13 of them being low-firing neurons). Only histograms of the spatial positions and spike waveform features that were significantly different between the outlier neurons (yellow) and the general population of neurons (blue) are shown (two-sample Kolmogorov-Smirnov test with Bonferroni correction for multiple comparisons; * and ** indicate corrected p-values of <0.05 and <0.01). Shaded areas: the area between 20th and 80th percentiles of the neurons' locations. **(c)** (Left) Histogram of firing rate changes during the reaction period from the pre-stimulus period (Figure 4e, Figure 4-supplemental 1c) for task-modulated (orange) and non-modulated (gray) neurons. (Right) Spatial positions of task-modulated and non-modulated LP neurons, with histograms of significant features (here, y and z positions) shown. **(d)** Same as **c** but using the left vs. right movement test (Figure 4e and Figure 4-supplemental 1f) to identify task-modulated units.

Figure 6–Figure supplement 1. High-firing and task-modulated VISa/am neurons.

Figure 6–Figure supplement 2. High-firing and task-modulated CA1 neurons.

Figure 6–Figure supplement 3. High-firing and task-modulated DG and PO neurons.

Figure 6–Figure supplement 4. Time-course and spatial position of neuronal Fano Factors.

Figure 6–Figure supplement 5. Neuronal subtypes and firing rates.

A multi-task neural network accurately predicts activity and quantifies sources of neural variability

As discussed above, variability in neural activity between labs or between sessions can be due to many factors. These include differences in behavior between animals, differences in probe placement between sessions, and uncontrolled differences in experimental setups between labs. How can we quantify and distinguish these different sources of variability? Simple linear regression models or generalized linear models (GLMs) are likely too inflexible to capture the nonlinear contributions that many of these variables, including lab IDs and spatial positions of neurons, might make to neural activity. On the other hand, fitting a different nonlinear regression model (involving many covariates) individually to each recorded neuron would be computationally expensive and could lead to poor predictive performance due to overfitting.

To estimate a flexible nonlinear model given constraints on available data and computation time, we adapt an approach that has proven useful in the context of sensory neuroscience (*McIntosh et al., 2016*; *Batty et al., 2016*; *Cadena et al., 2019*). We use a "multi-task" neural network (MTNN; Figure 7a) that takes as input a set of covariates (including the lab ID, the neuron's 3D spatial position in standardized CCF coordinates, the animal's estimated pose extracted from behavioral video monitoring, feedback times, and others; see Table 2 for a full list). The model learns a set of nonlinear features (shared over all recorded neurons) and fits a Poisson regression model on this shared feature space for each neuron. With this approach we effectively solve multiple nonlinear regression tasks simultaneously; hence the "multi-task" nomenclature. The model extends simpler regression approaches by allowing nonlinear interactions between covariates. In particular, previous reduced-rank regression approaches (*Kobak et al., 2016*; *Izenman, 1975*) can be seen as a special case of the multi-task neural network, with a single hidden layer and linear weights in each layer.

Figure 7b shows model predictions on held-out trials for a single neuron in VISa/am. We plot the observed and predicted peri-event time histograms and raster plots, split into left vs. right trials (only the plots for left trials are shown). As a visual overview of which behavioral covariates are correlated with the MTNN prediction of this neuron's activity on each trial, the predicted raster plot and various behavioral covariates that are input into the MTNN are shown in Figure 7c. Overall, the MTNN approach accurately predicts the observed firing rates. When the MTNN and GLMs are trained on movement, task-related, and prior covariates, the MTNN slightly outperforms the GLMs on predicting the firing rate of held-out test trials (See Figure 7-supplemental 1b).

Next we use the predictive model performance to quantify the contribution of each covariate to the fraction of variance explained by the model. Following *Musall et al.* (*2019*), we run two complementary analyses to quantify these effect sizes: *single-covariate fits*, in which we fit the model using just one of the covariates, and *leave-one-out fits*, in which we train the model with one of the covariates left out and compare the predictive explained to that of the full model. As an extension

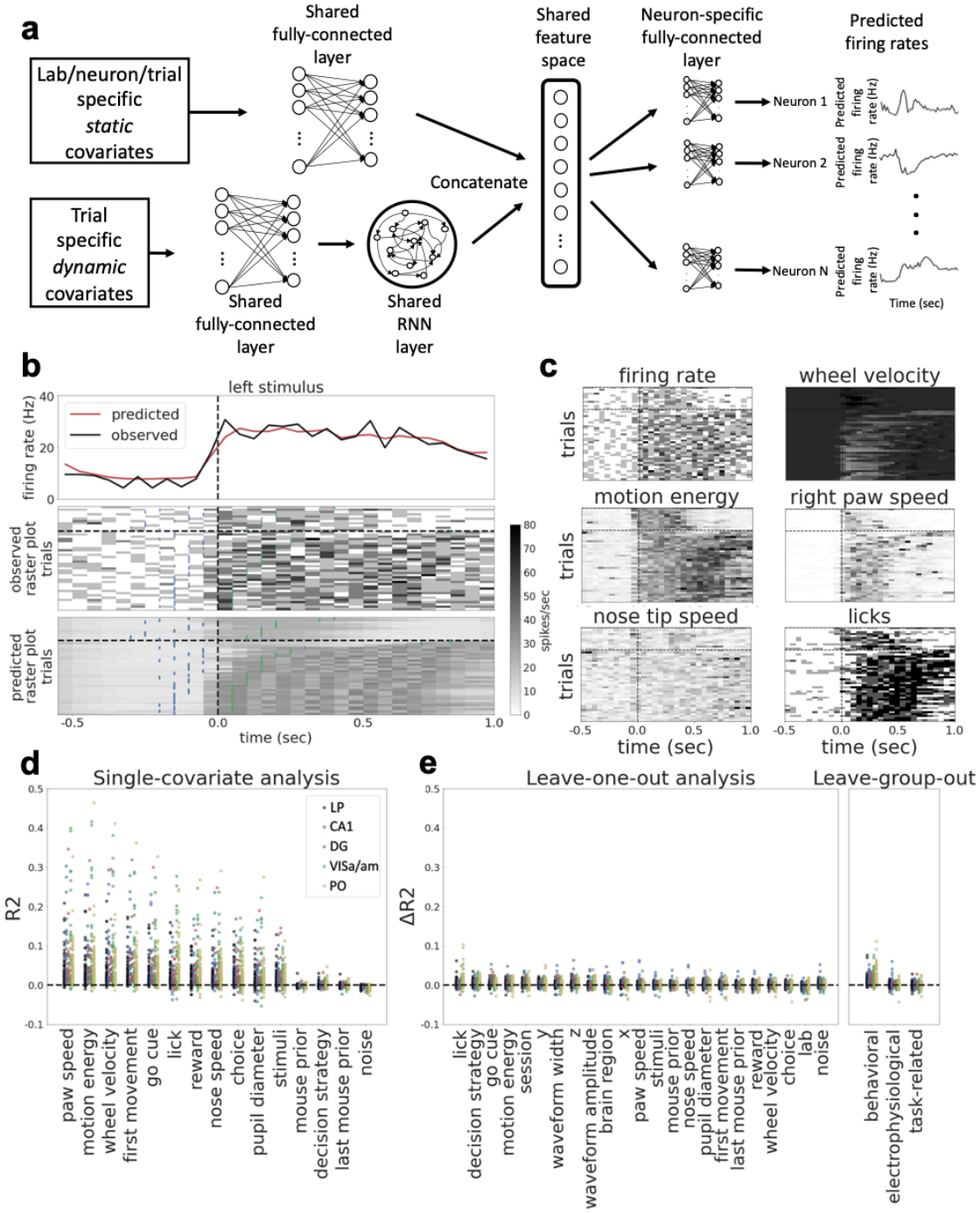

Figure 7. Single-covariate, leave-one-out, and leave-group-out analyses show the contribution of each (group of) covariate(s) to the MTNN model. Lab and session IDs have low contributions to the model. (a) We adapt a MTNN approach for neuron-specific firing rate prediction. The model takes in a set of covariates and outputs time-varying firing rates for each neuron for each trial. See Table 2 for a full list of covariates. **(b)** MTNN model estimates of firing rates (50 ms bin size) of a neuron in VISa/am from an example subject during held-out test trials. The trials with stimulus on the left are shown and are aligned to the first movement onset time (vertical dashed lines). We plot the observed and predicted PETHs and raster plots. The blue ticks in the raster plots indicate stimulus onset, and the green ticks indicate feedback times. The trials above (below) the black horizontal dashed line are incorrect (correct) trials, and the trials are ordered by reaction time. The trained model does well in predicting the (normalized) firing rates. The MTNN prediction quality measured in R^2 is 0.33 on held-out test trials and 0.93 on PETHs of held-out test trials. **(c)** We plot the MTNN firing rate predictions along with the raster plots of behavioral covariates, ordering the trials in the same manner as in **(b)**. We see that the MTNN firing rate predictions are modulated synchronously with several behavioral covariates, such as wheel velocity and paw speed.

Figure 7 *(previous page)*. (d) Single-covariate analysis, colored by the brain region. Each dot corresponds to a single neuron in each plot. **(e)** Leave-one-out and leave-group-out analyses, colored by the brain region. The analyses are run on 239 responsive neurons across 20 sessions. The leave-one-out analysis shows that lab/session IDs have low effect sizes on average, indicating that within and between-lab random effects are small and comparable. The "noise" covariate is a dynamic covariate (white noise randomly sampled from a Gaussian distribution) and is included as a negative control: the model correctly assigns zero effect size to this covariate. Covariates that are constant across trials (e.g., lab and session IDs, neuron's 3D spatial location) are left out from the single-covariate analysis.

Figure 7–Figure supplement 1. Scatter plot of MTNN prediction quality (R^2) vs. mean firing rate (spikes/sec); MTNN slightly outperforms GLMs on predicting the firing rates of held-out test trials; PETHs and MTNN predictions for held-out test trials

Figure 7–Figure supplement 2. MTNN prediction quality on the data simulated from GLMs is comparable to the GLMs' prediction quality. The effect sizes computed by the MTNN leave-one-out analysis are similar to the effect sizes computed by the GLMs' leave-one-out analysis

Figure 7–Figure supplement 3. Pairwise scatterplots of MTNN single-covariate effect sizes.

of the leave-one-out analysis, we run the *leave-group-out analysis*, in which we quantify the contribution of each group of covariates (electrophysiological, task-related, and movement) to the model performance. Using data simulated from GLMs, we first validate that the MTNN leave-one-out analysis is able to partition and explain different sources of neural variability (See Figure 7-supplemental 2).

We then run single-covariate, leave-one-out, and leave-group-out analyses to quantify the contributions of the covariates

Covariate Name	**Type**	**Group**	**Note**
Lab ID	Categorical / Static		
Session ID	Categorical / Static		
Neuron 3D spatial position	Real / Static	Electrophysiological	In standardized CCF coordinates
Neuron amplitude	Real / Static	Electrophysiological	Template amplitude
Neuron waveform width	Real / Static	Electrophysiological	Template width
Paw speed	Real / Dynamic	Movement	Inferred from DLC
Nose speed	Real / Dynamic	Movement	Inferred from DLC
Pupil diameter	Real / Dynamic	Movement	Inferred from DLC
Motion energy	Real / Dynamic	Movement	
Stimulus	Real / Dynamic	Task-related	Stimulus side, contrast and onset timing
Go cue	Binary / Dynamic	Task-related	
First movement	Binary / Dynamic	Task-related	
Choice	Binary / Dynamic	Task-related	
Feedback	Binary / Dynamic	Task-related	
Wheel velocity	Real / Dynamic	Movement	
Mouse Prior	Real / Static		Mouse's prior belief
Last Mouse Prior	Real / Static		Mouse's prior belief in previous trial
Lick	Binary / Dynamic	Movement	Inferred from DLC
Decision Strategy	Real / Static		Decision-making strategy (*Ashwood et al., 2021*)
Brain region	Categorical / Static	Electrophysiological	5 repeated site regions

Table 2. List of covariates input to the multi-task neural network.

listed in Table 2 to the predictive performance of the model on held-out test trials. The results are summarized in Figure 7d and 7e. According to the single-covariate analysis (Figure 7d), paw speed, face motion energy (derived from behavioral video), wheel velocity, and first movement onset timing can individually explain about 5% of variance of the neurons on average. The leave-one-out analysis (Figure 7e left) shows that most covariates have low unique contribution to the predictive power. This is because many covariates are correlated and are capable of capturing variance in the neural activity even if one of the covariates is dropped (See behavioral raster plots in Figure 7c). According to the leave-group-out analysis (Figure 7e right), the “movement" covariates as a group have the highest unique contribution to the model’s performance while the task-related and electrophysiological variables have close-to-zero unique contribution. Most importantly, the leave-one-out analysis shows that lab and session IDs, conditioning on the covariates listed in Table 2, have close to zero effect sizes, indicating that within-lab and between-lab random effects are small and comparable.

Discussion

We set out to test whether the results from an electrophysiology experiment could be reproduced across geographically separated laboratories. We observed notable variability in the position of the electrodes in the brain despite efforts to target the same location. Still, after applying stringent quality-control criteria (including the RIGOR standards, Table 1), we found that electrophysiological features such as neuronal yield, firing rate, and normalized LFP power were largely reproducible across laboratories; their intra-lab distributions did not differ more than expected by chance. Similarly, the proportion of cells whose responses are tuned to behaviorally-relevant task events was reproducible across labs (although the exact firing rate distributions underlying these proportions was somewhat more variable). Finally, a multi-task neural network approach predicted the firing rates of different units across sessions, and again, the within-lab random effects estimated by this model were comparable to between-lab random effects. Taken together, our results suggest that careful standardization can lead to reproducible electrophysiological results across laboratories.

The reassuring absence of systematic differences across labs argues that standardization of procedures is a helpful step in generating reproducible results. Interestingly, although our standardization eliminated most systematic differences, we cannot determine whether the reproducibility we observed was driven by person-to-person standardization or lab-to-lab standardization. Most likely, both factors contributed: all lab personnel received standardized instructions for how to implant head bars, train mice, and train animals, which likely reduced personnel-driven differences. In addition, our use of standardized instrumentation and software minimized lab-to-lab differences that might normally be present.

Reproducibility in our electrophysiology studies was further enhanced by rigorous quality control metrics that ultimately led us to exclude a significant fraction of datasets (54/96 sessions). Quality control was enforced for diverse aspects of the experiments, including histology, behavior, targeting, neuronal yield, and the total number of completed sessions. Among these measures, recordings with high noise and low neuronal yield were significantly represented in sessions that were excluded. A number of issues contributed here, including artifacts present in the recordings, inadequate grounding, and a decline in craniotomy health; all of these can potentially be improved with experimenter experience. A few quality control metrics were specific to our experiments (and thus not listed in Table 1). For instance, we excluded sessions with fewer than 400 trials which could be too stringent (or not stringent enough) for other experiments.

These observations suggest that future experiments would be more consistently reproducible if researchers followed, or at least reported, a number of agreed upon criteria, such as the RIGOR standards we define in Table 1. This approach has been successful in other fields: for instance, the neuroimaging field has agreed upon a set of guidelines for “best practices," and has identified factors that can impede those practices (*Nichols et al., 2017*). The genomics field likewise adopted the Minimum Information about a Microarray Experiment (MIAME) standard, designed to ensure that data from microarrays could be meaningfully interpreted and experimentally verified (*Brazma et al., 2001*). Finally, the autophagy community have standards for experiments that were established in 2008 (*Klionsky, 2016*). Our work here suggests the creation of a similar set of standards for electrophysiology and behavioral experiments would be beneficial. These could include expectations for reporting (such as histological information and behavioral trial numbers) as well as suggestions for minimizing variability (e.g., agreed upon standards for the noise level that would exclude a recording). We propose that our "Recording Inclusion metrics and Guidelines for Optimal Reproducibility" (RIGOR, Table 1) constitute a set of criteria that could be adopted across the field to improve reproducibility.

Establishment of such standards has the potential to enhance lab-to-lab reproducibility, but experiment-to-experiment variability may not be entirely eliminated. A large-scale effort to enhance reproducibility in *C. elegans* aging studies successfully replicated average lifespan curves across 3 labs by standardizing experimental methods such as handling of organisms and notation of age (e.g. when egg is hatched vs laid) (*Lithgow et al., 2017*; *Lucanic et al., 2017*). Still, variability in the lifespan curves of individual worms nevertheless persisted, warranting further studies to understand what molecular differences might explain this. Similarly, we observed no systematic difference across labs in either electrophysiological measures (Fig 3d) or functional responses (Figure 5j), but nonetheless found considerable variability across experiments within each lab (Figure 4g).

We found probe targeting to be a large source of variability, driven by micro-manipulator positioning and anatomical discrepancies. One possibility is that some aspect of our workflow led us to discard more insertions than would be typical.

However, another possibility is that our workflow uncovered a weakness in the standard targeting methods. Specifically, we used an automated histological pipeline combined with alignment and tracing that required agreement between multiple users. This approach, which exceeds what is done in many experimental labs, revealed for instance that much of the variance in targeting was due to the probe entry positions at the brain surface, which were randomly displaced across the dataset. The source of this variance could be due to a discrepancy in skull landmarks compared to the underlying brain anatomy. Accuracy in placing probes along a planned trajectory is therefore limited by this variability (about 400µm). Probe angle also showed a small degree of variance and a bias in both anterior-posterior and medio-lateral direction, indicating that the Allen Common Coordinate Framework (CCF) (*Wang et al., 2020*) and stereotaxic coordinate systems are slightly offset. Minimizing variance in probe targeting is an important element in increasing reproducibility, as slight deviations in probe entry position and angle can lead to samples from different populations of neurons. Our approach suggests a path forward to minimize these biases: probe angles must be carefully computed from the CCF, as the CCF and stereotaxic coordinate systems do not define the same coronal plane angle. Small differences in probe location may be responsible for other studies arriving at different conclusions, highlighting the need for agreed upon methods for targeting specific areas (*Rajasethupathy et al., 2015*; *Andrianova et al., 2022*).

Our results also highlight the critical importance of reproducible histological processing and subsequent probe alignment. Specifically, we used a centralized histology and registration pipeline to assign each recording site on each probe to a particular anatomical location, based on registration of the histological probe trajectories to the CCF and the electrophysiological features recorded at each site. This differs from previous approaches, in which stereotaxic coordinates alone were used to target an area of interest and exclusion criteria were not specified; see e.g. (*Najafi et al., 2020*; *Harvey et al., 2012*; *Goard et al., 2016*; *Raposo et al., 2014*; *Erlich et al., 2015*). The reliance on stereotaxic coordinates for localization, instead of standardized histological registration, is a possible explanation for conflicting results across laboratories in previous literature. Our results speak to the importance of adopting standardized procedures more broadly across laboratories.

A major contribution of our work is open-source data and code: we share our full dataset (link to data portal) and suite of analysis tools for quantifying reproducibility (link to code repository) and computing the RIGOR standards. The analyses here required significant improvements in data architecture, visualization, spike sorting, histology image analysis, and video analysis. Our analyses uncovered major gaps and issues in the existing toolsets that required improvements (see Methods and *The International Brain Laboratory* (*2021a*,b) for full details); the large-scale dataset analyzed here proved to be a useful stress test pointing to improved analysis pipelines. For example, we improved existing spike sorting pipelines with regard to scalability, reproducibility, and stability. These improvements contribute towards advancing automated spike sorting, and move beyond subjective manual curation, which scales poorly and limits reproducibility. We anticipate that our open-source dataset will play an important role in further improvements to these pipelines and also the development of further methods for modeling the spike trains of many simultaneously recorded neurons across multiple brain areas and experimental sessions.

Scientific advances rely on the reproducibility of scientific findings. The current study demonstrates that reproducibility is attainable for large-scale neural recordings during a standardized perceptual detection task across ten laboratories. We offer several recommendations to increase reproducibility, including (1) standardized protocols for data collection, (2) protocols for data processing, and (3) rigorous data quality metrics. Furthermore, we have made improvements in data architecture and processing, now available to the public. Our study provides a framework for the collection and analysis of large neural datasets in a reproducible manner that will play a key role as neuroscience continues to move towards increasingly complex datasets.

Resources

Data access

Please visit https://int-brain-lab.github.io/iblenv/notebooks_external/data_release_repro_ephys.html to access the data used in this article. Please visit the visualisation website https://viz.internationalbrainlab.org/app to view the data (use the tab *Repeated site*).

Code repository

Please visit https://github.com/int-brain-lab/paper-reproducible-ephys/ to access the code used to produce the results and figures presented in this article.

Protocols and pipelines

Please visit https://figshare.com/projects/Reproducible_Electrophysiology/138367 to access the protocols and pipelines used in this article.

Quality control and data inclusion

Please see this spreadsheet (link) for a comprehensive overview of which recordings are used in what figure panels, as well as the reasons for inclusion or exclusion.

Methods and Materials

All procedures and experiments were carried out in accordance with local laws and following approval by the relevant institutions: the Animal Welfare Ethical Review Body of University College London; the Institutional Animal Care and Use Committees of Cold Spring Harbor Laboratory, Princeton University, University of California at Los Angeles, and University of California at Berkeley; the University Animal Welfare Committee of New York University; and the Portuguese Veterinary General Board.

Animals

Mice were housed under a 12/12 h light/dark cycle (normal or inverted depending on the laboratory) with food and water available ad libitum, except during behavioural training days. Electrophysiological recordings and behavioural training were performed during either the dark or light phase of the cycle depending on the laboratory. N=48 adult mice (C57BL/6, male and female, obtained from either Jackson Laboratory or Charles River) were used in this study. Mice were aged 17-41 weeks and weighed 16.4-34.5 g on the day of the headbar implant surgery.

Materials and apparatus

Briefly, each lab installed a standardized electrophysiological rig (named 'ephys rig' throughout this text), which differed slightly from the apparatus used during behavioral training (*The International Brain Laboratory et al., 2021*). The general structure of the rig was constructed from Thorlabs parts and was placed inside a custom acoustical cabinet clamped on an air table (Newport, M-VIS3036-SG2-325A). A static head bar fixation clamp and a 3D-printed mouse holder were used to hold a mouse such that its forepaws rest on the steering wheel (86652 and 32019, LEGO) (*The International Brain Laboratory et al., 2021*). Silicone tubing controlled by a pinch valve (225P011-21, NResearch) was used to deliver water rewards to the mouse. The display of the visual stimuli occurred on a LCD screen (LP097Q × 1, LG). To measure the precise times of changes in the visual stimulus, a patch of pixels on the LCD screen flipped between white and black at every stimulus change, and this flip was captured with a photodiode (Bpod Frame2TTL, Sanworks). Ambient temperature, humidity, and barometric air pressure were measured with the Bpod Ambient module (Sanworks), wheel position was monitored with a rotary encoder (05.2400.1122.1024, Kubler).

Videos of the mouse were recorded from 3 angles (left, right and body) with USB cameras (CM3-U3-13Y3M-CS, Point Grey). The left camera acquires at 60Hz; full resolution (1280 x1024), right camera at 150Hz; half resolution (640x512), and body camera

at 30Hz; half resolution (640Hzx512). A custom speaker (Hardware Team of the Champalimaud Foundation for the Unknown, V1.1) was used to play task-related sounds, and an ultrasonic microphone (Ultramic UM200K, Dodotronic) was used to record ambient noise from the rig. All task-related data was coordinated by a Bpod State Machine (Sanworks). The task logic was programmed in Python and the visual stimulus presentation and video capture were handled by Bonsai (*Lopes et al., 2015*) utilizing the Bonsai package BonVision (*Lopes et al., 2021*).

All recordings were made using Neuropixels probes (Imec, 3A and 3B models), advanced in the brain using a micromanipulator (Sensapex, uMp-4) tilted by a 15 degree angle from the vertical line. The aimed electrode penetration depth was 4.0 mm. Data were acquired via an FPGA (for 3A probes) or PXI (for 3B probes, National Instrument) system and stored on a PC.

Headbar implant surgery

Briefly, mice were placed in an induction box with 3-4% isoflurane and maintained at 1.5-2% isoflurane. Saline 10mg/kg subcutaneously is given each hour. The mouse is placed in the stereotaxic frame using ear bars placed in the ridge posterior to the ear canal. The mouse is then prepped for surgery, removing hair from the scalp using epilation creme. Much of the underlying periosteum was removed and bregma and lambda were marked. Then the head was positioned such that there was a 0 degree angle between bregma and lambda in all directions. Lateral and middle tendons are removed using fine forceps. The head bar was then placed in one of three stereotactically defined locations and cemented in place These locations are: AP -6.90, ML +/- 1.25(curved headbar placed caudally onto cerebellum), AP +1.36, ML +/- 1.25 (curved headbar placed rostrally onto frontal zones), and AP -2.95, ML +/-1.25 (straight headbar placed centrally).. The location of the future craniotomies were measured using a pipette referenced to bregma, and marked on the skull using either a surgical blade or pen. A small amount of vetbond was applied to the edges of the skin wound to seal it off and create more surface area. The exposed skull was then covered with cement and clear UV curing glue, ensuring that the remaining scalp was unable to retract from the implant.

Behavioral training and habituation to the ephys rig

All recordings performed in this study were done in expert mice. To reach this status, animals were habituated for three days and trained for several days in the equal probability task version where the Gabor patch appears on the right or left side of the screen with equal probability. Animals are trained to move the visual stimulus controlled by a wheel toward the center of the screen. Animals must reach a ’trained 1b’ status wherein each of the three consecutive sessions, the mouse completed over 400 trials and performed over 90% on the easy (contrast >= 50%) trials. Additionally, the median reaction time across these sessions must be below 2 seconds for the 0% contrast. Lastly, a psychometric curve is fitted with four parameters bias, lapse right, lapse left and threshold, must meet the following criteria: the absolute bias must be below 10, the threshold below 20, and each lapse below 0.1. Once these conditions are met, animals progress to ’biasedChoiceWorld’ in which they are first presented with an unbiased block of trials and subsequently blocks are from either of two biased blocks: Gabor patch is presented on the left and right with probabilities of 0.2 and 0.8 (20:80) respectively, and in the other block type the Gabor patch is presented on the left and right with probabilities of 0.8 and 0.2 (80:20) respectively.In summary, once mice learned the biasedChoiceWorld task (criteria ‘ready4ephysRig’ reached), they were habituated to the electrophysiology rig. Briefly, this criterion is met by performing three consecutive sessions that meet ’trained 1b’ status. Additionally, psychometric curves (separately fit for each block type) have bias shifts < 5%, and lapse rates measured on asymmetric blocks are below 0.1. Their first requirement was to perform one session of biasedChoiceWorld on the electrophysiology rig, with at least 400 trials and 90% correct on easy contrasts (collapsing across block types). Once this criterion was reached, time delays were introduced at the beginning of the session; these delays served to mimic the time it would take to insert electrodes in the brain. To be included in subsequent sessions, mice were required to maintain performance for 3 subsequent sessions (same criterion as ‘ready4ephysRig’), with a minimum of one session with a 15-minute pre-session delay.

In this study, electrophysiology sessions were considered in the analysis if the mice performed at least 400 trials.

Electrophysiological recording using Neuropixels probes

Data acquisition

Briefly, upon the day of electrophysiological recording, the animal was anaesthetised using isoflurane and surgically prepared. The UV glue was removed using ethanol and a biopsy punch or scalpel blade. Exposed skull was then checked for infection. A test was made to check whether the implant could hold liquid without leaking to ensure that the brain does not dry during the recording. Subsequently, a grounding pin was cemented to the skull using Metabond. One or two craniotomies (1 × 1 mm) were made over the marked locations using either a biopsy punch or drill. The dura was left intact, and the brain was lubricated with ACSF. DuraGel was applied over the dura as a moisturising sealant, and covered with a layer of Kwikcast. The mouse was administered analgesics subcutaneously, and left to recover in a heating chamber until locomotor and grooming activity were fully recovered.

Once the animal was recovered from the craniotomy, it was fixed in the apparatus. Once a craniotomy was made, up to 4 subsequent recording sessions were made in that same craniotomy. Up to two probes were implanted in the brain on a given session.

Probe track labeling

CM-Dil (V22888 Thermofisher) was used to label probes for subsequent histology. Store CM-Dil in the freezer -at 20C until ready for use. On the day of recording, thaw CM-Dil at room temperature, protecting it from light. Labeling took place under a microscope while the Neuropixels probe was secured onto a micromanipulator, electrode sites facing up. 1uL of CM-Dil was placed onto either a coverslip or parafilm. Using the micromanipulator, the probe tip was inserted into the drop of dye with care taken to not get dye onto the electrode sites. For Neuropixels probes, the tip extends about 150um from the first electrode site. The tip is kept in the dye until the drop dries out completely (approximately 30 seconds) and then the micromanipulator is slowly retracted to remove the probe).

Spike sorting

Raw electrophysiological recordings were initially saved in a flat uncompressed binary format, representing a storage of 1.3GB per minute. To save disk space and achieve better transfer speeds we utilized simple lossless compression to achieve a compression ratio between 2x and 3x. In many cases, we encounter line noise due to voltage leakage on the probe. This translates into large "stripes" of noise spanning the whole probe. To reduce the impact of these noise "stripes" we perform three main pre-processing steps including: (1) correction for "sample shift" along the length of the probe by aligning the samples with a frequency domain approach; (2) automatic detection, rejection and interpolation of failing channels; (3) application of a spatial “de-striping" filter. After these preprocessing steps, spike sorting was performed using a modified version of the Kilosort 2.5 algorithm (*Steinmetz et al., 2021*). At this step, we apply registration, clustering, and spike deconvolution. We found it necessary to improve the original code in several aspects (e.g., improved modularity and documentation, and better memory handling for datasets with many spikes) and developed an open-source Python port; the code repository is here: (*The International Brain Laboratory, 2021b*). See *The International Brain Laboratory et al.* (*2022*) for full details.

Single cluster quality metrics

To determine whether a single cluster will be used in downstream analysis, we used three metrics: the refractory period, an amplitude cut-off estimate, and the median of the amplitudes. First, we developed a metric which estimates whether a neuron is contaminated by refractory period violations (indicating potential overmerge problems in the clustering step) without assuming the length of the refractory period. For each of the many refractory period lengths, we compute the number of spikes (refractory period violations) that would correspond to some maximum acceptable amount of contamination (chosen as 10%). We then compute the likelihood of observing fewer than this number of spikes in that refractory period under the assumption of Poisson spiking. For a neuron to pass this metric, this likelihood that our neuron is less than 10% contaminated, must be larger than 90% for any one of the possible refractory period lengths.

Next, we compute an amplitude cut-off estimate. This metric estimates whether an amplitude distribution is cut off by thresholding in the deconvolution step (thus leading to a large fraction of missed spikes). To do so, we compare the lowest bin of the histogram (the number of neurons with the lowest amplitudes), to the bins in the highest quantile of the distribution (defined as the top 1/4 of bins higher than the peak of the distribution.) Specifically, we compute how many standard deviations the height of the low bin falls outside of the mean of the height of the bins in the high quantile. For a neuron to pass this metric, this value must be less than 5 standard deviations, and the height of the lowest bin must be less than 10% of the height of the peak histogram bin.

Finally, we compute the median of the amplitudes. For a neuron to pass this metric, the median of the amplitudes must be larger than 50 uV.

Local field potential (LFP)

Concurrently with the action potential band, each channel of the Neuropixel probe recorded a low-pass filtered trace at a sampling rate of 2500 Hz. A denoising was applied to the raw data, comprising four steps. First a Butterworth low-cut filter is applied on the time traces, with 2Hz corner frequency and order 3. Then a subsample shift is applied to rephase each channel according to the time-sampling difference due to sequential sampling of the hardware. Then faulty bad channels are automatically identified, removed and interpolated. At last the median reference is subtracted at each time sample. See *The International Brain Laboratory et al.* (*2022*) for full details. After this processing, the power spectral density at different frequencies was estimated per channel using the Welch's method with partly overlapping Hanning windows of 1024 samples. Power spectral density (PSD) was converted into dB as follows:

$$dB = 10 * log(PSD) \tag{1}$$

Serial section two-photon imaging

Mice were given a terminal dose of pentobarbital intraperitoneally. Toe-pinch is performed as confirmation that the mouse is under before proceeding with the surgical procedure. Thoracic cavity is opened, atrium is cut, and PBS followed by 4% formaldehyde solution (Thermofisher 28908) in 0.1M PB pH 7.4 is perfused through the left ventricle. Whole mouse brain was dissected, and post-fixed in the same fixative for a minimum of 24 hours at room temperature. Tissues were washed and stored for up to 2-3 weeks in PBS at 4C, prior to shipment to the Sainsbury Wellcome Centre for image acquisition.

The brains were embedded in agarose and imaged in a water bath filled with 50 mM PB using a 4 kHz resonant scanning serial section two-photon microscopy (*Ragan et al., 2012*; *Economo et al., 2016*). The microscope was controlled with ScanImage Basic (Vidrio Technologies, USA), and BakingTray, a custom software wrapper for setting up the imaging parameters (*Campbell, 2020*). Image tiles were assembled into 2D planes using StitchIt (*Campbell, 2021*). Whole brain coronal image stacks were acquired at a resolution of 4.4 x 4.4 x 25.0 µm in XYZ (Nikon 16x NA 0.8), with a two-photon laser wavelength of 920 nm, and approximately 150 mW at the sample. The microscope cut 50 µm sections using a vibratome (Leica VT1000) but imaged two optical planes within each slice at depths of about 30 µm and 55 µm from the tissue surface using a PIFOC. Two channels of image data were acquired simultaneously using Hamamatsu R10699 multialkali PMTs: 'Green' at 525 nm $\pm$25 nm (Crhoma ET525/50m); 'Red' at 570 nm low pass (Chroma ET570lp).

Whole brain images were downsampled to 25 µm isotropic voxels and registered to the adult mouse Allen common coordinate framework (*Wang et al., 2020*) using BrainRegister (*West, 2021*), an elastix-based (*Klein et al., 2010*) registration pipeline with optimised parameters for mouse brain registration. Two registrations are performed, samples are registered to the CCF template image and the CCG template is registered to the sample.

Probe track tracing and alignment

Tracing of Neuropixels electrode tracks is performed on registered image stacks. This is performed by the experimenter and an additional member. Neuropixels probe tracks were manually traced to yield a probe trajectory using Lasagna (*Campbell et al., 2020*), a Python-based image viewer equipped with a plugin tailored for this task. Tracing was performed on the merged images on the green (auto-fluorescence) and red (CM-Dil labeling) channels, using both coronal and sagittal views. Traced probe track

data was uploaded to an Alyx server (*Rossant et al., 2021*); a database designed for experimental neuroscience laboratories. Neuropixels channels were then manually aligned to anatomical features along the trajectory using electrophysiological landmarks with a custom electrophysiology alignment tool (*Faulkner, 2020*) (*Liu et al., 2021*).

Permutation tests and power analysis

We use permutation tests to study the reproducibility of neural features across laboratories. To this end, we first defined a test statistic that is sensitive to systematic deviations in the distributions of features between laboratories: the maximum absolute difference between the cumulative distribution function (CDF) of a neural feature within one lab and the CDF across all other labs (similar to the test statistic used for a Kolmogorov–Smirnov test). For the CDF, each mouse might contribute just a single value (e.g. in the case of the deviations from the target region), or a number for every neuron in that mouse (e.g. in the case of comparing firing rate differences during specific time-periods). The deviations between CDFs from all the individual labs are then reduced into one number by considering only the deviation of the lab with the strongest such deviation, giving us a metric that quantifies the difference between lab distributions. The null hypothesis is that there is no difference between the different laboratory distributions, i.e. the assignment of mice to laboratories is completely random. We sampled from the corresponding null distribution by permuting the assignments between laboratories and mice randomly 50,000 times (leaving the relative numbers of mice in laboratories intact) and computing the test statistic on these randomised samples. Given this sampled null distribution, the p-value of the permutation test is the proportion of the null distribution that has more extreme values than the test statistic that was computed on the real data.

For the power analysis in Fig. 2, the goal was to find how strongly we have to shift all values (firing rate modulations or Fano factors) within the individual labs, in order to create a significant p-value for any given test. This grants us a better understanding of the workings and limits of our test. As we chose an α level of 0.01, we needed to find the perturbations that gave a p-value < 0.01. To achieve this for a given test and a given lab, we took the values of every neuron within that lab, and shifted them all up or down by a certain amount. We used binary search to find the exact points at which such an up- or down-shift caused the test to become significant. This analysis tells us exactly at which points our test becomes significant, and importantly makes sure that our permutation test is actually sensitive enough to pick up on deviations of certain magnitudes. It may seem counter intuitive that some tests allow for larger deviations than others, or that even within the same test some labs have a different range of possible perturbations than others. This is because the test considers the entire distribution of values, resulting in possibly complex interactions between the labs. Precisely because of these interactions of the data with the test, we performed a thorough power analysis to ensure that our procedure is sufficiently sensitive to across-lab variations. The bottom row of Fig. 2 shows the overall distribution of permissible shifts, the large majority of which is below one standard deviation of the corresponding lab distribution.

Dimensionality reduction of PETHs via principal component analysis

In Figure 5 we use principal component analysis (PCA) to embed PETHs into a two-dimensional feature space for visualization and further analysis. Our overall approach is to compute PETHs, split into fast-reaction-time and slow-reaction-time trials, then concatenate these PETH vectors for each cell to obtain an informative summary of each cell's activity. Next we stack these double PETHs from all labs into a single matrix and use PCA to obtain a low-rank approximation of this PETH matrix.

In detail, the two PETHs consist of one averaging fast reaction time ($< 0.15sec$) trials and the other slow reaction time ($> 0.15sec$) trials, each of length T time steps. We used 20 ms bins, from −0.5 sec to 1.5 sec relative to motion onset, so $T = 100$. We also performed a simple normalization on each PETH, dividing the firing rates by the baseline firing rate (prior to motion onset) of each cell plus a small positive offset term (to avoid amplifying noise in very low-firing cells), following *Steinmetz et al.* (*2021*).

Let the stack of these double PETH vectors be Y, being a $N \times 2T$ matrix, where N is the total number of neurons recorded across 5 brain regions and labs. Running principal components analysis (PCA) on Y (singular value decomposition) is used to obtain the low-rank approximation $UV \approx Y$. This provides a simple low-d embedding of each cell: U is $N \times k$, with each row of U representing a k-dimensional embedding of a cell that can be visualized easily across labs and brain regions. V is $k \times 2T$ and

corresponds to the k temporal basis functions that PCA learns to best approximate Y. Figure 5(a) shows two cells of Y and the corresponding PCA approximation from UV.

The scatter plots in Figure 5 show the embedding U across labs and brain regions, with the embedding dimension $k = 2$. Each $k \times 1$ vector in U, corresponding to a single cell, is assigned to a single dot in Figure 5c.

Linear regression model to quantify the contribution of spatial and spike features to variability

To fit a linear regression model to the session-averaged firing rate of neurons, for each brain region, we used a Nx5 predictor matrix where N is the number of recorded neurons within the region. The five columns contain the following five covariates for each neuron: x, y, z position, spike amplitude, and spike peak-to-trough duration. The Nx1 observation matrix consisted of the average firing rate for each neuron throughout the entire recording period. The linear model was fit using ordinary least-squares without regularization. The unadjusted coefficient of determination (R^2) was used to report the level of variability in neuronal firing rates explained by the model.

Video analysis

In the recording rigs, we used three cameras, one called 'left' at full resolution (1280x1024) and 60 Hz filming the mouse from one side, one called 'right' at half resolution (640x512) and 150 Hz, filming the mouse symmetrically from the other side, and one called 'body' filming the trunk of the mouse from above. Several quality control metrics were developed to detect video issues such as poor illumination or accidental misplacement of the cameras.

We used DeepLabCut (*Mathis et al., 2018*) to track various body parts such as the paws, nose, tongue, and pupil. The pipeline first detects 4 regions of interest (ROI) in each frame, crops these ROIs using ffmpeg (*Tomar, 2006*) and applies a separate network for each ROI to track features. For each side video we track the following points:

- ROI eye:

  `'pupil_top_r', 'pupil_right_r', 'pupil_bottom_r', 'pupil_left_r'`

- ROI mouth:

  `'tongue_end_r', 'tongue_end_l'`

- ROI nose:

  `'nose_tip'`

- ROI paws:

  `'paw_r', 'paw_l'`

The right side video was flipped and spatially up-sampled to look like the left side video, such that we could apply the same DeepLabCut networks. The code is available here: (*The International Brain Laboratory, 2021a*).

Extensive curating of the training set of images for each network was required to obtain reliable tracking across animals and laboratories. We annotated in total more than 10K frames, across several iterations, using a semi-automated tracking failure detection approach, which found frames with temporal jumps, three-dimensional re-projection errors when combining both side views, and heuristic measures of spatial violations. These selected 'bad' frames were then annotated and the network re-trained. To find further raw video and DeepLabCut issues, we inspected trial-averaged behaviors obtained from the tracked features, such as licking aligned to feedback time, paw speed aligned to stimulus onset and scatter plots of animal body parts across a session superimposed onto example video frames. See *The International Brain Laboratory* (*2021a*) for full details.

Despite the large labeled dataset and multiple network retraining iterations described above, DeepLabCut was not able to achieve sufficiently reliable tracking of the paws or pupils. Therefore we used an improved tracking method for these body parts (*Biderman et al., 2023*), trained on the same final labeled dataset used to train DeepLabCut.

Multi-task neural network model to quantify sources of variability

Data preprocessing

For the multi-task neural network (MTNN) analysis, we used data from 20 sessions recorded in CCU, CSHL (C), SWC, UCL, and Berkeley. We filtered out the sessions with unreliable behavioral traces from video analysis, and selected labs with at least 4 sessions for the MTNN analysis. For the labs with more than 4 sessions, we randomly subsampled 4 sessions. We included various covariates in our feature set (e.g. go-cue signals, stimulus/reward type, Deep Lab Cut behavioral outputs). For the "decision strategy" covariate, we used the posterior estimated state probabilities of the 4-state GLM-HMMs trained on the sessions used for the MTNN analysis (*Ashwood et al., 2021*). Both biased and unbiased data were used when training the 4-state model. For each session, we first filtered out the trials where no choice is made. We then selected the trials whose stimulus onset time is no more than 0.4 seconds before the first movement onset time and feedback time is no more than 0.9 seconds after the first movement onset time. Finally, we selected responsive neurons whose mean firing rate is greater than 5 spikes/second for further analyses. For sessions with more than 15 such responsive neurons, we randomly sampled 15 neurons, in order to keep the MTNN training time at a reasonable level while also preventing the sessions with relatively high number of neurons from dominating the losses and updates during MTNN training. The lab IDs and session IDs were each encoded in a "one-hot" format (i.e., each lab (or session) is encoded as a length 4 one-hot vector). For the leave-one-out effect size of the session IDs, we compared the model trained with all of the covariates in Table 2 against the model trained without the session IDs. For the leave-one-out effect size of the lab IDs, we compared the model trained without the lab IDs against the model trained without both the lab and session IDs. We prevented the lab and session IDs from containing overlapping information with this encoding scheme, where the lab IDs cannot be predicted from the session IDs, and vice versa, during the leave-one-out analysis.

Model Architecture

Given a set of covariates in Table 2, the MTNN predicts the target sequence of firing rates from 0.5 seconds before first movement onset to 1 second after, with bin width set to 50 ms (30 time bins). More specifically, a sequence of feature vectors $x_{\text{dynamic}} \in \mathbb{R}^{D_{\text{dynamic}} \times T}$ that include dynamic covariates, such as Deep Lab Cut (DLC) outputs, and wheel velocity, and a feature vector $x_{\text{static}} \in \mathbb{R}^{D_{\text{static}}}$ that includes static covariates, such as the lab ID, neuron's 3-D location, are input to the MTNN to compute the prediction $y^{pred} \in \mathbb{R}^T$, where D_{static} is the number of static features, D_{dynamic} is the number of dynamic features, and T is the number of time bins. The MTNN has initial layers that are shared by all neurons, and each neuron has its designated final fully-connected layer.

Given the feature vectors x_{dynamic} and x_{static} for session s and neuron u, the model predicts the firing rates y^{pred} by:

$$e_{\text{static}} = f(w_{\text{static}}^T x_{\text{static}} + b_{\text{static}}) \tag{2}$$

$$e_{\text{dynamic}} = f(w_{\text{dynamic}}^T x_{\text{dynamic}} + b_{\text{dynamic}}) \tag{3}$$

$$h_t^{(forward)} = max(0, U_1 e_{\text{dynamic},t} + V_1 h_{t-1}^{(forward)} + b_{forward}) \tag{4}$$

$$h_t^{(backward)} = max(0, U_2 e_{\text{dynamic},t} + V_2 h_{t+1}^{(backward)} + b_{backward}) \tag{5}$$

$$y_t^{pred} = f(w_{(s,u)}^T \text{concat}(e_{\text{static}}, h_t^{(forward)}, h_t^{(backward)}) + b_{(s,u)}) \tag{6}$$

where f is the activation function. Eqn. (2) and Eqn. (3) are the shared fully-connected layers for static and dynamic covariates, respectively. Eqn. (4) and Eqn. (5) are the shared one-layer bidirectional recurrent neural networks (RNNs) for dynamic covariates, and Eqn. (6) is the neuron-specific fully-connected layer, indexed by (s, u). Each part of the MTNN architecture can have an arbitrary number of layers. For our analysis, we used two fully-connected shared layers for static covariates (Eqn. (2)) and three-layer bidirectional RNNs for dynamic covariates, with the embedding size set to 64.

Model training

The model was implemented in PyTorch and trained on a single GPU. The training was performed using Stochastic Gradient Descent on the Poisson negative loglikelihood (Poisson NLL) loss with learning rate set to 0.1, momentum set to 0.9, and weight decay set to 10^{-15}. We used a learning rate scheduler such that the learning rate for the i-th epoch is 0.1×0.95^i, and the dropout

rate was set to 0.15. We also experimented with mean squared error (MSE) loss instead of Poisson NLL loss, and the results were similar. The batch size was set to 512.

The dataset consists of 20 sessions, 239 neurons and 6480 active trials in total. For each session, 20% of the trials are used as the test data and the remaining trials are split 20:80 for the validation and training sets. During training, the performance on the held-out validation set is checked after every 3 passes through the training data. The model is trained for 100 epochs, and the model parameters with the best performance on the held-out validation set are saved and used for predictions on the test data.

Simulated experiments

For the simulated experiment in Figure 7-supplemental 2, we first trained GLMs on the same set of 239 responsive neurons from 20 sessions used for the analysis in Figure 7d and 7e, with a reduced set of covariates consisting of stimulus timing, stimulus side and contrast, first movement onset timing, feedback type and timing, wheel velocity, and mouse's priors for the current and previous trials. The kernels of the trained GLMs show the contribution of each of the covariates to the firing rates of each neuron. For each simulated neuron, we used these kernels of the trained GLM to simulate its firing rates for 400 randomly initialized trials. The random trials were 1.5 seconds long with 50 ms bin width. For all trials, the first movement onset timing was set to 0.5 second after the start of the trial, and the stimulus contrast, side, onset timing and feedback type, timing were randomly sampled. We used wheel velocity traces and mouse's priors from real data for simulation. We finally ran the leave-one-out analyses with GLMs/MTNN on the simulated data and compared the effect sizes estimated by GLMs and MTNN.

Acknowledgments

This work was supported by grants from the Wellcome Trust (209558 and 216324), National Institutes of Health (1F32MH123010, 1U19NS123716, including a Diversity Supplement) and the Simons Foundation. We thank R. Poldrack, T. Zador, P. Dayan, and C. Hurwitz for helpful comments on the manuscript. The production of all IBL Platform Papers is led by a Task Force, which defines the scope and composition of the paper, assigns and/or performs the required work for the paper, and ensures that the paper is completed in a timely fashion. The Task Force members for this platform paper include authors SAB, GC, AC, MFD, HDL, MF, GM, LP, NR, MS, NS, MT, and SW.

Diversity Statement

We support inclusive, diverse and equitable conduct of research. One or more of the authors of this paper self-identifies as a member of an underrepresented ethnic minority in science. One or more of the authors self-identifies as a member of the LGBTQIA+ community.

References

Andrianova L, Yanakieva S, Margetts-Smith G, Kohli S, Brady ES, Aggleton JP, Craig MT. No evidence from complementary data sources of a direct projection from the mouse anterior cingulate cortex to the hippocampal formation. bioRxiv. 2022; .

Ashwood ZC, Roy NA, Stone IR, Churchland AK, Pouget A, Pillow JW, et al. Mice alternate between discrete strategies during perceptual decision-making. bioRxiv. 2021; p. 2020–10.

Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016; 533(7604). doi: 10.1038/533452a.

Barry C, Ginzberg LL, O'Keefe J, Burgess N. Grid cell firing patterns signal environmental novelty by expansion. Proceedings of the National Academy of Sciences. 2012; 109(43):17687–17692.

Batty E, Merel J, Brackbill N, Heitman A, Sher A, Litke A, Chichilnisky E, Paninski L. Multilayer recurrent network models of primate retinal ganglion cell responses. ICLR. 2016; .

Biderman D, Whiteway MR, Hurwitz C, Greenspan NR, Lee RS, Vishnubhotla A, Schartner M, Huntenburg JM, Khanal A, Meijer GT, et al. Lightning Pose: improved animal pose estimation via semi-supervised learning, Bayesian ensembling, and cloud-native open-source tools. bioRxiv. 2023; p. 2023–04.

Bragin A, Jando G, Nadasdy Z, van Landeghem M, Buzsáki G. Dentate EEG spikes and associated interneuronal population bursts in the hippocampal hilar region of the rat. Journal of Neurophysiology. 1995; 73(4):1691–1705. doi: 10.1152/jn.1995.73.4.1691.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nature genetics. 2001; 29(4):365–371.

Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, Ecker AS. Deep convolutional models improve predictions of macaque V1 responses to natural images. PLOS Computational Biology. 2019; 15(4):1–27. doi: 10.1371/journal.pcbi.1006897.

Campbell R, BakingTray; 2020. https://github.com/SainsburyWellcomeCentre/BakingTray, doi: https://doi.org/10.5281/zenodo.3631609.

Campbell R, StitchIt; 2021. https://github.com/SainsburyWellcomeCentre/StitchIt, doi: https://zenodo.org/badge/latestdoi/57851444.

Campbell R, Blot A, Rousseau C, Winter O, Lasagna; 2020. https://github.com/SainsburyWellcomeCentre/lasagna, doi: 10.5281/zenodo.3941894.

Chen G, Manson D, Cacucci F, Wills TJ. Absence of visual input results in the disruption of grid cell firing in the mouse. Current Biology. 2016; 26(17):2335–2342.

Chung JE, Joo HR, Fan JL, Liu DF, Barnett AH, Chen S, Geaghan-Breiner C, Karlsson MP, Karlsson M, Lee KY, Liang H, Magland JF, Pebbles JA, Tooker AC, Greengard LF, Tolosa VM, Frank LM. High-Density, Long-Lasting, and Multi-region Electrophysiological Recordings Using Polymer Electrode Arrays. Neuron. 2019; 101(1):21–31.e5. https://www.sciencedirect.com/science/article/pii/S0896627318309930, doi: https://doi.org/10.1016/j.neuron.2018.11.002.

Churchland AK, Kiani R, Chaudhuri R, Wang XJ, Pouget A, Shadlen MN. Variance as a signature of neural computations during decision making. Neuron. 2011; 69(4):818–31. http://www.ncbi.nlm.nih.gov/pubmed/21338889, doi: 10.1016/j.neuron.2010.12.037.

Churchland MM, Yu BM, Cunningham JP, Sugrue LP, Cohen MR, Corrado GS, Newsome WT, Clark AM, Hosseini P, Scott BB, Bradley DC, Smith MA, Kohn A, Movshon JA, Armstrong KM, Moore T, Chang SW, Snyder LH, Lisberger SG, Priebe NJ, et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. Nat Neurosci. 2010; 13(3):369–78. http://www.ncbi.nlm.nih.gov/pubmed/20173745, doi: 10.1038/nn.2501.

Dragoi G, Tonegawa S. Preplay of future place cell sequences by hippocampal cellular assemblies. Nature. 2011; 469(7330):397–401.

Economo MN, Clack NG, Lavis LD, Gerfen CR, Svoboda K, Myers EW, Chandrashekar J. A platform for brain-wide imaging and reconstruction of individual neurons. eLife. 2016; 5(e10566). doi: 10.7554/eLife.10566.

Erlich JC, Brunton BW, Duan CA, Hanks TD, Brody CD. Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat. Elife. 2015; 4:e05457.

Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA. Investigating the replicability of preclinical cancer biology. eLife. 2021; 10:e71601. doi: 10.7554/eLife.71601, publisher: eLife Sciences Publications, Ltd.

Faulkner M, Ephys Atlas GUI; 2020. https://github.com/int-brain-lab/iblapps/tree/master/atlaselectrophysiology.

Goard MJ, Pho GN, Woodson J, Sur M. Distinct roles of visual, parietal, and frontal motor cortices in memory-guided sensorimotor decisions. elife. 2016; 5:e13764.

Grosmark AD, Buzsáki G. Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences. Science. 2016; 351(6280):1440–1443.

Hafting T, Fyhn M, Molden S, Moser MB, Moser EI. Microstructure of a spatial map in the entorhinal cortex. Nature. 2005; 436(7052):801–806.

Harvey CD, Coen P, Tank DW. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature. 2012; 484(7392):62–68.

Izenman AJ. Reduced-rank regression for the multivariate linear model. Journal of multivariate analysis. 1975; 5(2):248–264.

Jia X, Siegle JH, Bennett C, Gale SD, Denman DJ, Koch C, Olsen SR. High-density extracellular probes reveal dendritic backpropagation and facilitate neuron classification. Journal of Neurophysiology. 2019; 121(5):1831–1847. https://doi.org/10.1152/jn.00680.2018, doi: 10.1152/jn.00680.2018, pMID: 30840526.

Jun JJ, Steinmetz NA, Siegle JH, Denman DJ, Bauza M, Barbarits B, Lee AK, Anastassiou CA, Andrei A, Aydın Ç, et al. Fully integrated silicon probes for high-density recording of neural activity. Nature. 2017; 551(7679):232–236.

Klein S, Staring M, Murphy K, Viergever MA, Pluim JPW. elastix: a toolbox for intensity based medical image registration. IEEE Transactions on Medical Imaging. 2010; 29(1):196–205. doi: 10.1109/TMI.2009.2035616.

Klionsky DJ. Developing a set of guidelines for your research field: a practical approach. Mol Biol Cell. 2016; 27(5):733–8. https://www.ncbi.nlm.nih.gov/pubmed/26915690, doi: 10.1091/mbc.E15-09-0618.

Kobak D, Brendel W, Constantinidis C, Feierstein CE, Kepecs A, Mainen ZF, Qi XL, Romo R, Uchida N, Machens CK. Demixed principal component analysis of neural population data. Elife. 2016; 5:e10989.

Li X, Ai L, Giavasis S, Jin H, Feczko E, Xu T, Clucas J, Franco A, Sólon Heinsfeld A, Adebimpe A, Vogelstein JT, Yan CG, Esteban O, Poldrack RA, Craddock C, Fair D, Satterthwaite T, Kiar G, Milham MP. Moving Beyond Processing and Analysis-Related Variation in Neuroscience. bioRxiv. 2021; https://www.biorxiv.org/content/early/2021/12/03/2021.12.01.470790, doi: 10.1101/2021.12.01.470790.

Lithgow GJ, Driscoll M, Phillips P. A long journey to reproducible results. Nature. 2017; 548(7668):387–388. https://www.ncbi.nlm.nih.gov/pubmed/28836615, doi: 10.1038/548387a.

Liu LD, Chen S, Hou H, West SJ, Faulkner M, Economo MN, Li N, Svoboda K, the International Brain Laboratory. Accurate localization of linear probe electrode arrays across multiple brains. eNeuro. 2021; 8(6). doi: 10.1523/ENEURO.0241-21.2021.

Liu Y, Dolan RJ, Kurth-Nelson Z, Behrens TE. Human replay spontaneously reorganizes experience. Cell. 2019; 178(3):640–652.

Lopes G, Bonacchi N, Frazão J, Neto JP, Atallah BV, Soares S, Moreira L, Matias S, Itskov PM, Correia PA, et al. Bonsai: an event-based framework for processing and controlling data streams. Frontiers in neuroinformatics. 2015; 9:7.

Lopes G, Farrell K, Horrocks EA, Lee CY, Morimoto MM, Muzzu T, Papanikolaou A, Rodrigues FR, Wheatcroft T, Zucca S, et al. Creating and controlling visual environments using BonVision. Elife. 2021; 10:e65541.

Lucanic M, Plummer WT, Chen E, Harke J, Foulger AC, Onken B, Coleman-Hulbert AL, Dumas KJ, Guo S, Johnson E, et al. Impact of genetic background and experimental reproducibility on identifying chemical compounds with robust longevity effects. Nature communications. 2017; 8(1):1–13.

Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, Bethge M. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature neuroscience. 2018; 21(9):1281–1289.

McIntosh LT, Maheswaranathan N, Nayebi A, Ganguli S, Baccus SA. Deep learning models of the retinal response to natural scenes. Advances in neural information processing systems. 2016; 29:1369.

Musall S, Kaufman MT, Juavinett AL, Gluf S, Churchland AK. Single-trial neural dynamics are dominated by richly varied movements. Nature neuroscience. 2019; 22(10):1677–1686.

Najafi F, Elsayed GF, Cao R, Pnevmatikakis E, Latham PE, Cunningham JP, Churchland AK. Excitatory and inhibitory subnetworks are equally selective during decision-making and emerge simultaneously during learning. Neuron. 2020; 105(1):165–179.

Nichols TE, Das S, Eickhoff SB, Evans AC, Glatard T, Hanke M, Kriegeskorte N, Milham MP, Poldrack RA, Poline JB, et al. Best practices in data analysis and sharing in neuroimaging using MRI. Nature neuroscience. 2017; 20(3):299–303.

Ólafsdóttir HF, Barry C, Saleem AB, Hassabis D, Spiers HJ. Hippocampal place cells construct reward related sequences through unexplored space. Elife. 2015; 4:e06063.

Penttonen M, Kamondi A, Sik A, Acsády L, Buzsáki G. Feed-forward and feed-back activation of the dentate gyrus in vivo during dentate spikes and sharp wave bursts. Hippocampus. 1997; 7(4):437–450.

Ragan T, Kadiri LR, Venkataraju KU, Bahlmann K, Sutin J, Taranda J, Arganda-Carreras I, Kim Y, Seung HS, Osten P. Serial two-photon tomography for automated ex vivo mouse brain imaging. Nat Methods. 2012; 9(3):255–8. doi: 10.1038/nmeth.1854.

Rajasethupathy P, Sankaran S, Marshel JH, Kim CK, Ferenczi E, Lee SY, Berndt A, Ramakrishnan C, Jaffe A, Lo M, Liston C, Deisseroth K. Projections from neocortex mediate top-down control of memory retrieval. Nature. 2015; 526(7575):653–659.

Raposo D, Kaufman MT, Churchland AK. A category-free neural population supports evolving demands during decision-making. Nature neuroscience. 2014; 17(12):1784–1792.

Rossant C, Winter O, Hunter M, Huntenburg J, Faulkner M, Wells M, Steinmetz N, Harris K, Bonacchi N, Alyx; 2021. https://github.com/cortex-lab/alyx.

Roth MM, Dahmen JC, Muir DR, Imhof F, Martini FJ, Hofer SB. Thalamic nuclei convey diverse contextual information to layer 1 of visual cortex. Nat Neurosci. 2016; 19(2):299–307.

Saalmann YB, Kastner S. Cognitive and perceptual functions of the visual thalamus. Neuron. 2011; 71(2):209–223.

Senzai Y, Buzsáki G. Physiological Properties and Behavioral Correlates of Hippocampal Granule Cells and Mossy Cells. Neuron. 2017; 93(3):691–704.e5. doi: 10.1016/j.neuron.2016.12.011.

Shin JD, Tang W, Jadhav SP. Dynamics of Awake Hippocampal-Prefrontal Replay for Spatial Learning and Memory-Guided Decision Making. Neuron. 2019; 104(6):1110–1125.e7. https://www.sciencedirect.com/science/article/pii/S0896627319307858, doi: https://doi.org/10.1016/j.neuron.2019.09.012.

Siegle JH, Jia X, Durand S, Gale S, Bennett C, Graddis N, Heller G, Ramirez TK, Choi H, Luviano JA, Groblewski PA, Ahmed R, Arkhipov A, Bernard A, Billeh YN, Brown D, Buice MA, Cain N, Caldejon S, Casal L, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. Nature. 2021; 592(7852):86–92.

Silva D, Feng T, Foster DJ. Trajectory events across hippocampal place cells require previous experience. Nature neuroscience. 2015; 18(12):1772–1779.

Steinmetz NA, Aydin C, Lebedeva A, Okun M, Pachitariu M, Bauza M, Beau M, Bhagat J, Böhm C, Broux M, Chen S, Colonell J, Gardner RJ, Karsh B, Kloosterman F, Kostadinov D, Mora-Lopez C, O'Callaghan J, Park J, Putzeys J, et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. Science. 2021; 372(6539):eabf4588.

Steinmetz NA, Zatka-Haas P, Carandini M, Harris KD. Distributed coding of choice, action and engagement across the mouse brain. Nature. 2019 Dec; 576(7786):266–273.

The International Brain Laboratory, iblvideo; 2021. https://github.com/int-brain-lab/iblvideo.

The International Brain Laboratory, pykilosort; 2021. https://github.com/int-brain-lab/pykilosort.

The International Brain Laboratory, Aguillon-Rodriguez V, Angelaki D, Bayer H, Bonacchi N, Carandini M, Cazettes F, Chapuis G, Churchland AK, Dan Y, Dewitt E, Faulkner M, Forrest H, Haetzel L, Hausser M, Hofer SB, Hu F, Khanal A, Krasniak C, Laranjeira I, et al. Standardized and reproducible measurement of decision-making in mice. eLife. 2021; 10:e63711. doi: 10.7554/eLife.63711.

The International Brain Laboratory, Banga K, Boussard J, Chapuis G, Faulkner M, Harris K, Huntenburg J, Hurwitz C, Lee HD, Paninski L, Rossant C, Roth N, Steinmetz N, Windolf C, Winter O. Spike sorting pipeline for the International Brain Laboratory. figshare. 2022; doi: 10.6084/m9.figshare.19705522.

Tolhurst DJ, Movshon JA, Dean AF. The statistical reliability of signals in single neurons in cat and monkey visual cortex. Vision Res. 1983; 23(8):775–85. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=6623937, doi: 0042-6989(83)90200-6 [pii].

Tomar S. Converting video formats with FFmpeg. Linux Journal. 2006; 2006(146):10.

Tsui J, Schwartz N, Ruthazer ES. A developmental sensitive period for spike timing-dependent plasticity in the retinotectal projection. Frontiers in synaptic neuroscience. 2010; 2:13.

Turk-Browne NB. The hippocampus as a visual area organized by space and time: A spatiotemporal similarity hypothesis. Vision research. 2019; 165:123–130.

Urai AE, Doiron B, Leifer AM, Churchland AK. Large-scale neural recordings call for new insights to link brain and behavior. Nature Neuroscience. 2022 Jan; 25(1):11–19. https://doi.org/10.1038/s41593-021-00980-9, doi: 10.1038/s41593-021-00980-9.

Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, Karp NA, Kas MJ, Schielzeth H, Van de Casteele T, Würbel H. Reproducibility of animal research in light of biological variation. Nature Reviews Neuroscience. 2020; 21(7):384–393. doi: 10.1038/s41583-020-0313-3.

de Vries SEJ, Lecoq JA, Buice MA, Groblewski PA, Ocker GK, Oliver M, Feng D, Cain N, Ledochowitsch P, Millman D, Roll K, Garrett M, Keenan T, Kuan L, Mihalas S, Olsen S, Thompson C, Wakeman W, Waters J, Williams D, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. Nature Neuroscience. 2020; 23(1):138–151. doi: 10.1038/s41593-019-0550-9.

Waaga T, Agmon H, Normand VA, Nagelhus A, Gardner RJ, Moser MB, Moser EI, Burak Y. Grid-cell modules remain coordinated when neural activity is dissociated from external sensory cues. Neuron. 2022; .

Wang Q, Ding SL, Li Y, Royall J, Feng D, Lesnar P, Graddis N, Naeemi M, Facer B, Ho A, Dolbeare T, Blanchard B, Dee N, Wakeman W, Hirokawa KE, Szafer A, Sunkin SM, Oh SW, Bernard A, Phillips JW, et al. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. Cell. 2020; 181(4):936–953.e20. doi: 10.1016/j.cell.2020.04.007.

West SJ, BrainRegister; 2021. https://github.com/stevenjwest/brainregister.

Zhang LI, Tao HW, Holt CE, Harris WA, Poo Mm. A critical window for cooperation and competition among developing retinotectal synapses. Nature. 1998; 395(6697):37–44.

Zhao Z, Zhu H, Li X, Sun L, He F, Chung JE, Liu DF, Frank L, Luan L, Xie C. Ultraflexible electrode arrays for months-long high-density electrophysiological mapping of thousands of neurons in rodents. Nature Biomedical Engineering. 2022 Oct; https://www.nature.com/articles/s41551-022-00941-y, doi: 10.1038/s41551-022-00941-y.

Supplementary figures

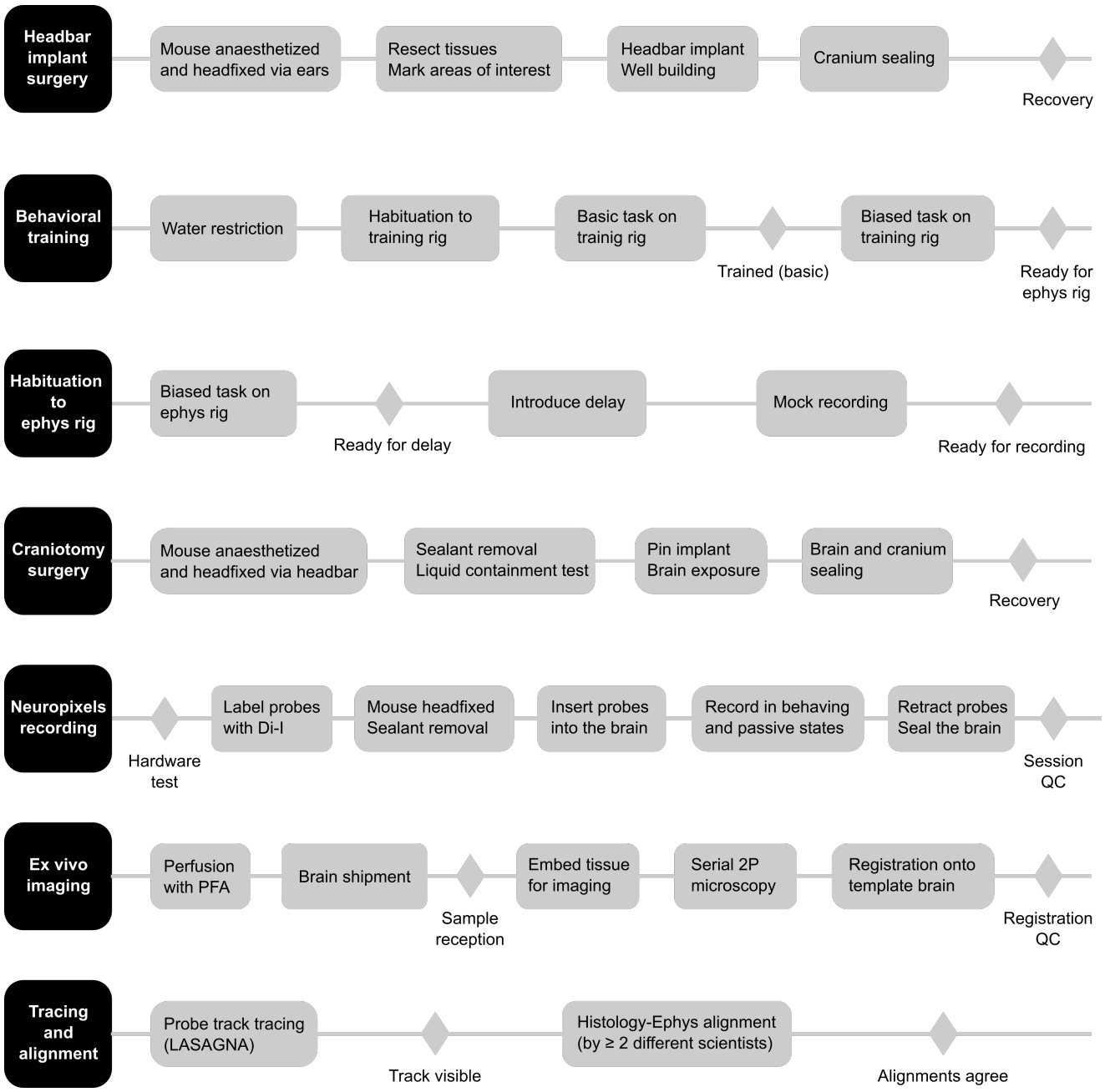

Figure 1–Figure supplement 1. Detailed experimental pipeline for the Neuropixels experiment. The experiment follows the main steps indicated in the left-hand black squares in chronological order from top to bottom. Within each main step, actions are undertaken from left to right; diamond markers indicate points of control.

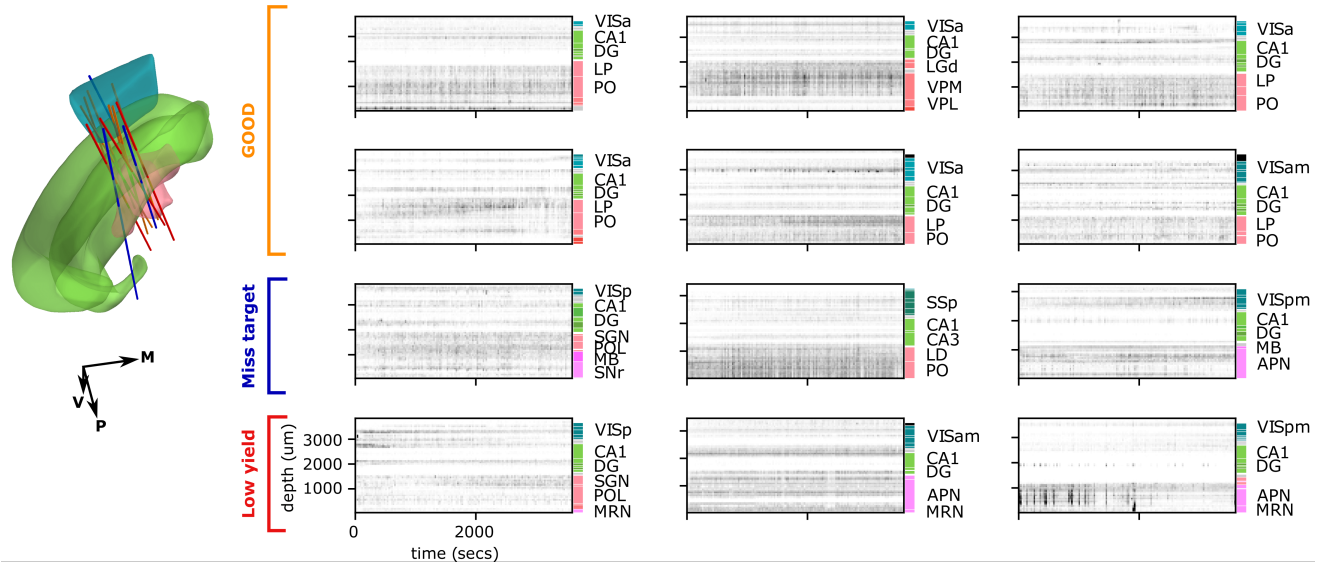

Figure 1–Figure supplement 2. (Left) 3D schematic of the probe insertions of the repeated site from 12 mice. Colors correspond to the quality of the probe insertion: good (yellow); blue (miss target); red (low yield). (Right) Spiking activity qualitatively appears heterogeneous across recordings. Example raster plots of neural activity recorded from the repeated site in 12 mice. The raster plots in the first top two rows originate from sessions marked as being of good quality. The middle and bottom rows are raster plots from recordings that were excluded, based either on the probe misplacement, or the low number of detected units. Allen Mouse CCF Labels: Anterior pretectal nucleus (APN); Dentate Gyrus (DG); Field CA1 (CA1); Field CA3 (CA3); Lateral dorsal nucleus of the thalamus (LD); Dorsal part of the lateral geniculate complex (LGd); Lateral posterior nucleus of the thalamus (LP); Midbrain (MB); Midbrain reticular nucleus (MRN); Posterior complex of the thalamus (PO); Posterior limiting nucleus of the thalamus (POL); Suprageniculate nucleus (SGN); Substantia nigra, reticular part (SNr); Primary somatosensory area (SSp); Ventral posterolateral nucleus of the thalamus (VPL); Ventral posteromedial nucleus of the thalamus (VPM); Anterior area (VISa); Anteromedial visual area (VISam); Posteromedial visual area (VISpm).

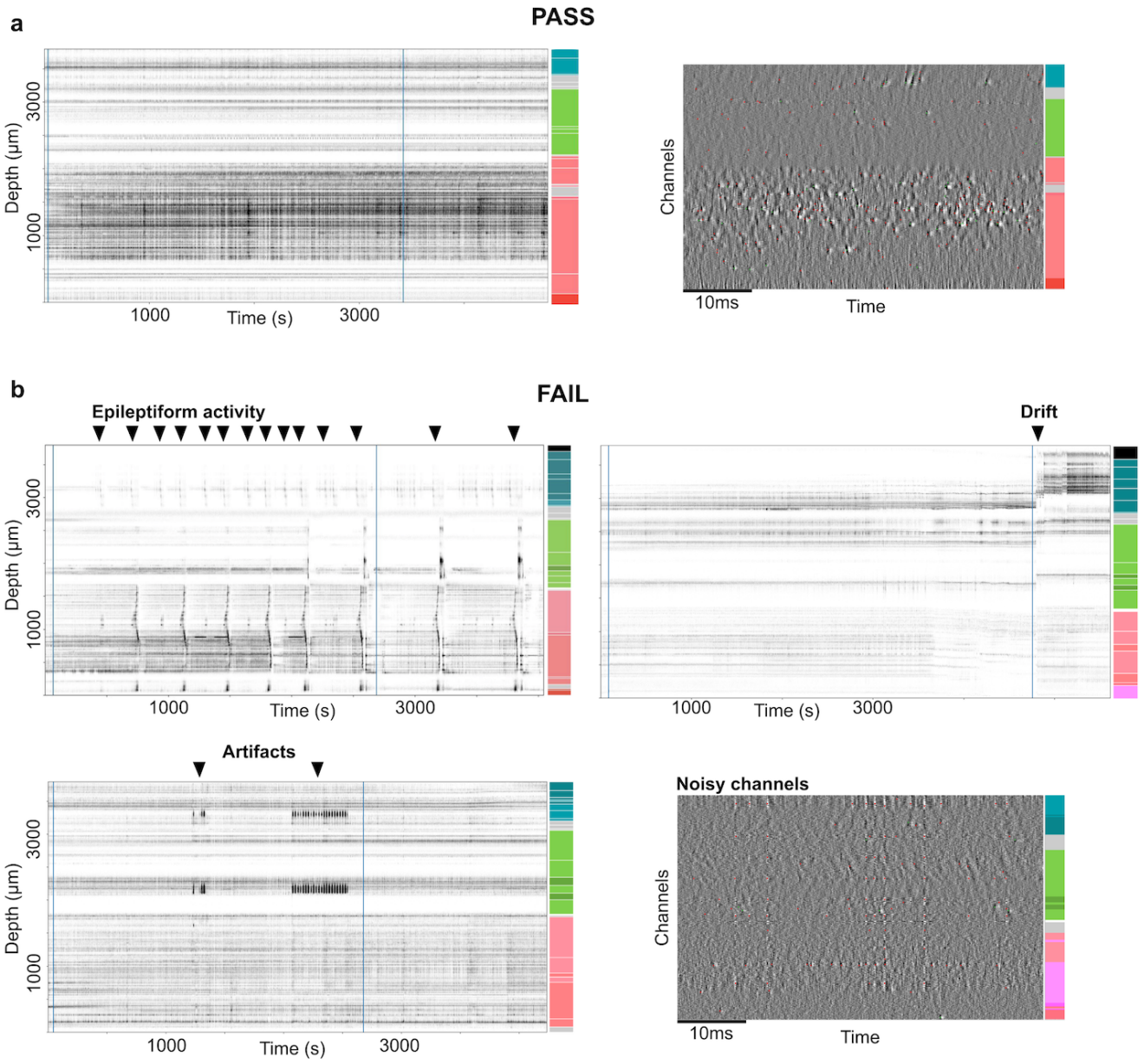

Figure 1–Figure supplement 3. (a) Example raster (left) and raw electrophysiology data snippet (right) for a recording that passes quality control. The blue lines on the raster plot mark the start and end of the behavioral task. **(b)** Example raster and raw data snippets for four recordings that fail quality control; either because of the presence of epileptic seizures (top-left), pronounced drift (top-right), artifacts (bottom-left), or large number of noisy channels (bottom-right).

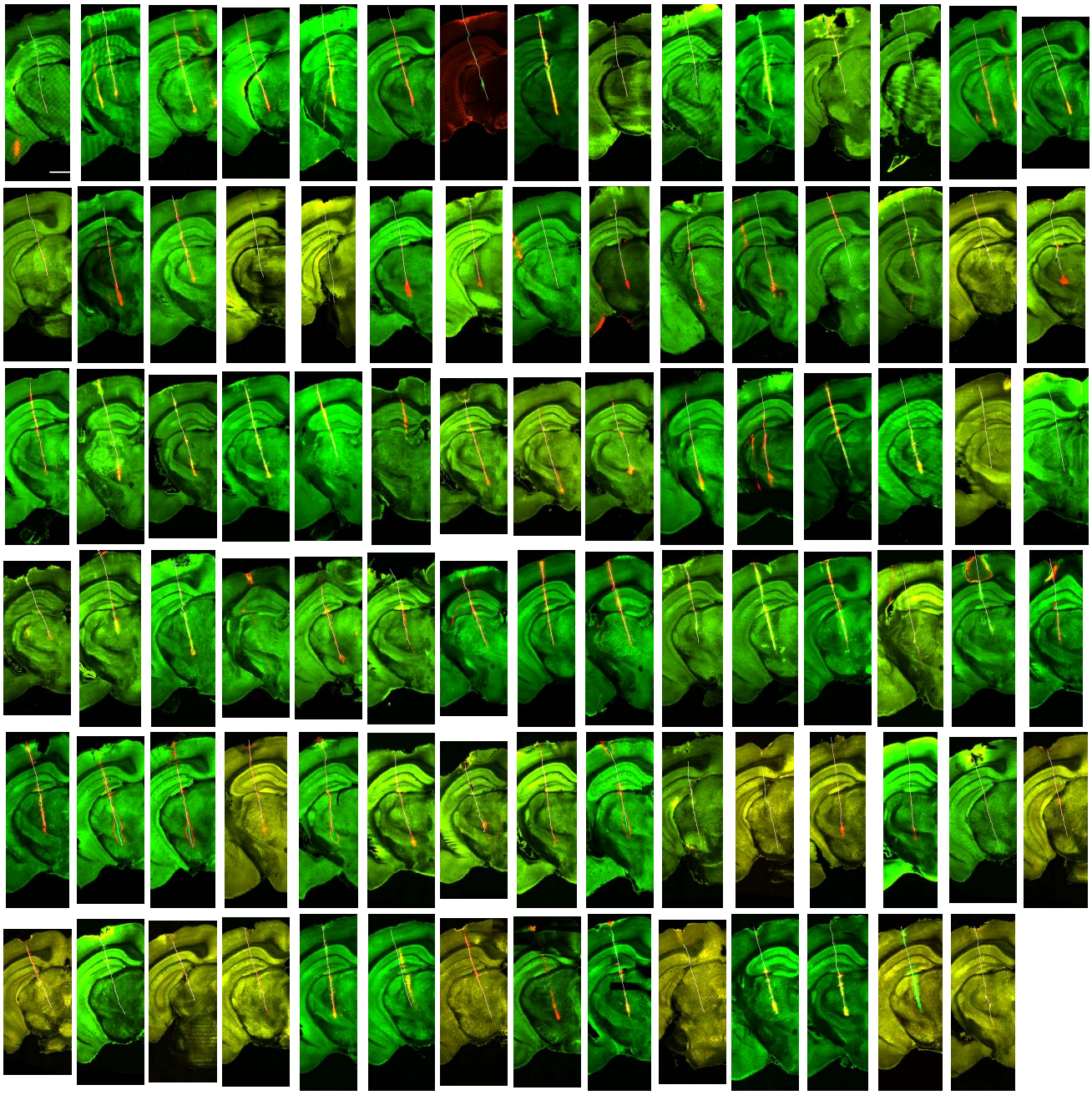

Figure 2–Figure supplement 1. Plots of all subjects with a repeated site insertion that were included in the analysis of probe placement. Coronal tilted slices are made along the linearly interpolated best-fit to the histology insertion, shown through the raw histology (green: auto-fluorescence data for image registration; red: cm-DiI fluorescence signal marking probe tracks). Traced probe tracks are highlighted in white. Scale bar: 1mm.

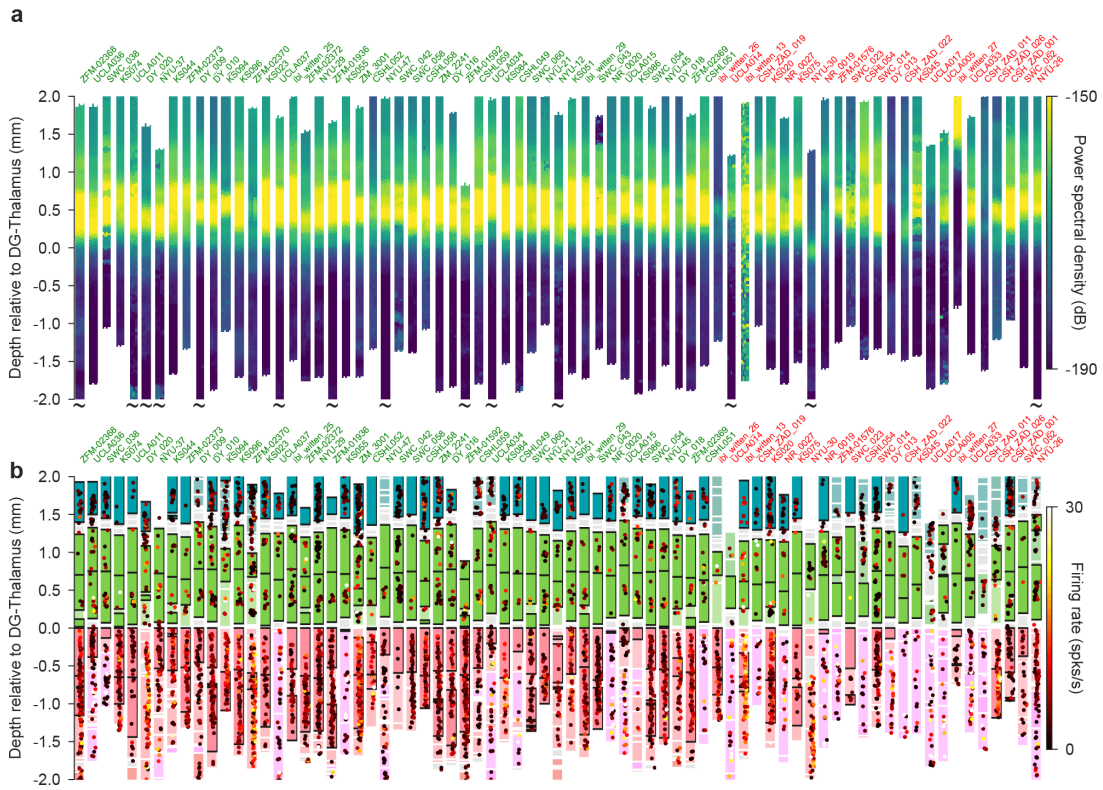

Figure 3–Figure supplement 1. (a, b) Probe plots as in Figure 3a,b. Above each probe plot is the name of the mouse, the color indicates whether the recording passed QC (green is pass, red is fail).

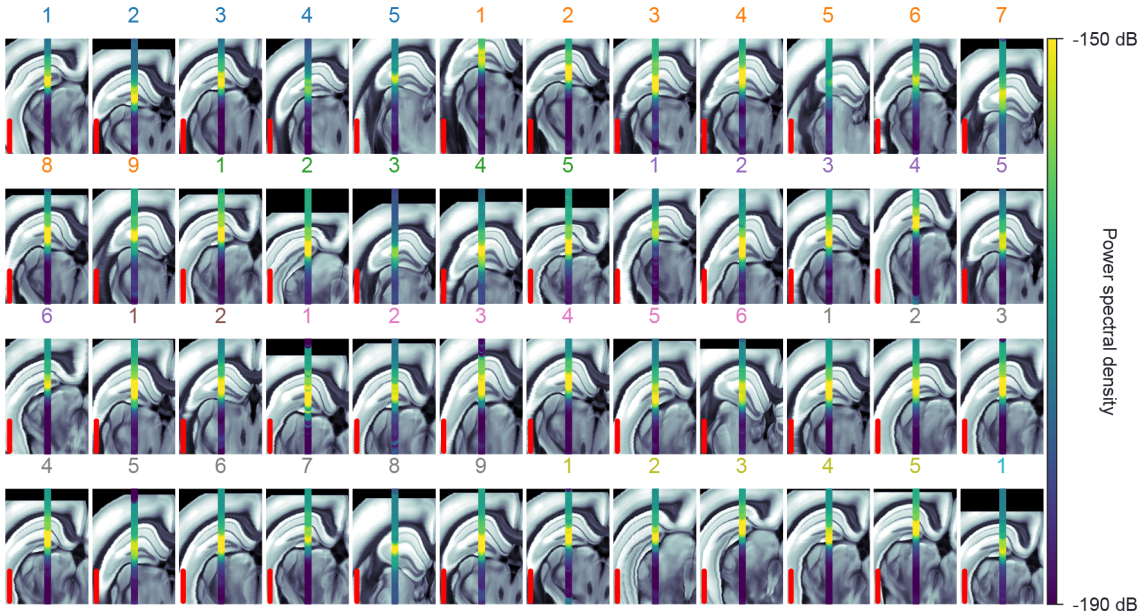

Figure 3–Figure supplement 2. Power spectral density between 20 and 80 Hz recorded along each probe shown in figure 3 overlaid on a coronal slice. Each coronal slice has been rotated so that the probe lies along the vertical axis. Colors correspond to probe insertions belonging to a single lab (Berkeley - blue; Champalimaud - orange; CSHL (C) - green; CSHL (Z) - red; NYU - purple; Princeton - brown; SWC - pink; UCL - grey; UCLA - yellow; UW - teal). Numbers above the image denote a recording session for individual mice. Red line is a 1 mm scalebar.

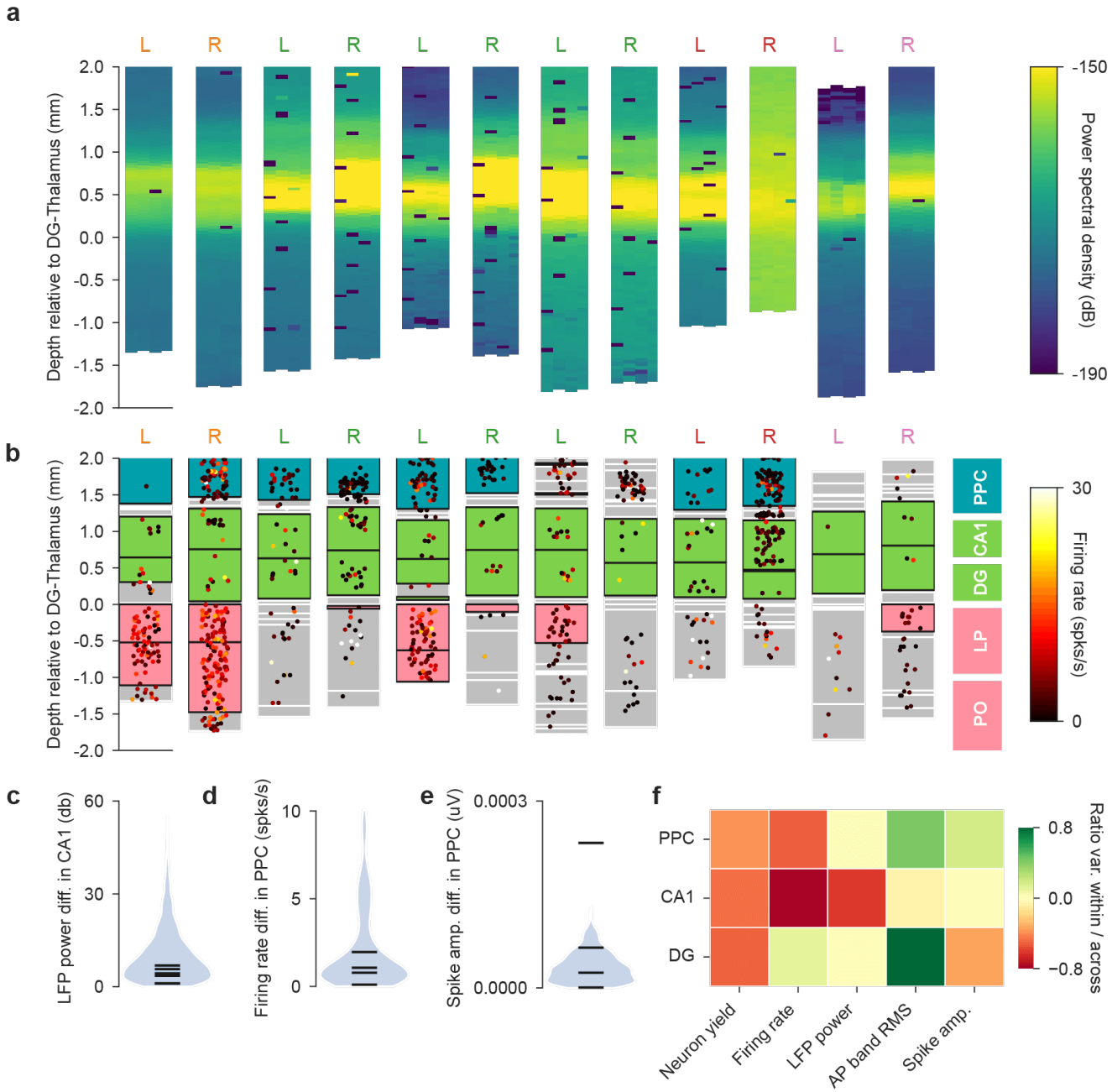

Figure 3–Figure supplement 3. Bilateral recordings of the repeated site in both hemispheres show within-animal variance is often smaller than across-animal variance. **(a, b)**, Power spectral density and neural activity of all bilateral recordings. L and R indicate the left and right probe of each bilateral recording. Each L/R pair is recorded simultaneously. The color indicates the lab (lab-color assignment identical to figure 3). **(c)** Within-animal variance is smaller than across-animal variance for LFP power. The across-animal variance is depicted as the distribution of all pair-wise absolute differences in LFP power between any two recordings in the entire dataset (blue shaded violin plot). The black horizontal ticks indicate where the bilateral recordings (within-animal variance) fall in this distribution. **(d, e)** Violin plots for firing rate and spike amplitude in VISa/am, similar analysis as in (c). **(f)** Whether within or across animal variance is larger is dependent on the metric and brain region; red colors indicate that within < across and green colors within > across. Variance is quantified here as the interquartile distance of the distributions in c-e.

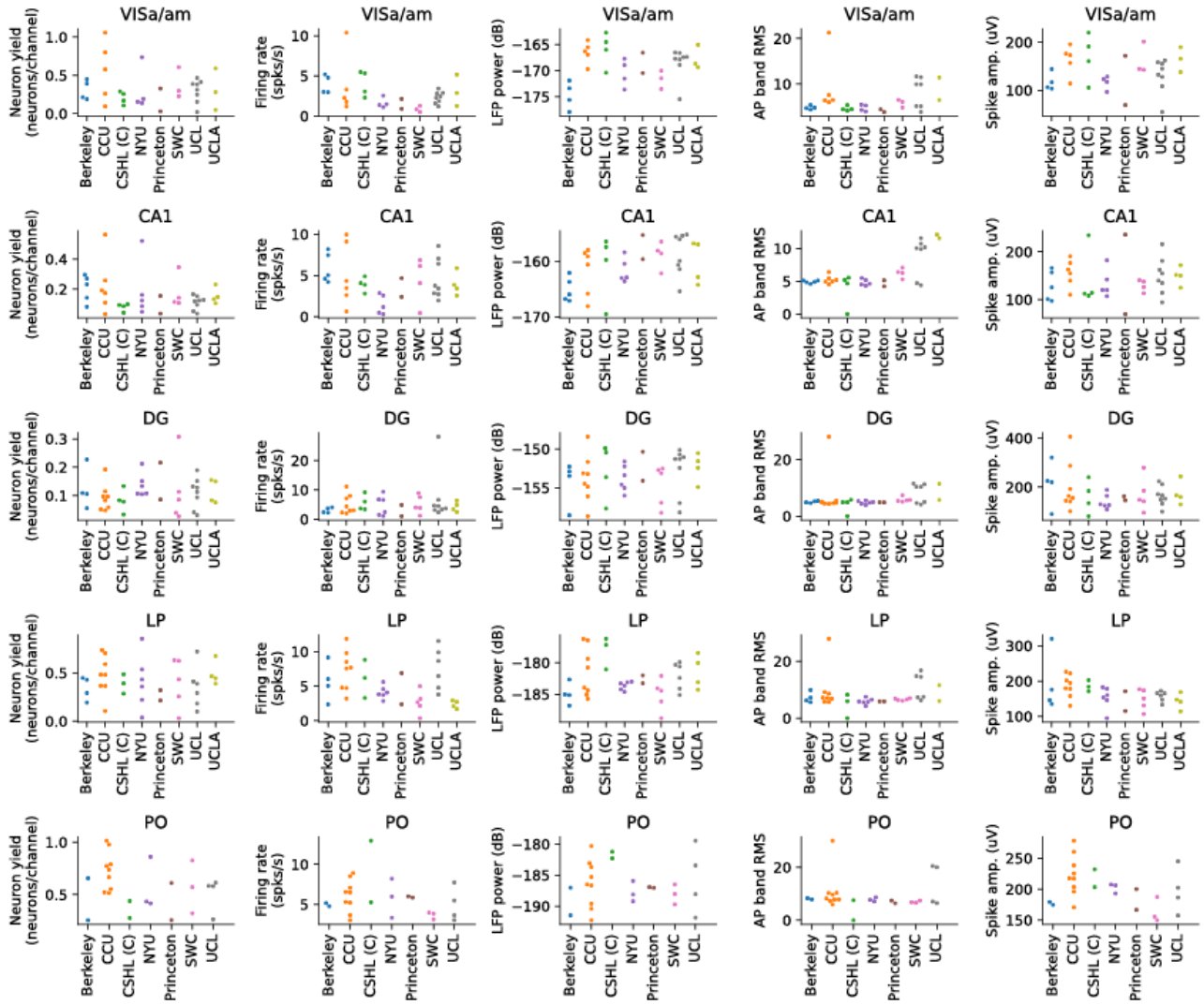

Figure 3–Figure supplement 4. All electrophysiological features (rows) per brain region (columns) that were used in the permutation test and decoding analysis of Figure 3. Each dot is a recording, colors indicate the laboratory.

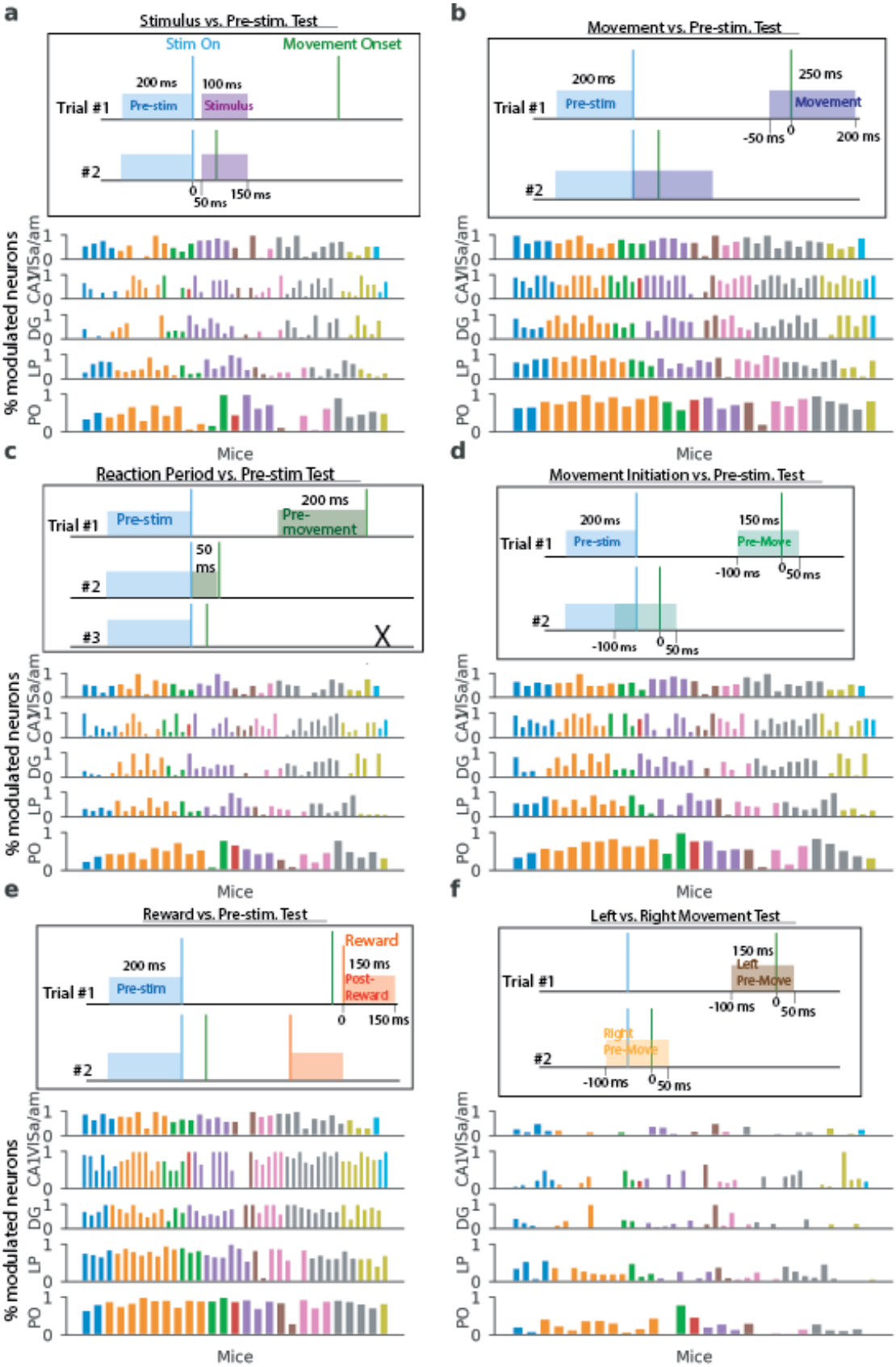

Figure 4–Figure supplement 1. (a)-(f) Schematics of six tests to quantify neuronal task modulation. Each panel includes a schematic of example trials showing potential caveats of each method. For instance, in **a** the stimulus period may or may not include movement, depending on the reaction time of the animal. In **c**, the reaction period is only calculated in trials with >50 ms between the stimulus and movement onset, and up to 200 ms before movement for trials with longer reaction times. Below each schematic, the proportion of task-modulated neurons for the test is shown, across mice and brain regions, colored by lab ID.

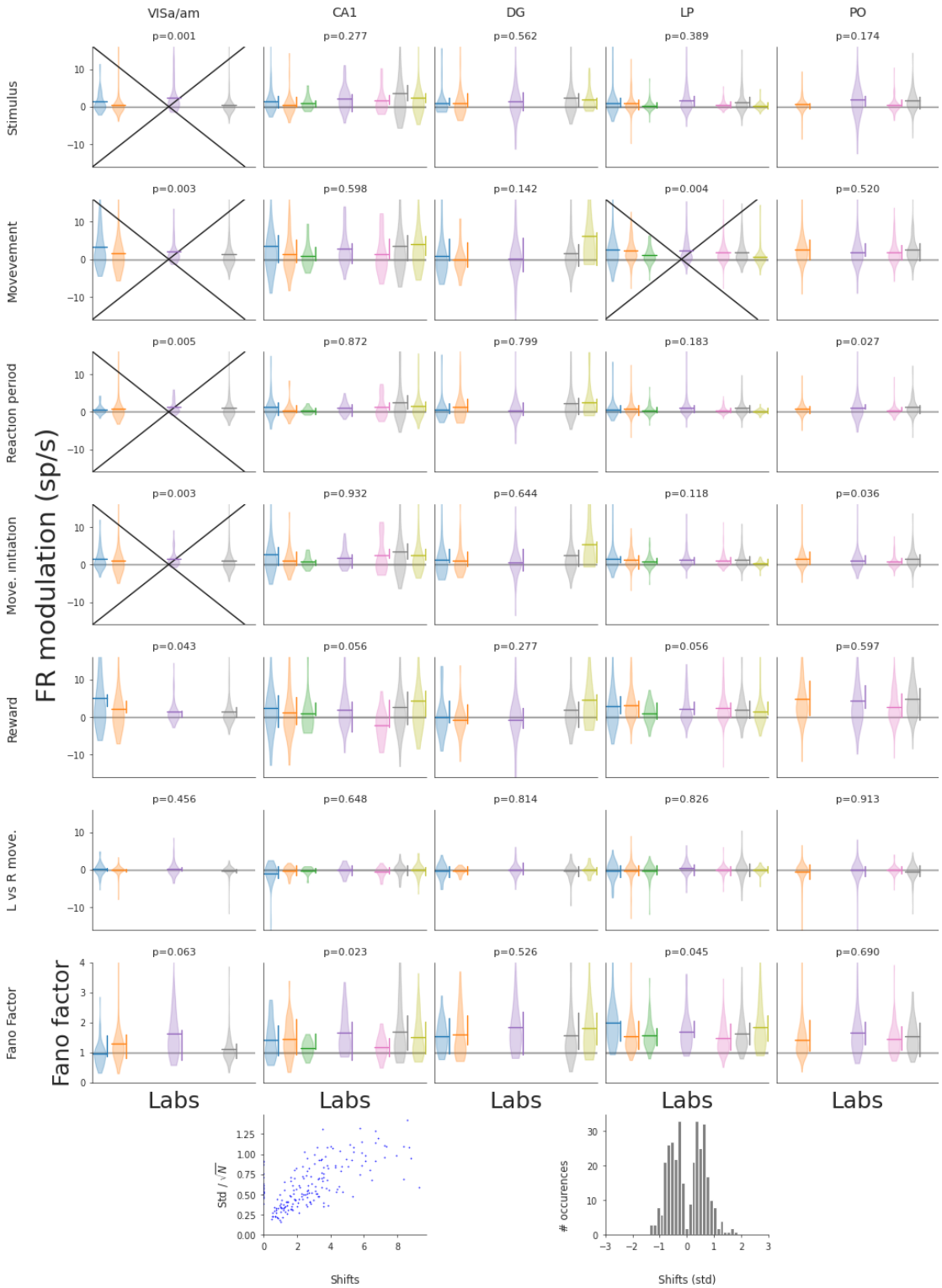

Figure 4–Figure supplement 2. For every test and every lab, we performed a power analysis to test how far the values of that lab would have to be shifted upwards or downwards to cause a significant permutation test (tests that were significant in that absence of such shifts are crossed out). Horizontal lines indicate the means of the lab distributions, vertical bars indicate the magnitude of the needed up- and downwards perturbations for a significant test (p-value < 0.01), the titles of the individual tests denote the p-value of the original unperturbed test. The magnitudes usually span a rather small range of permissible values, which means that our permutation testing procedure is sensitive to deviations of individual labs. The plot on the bottom left shows the correlation between shift size and standard deviation within the labs. In the bottom right is a histogram of the magnitude of shifts in units of the standard deviation of the corresponding distribution. Most shifts are below 1 standard deviations of the corresponding lab distribution

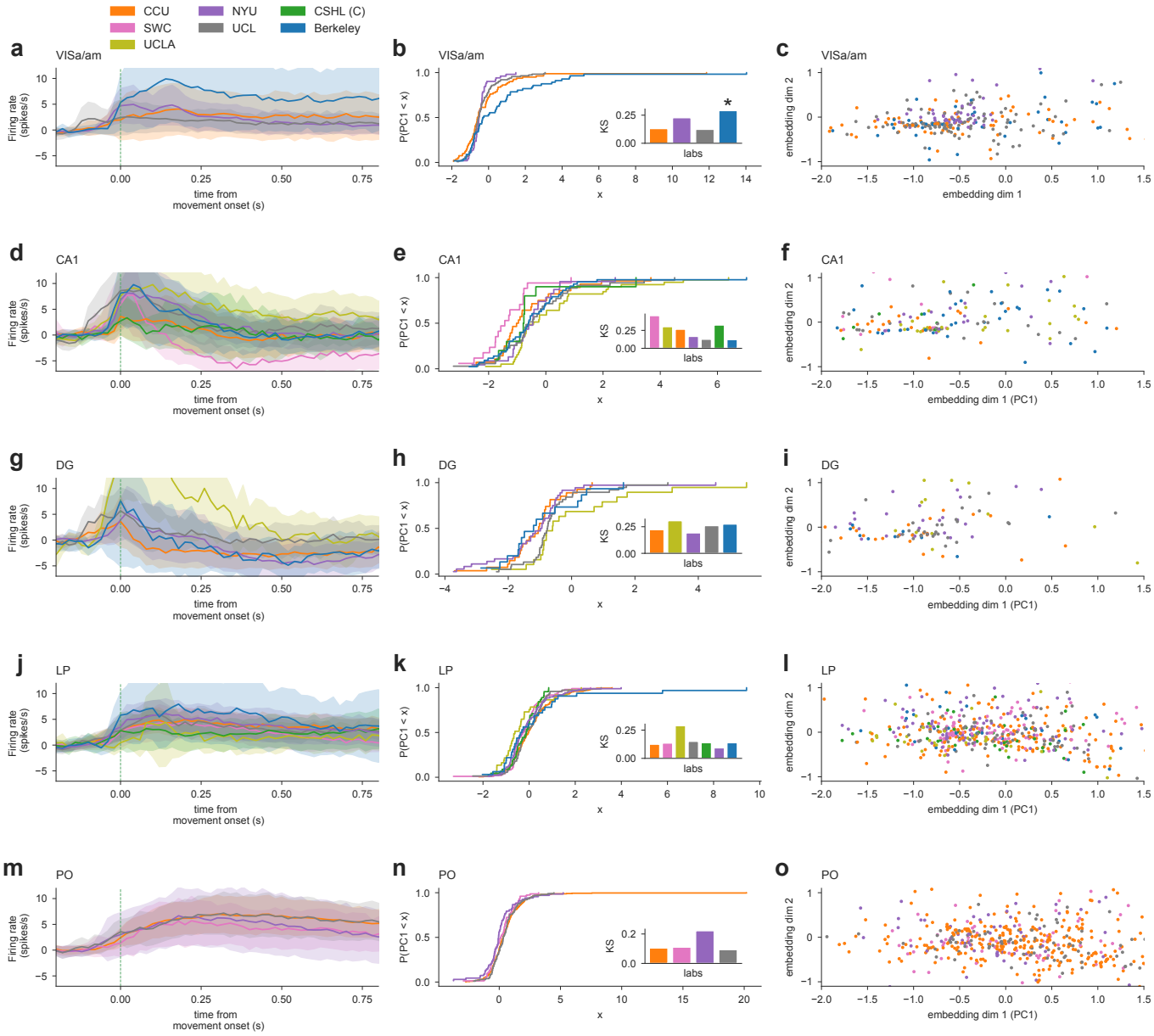

Figure 5–Figure supplement 1. Mean firing rates across all labs per region including VISa/am, CA1, DG, LP, and PO (panels a, d, g, j, m). In addition, the second column of panels (panels b, e, h, k, n) shows for each region the cumulative distribution function (CDF) of the first embedding dimension (PC1) per lab. The insets show the Kolmororov-Smirnov distance (KS) per lab from the distribution of all remaining labs pooled, annotated with an asterisk if $p < 0.01$ for the KS test (corrected for multiple comparisons per region). The third column of panels (c, f, i, l, o) displays the embedded activity of neurons from VISa/am, CA1, DG, LP, and PO. Only for Berkeley in VISa/am were dynamics significantly different from the mean of all remaining labs using the KS test.

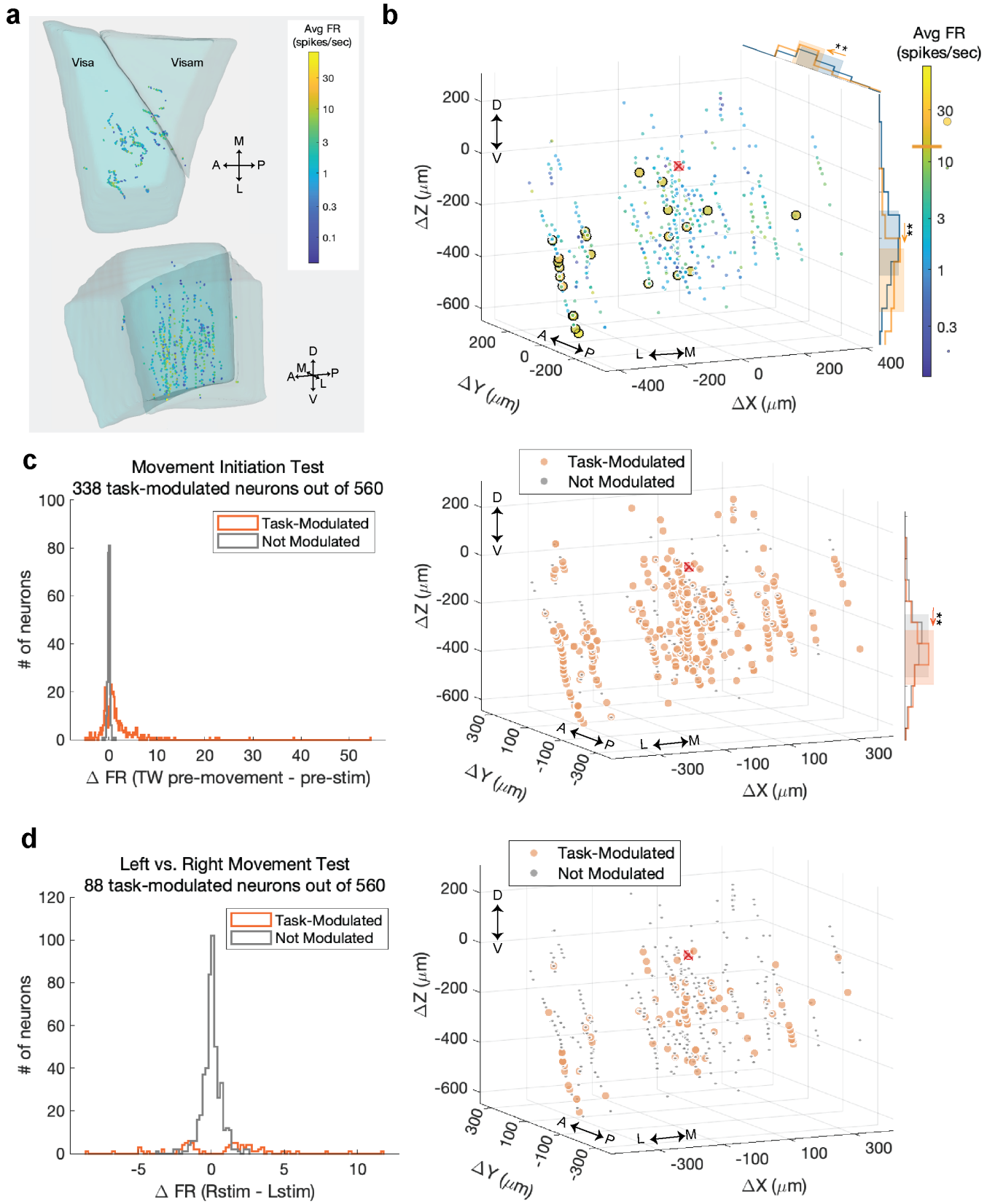

Figure 6–Figure supplement 1. High-firing and task-modulated VISa/am neurons were located in deeper layers than other VISa/am neurons. **(a-d)** Similar to Figure 6 but for VISa/am.

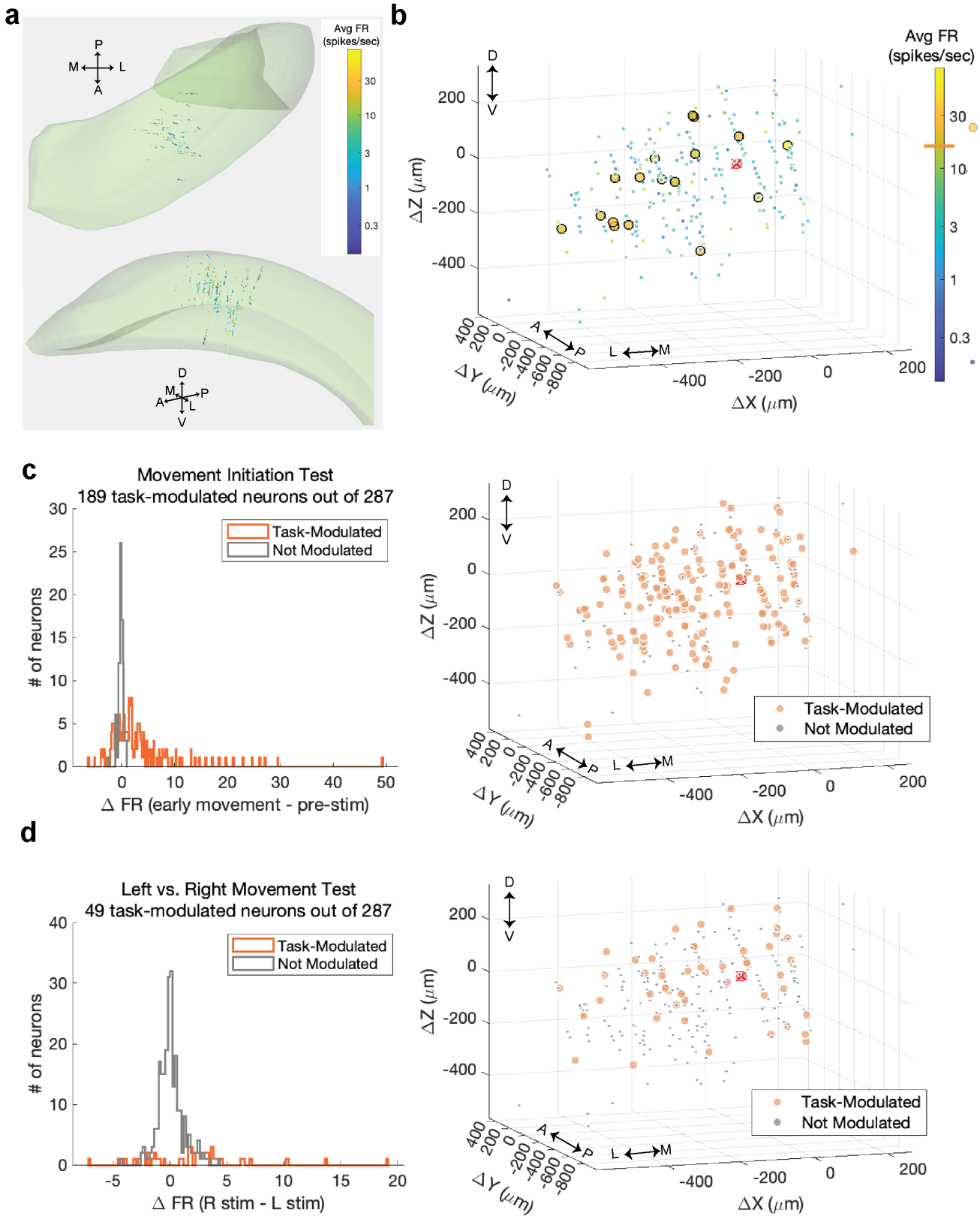

Figure 6–Figure supplement 2. High-firing and task-modulated CA1 neurons had no difference in spatial position or spike characteristics compared to regular firing and non-task-modulated neurons. **(a-d)** Similar to Figure 6 but for CA1.

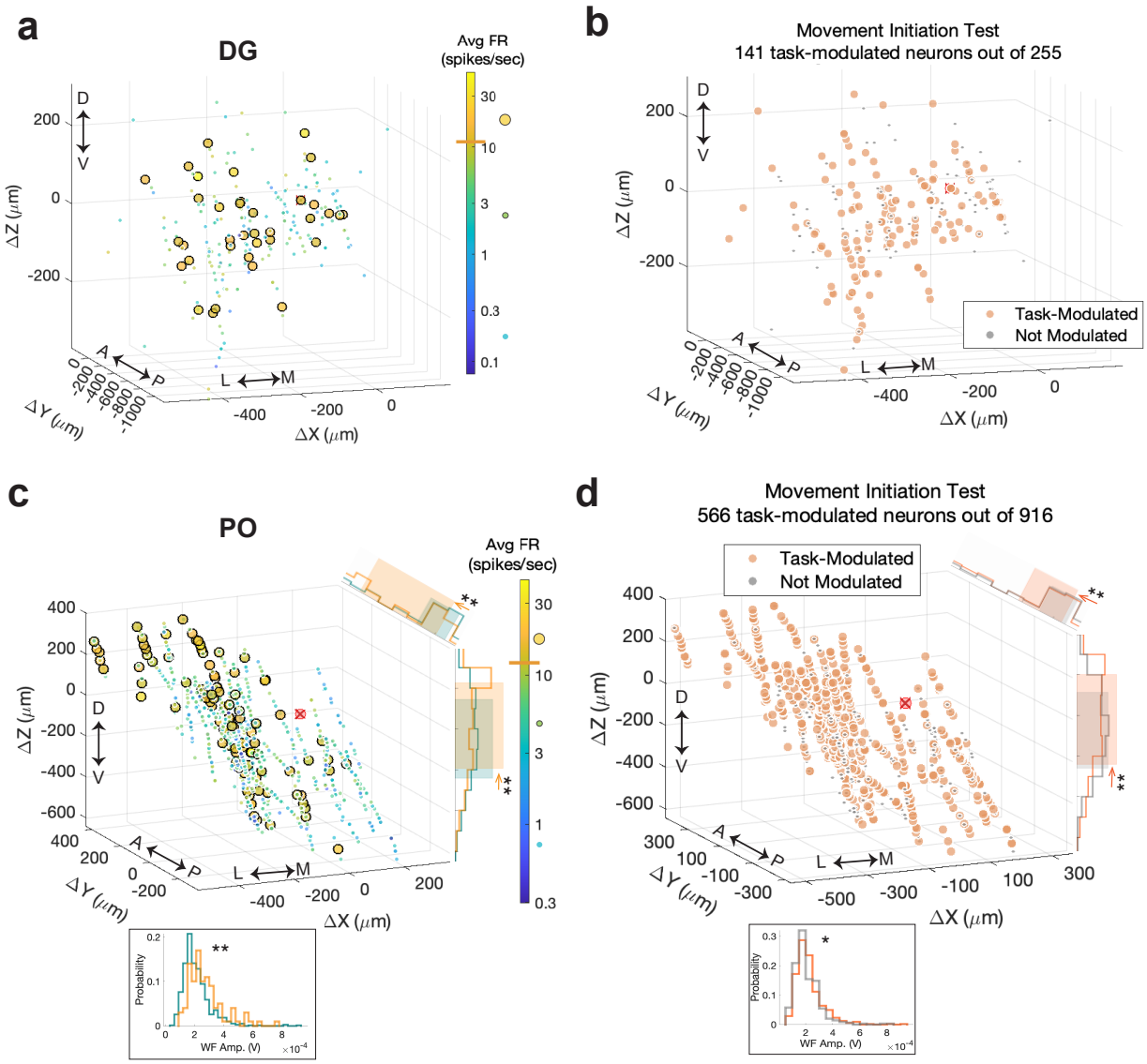

Figure 6–Figure supplement 3. Spatial positions and spike characteristics of outlier and task-modulated neurons in PO, but not DG, were different from other neurons. **(a)** Spatial positions of DG neurons plotted as distance from the planned target center of mass, indicated with the red x. Spatial positions and waveform features were not significantly different between the outliers (yellow) and the general population of neurons (blue). **(b)** Spatial positions and spike waveform features of task-modulated and non-modulated DG neurons are not different (using the movement initiation test). **(c-d)** Same as **a-b** but for PO neurons, where spatial position and spike amplitudes were significantly different between outliers and the general population of neurons, as well as between task-modulated and non-modulated neurons (as shown with the histograms).

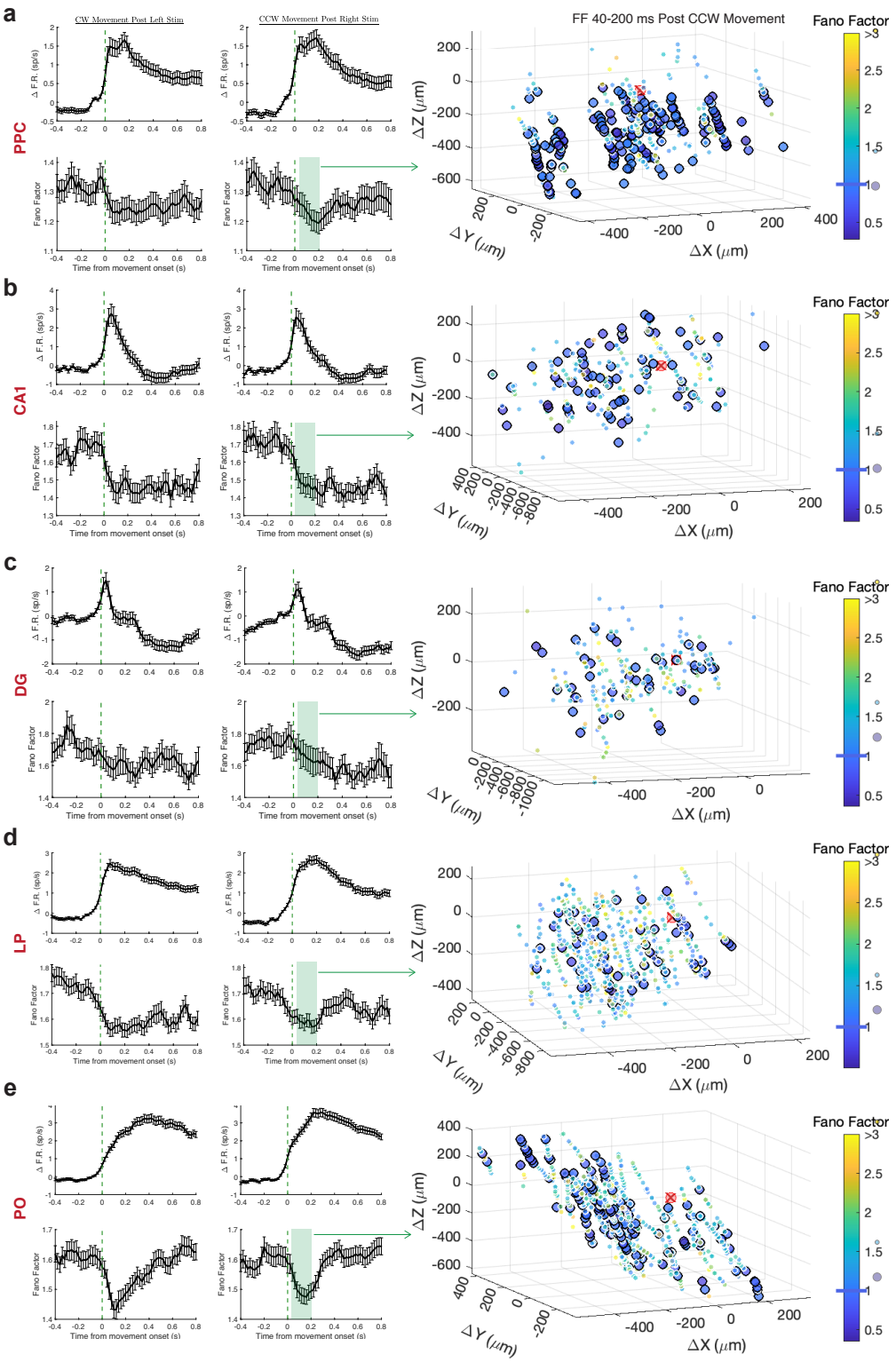

Figure 6–Figure supplement 4. Time-course and spatial position of neuronal Fano Factors. **(a)** *Left column*: Change in firing rate (top) and Fano Factor (bottom) averaged over all VISa/am neurons when aligned to movement onset after presentation of left or right full-contrast stimuli (correct trials only; Fano Factor calculation limited to neurons with a session-averaged firing rate >1 sp/sec). Error bars: standard error means between neurons. *Right column*: Neuronal Fano Factors (averaged over 40-200 ms post movement onset after right-side full-contrast stimuli) and their spatial positions. Larger circles indicate neurons with Fano Factor <1. **(b-e)** Same as **a** for CA1, DG, LP, and PO. Spatial position between high vs. low Fano Factor neurons was only significantly different in VISa/am (deeper neurons had lower Fano Factors) and PO (neurons with lower Fano Factors were positioned more laterally). In VISa/am, spike duration between high and low Fano Factor neurons was also significantly different, possibly due to cell type differences (neurons with shorter spike durations tended to have higher Fano Factors; histograms not shown).

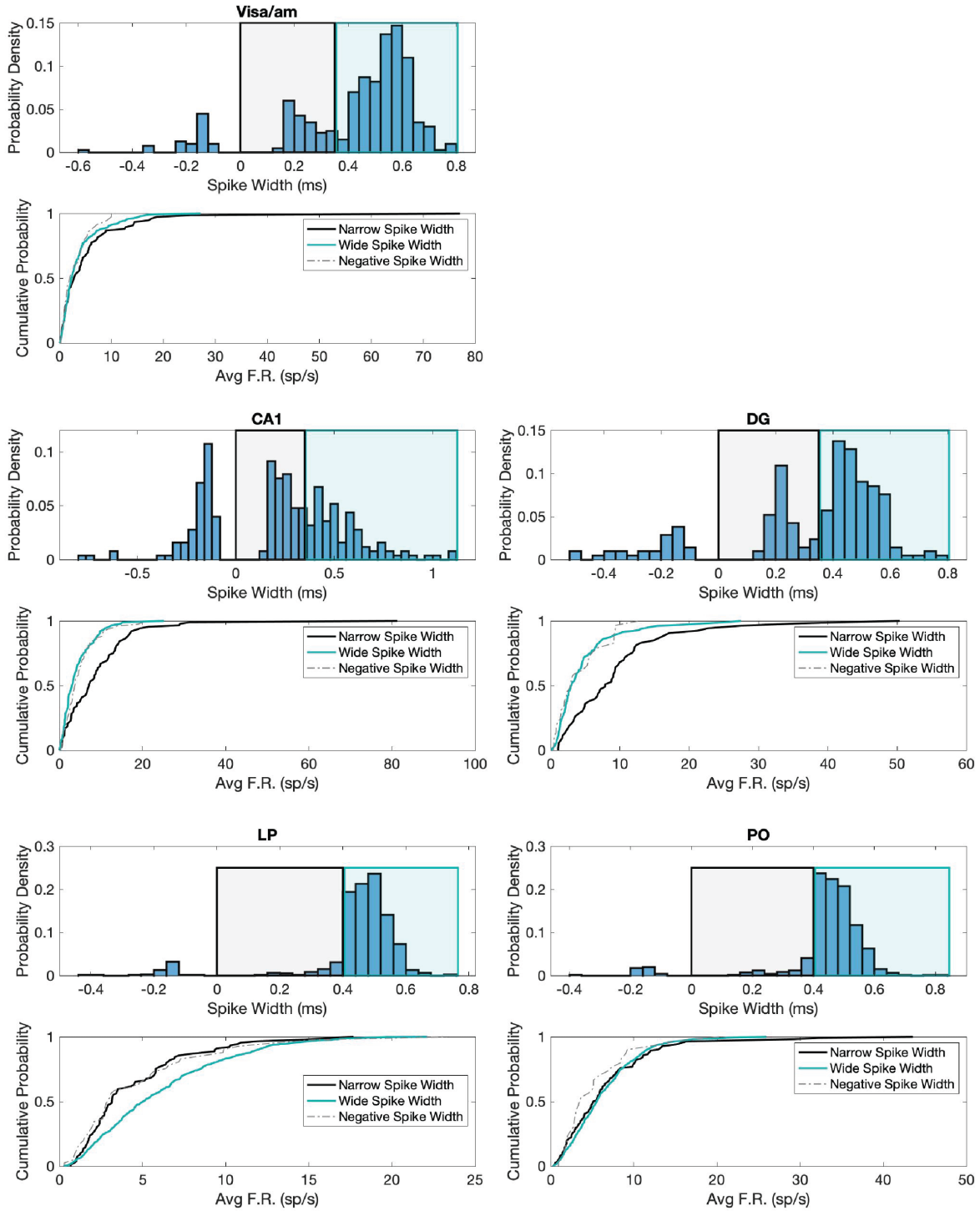

Figure 6–Figure supplement 5. High-firing neurons do not belong to a specific cell subtype. To identify putative Fast-Spiking (FS) and Regular-Spiking (RS) neuronal populations, we examined spike peak-to-trough durations (*Jia et al., 2019*). This distribution was bimodal in Visa/am, CA1, and DG, but not LP and PO (as expected from *Jia et al.* (*2019*)). This bimodality (indicated with the black and blue boxes) suggests distinct populations of FS and RS neurons only in cortical and hippocampal regions, which should have narrow (black) and wide (blue) spike widths, respectively. To confirm the distinct populations of FS and RS neurons, we next plotted the cumulative probability of firing rate for these two putative neuronal categories. Indeed, in cortex and hippocampus, neurons with narrow spikes tend to have higher firing rates (in black) while neurons with wider spikes have lower firing rates (in blue). In contrast, in LP and PO, we did not identify specific populations of neuronal subtypes using the spike waveform (to our knowledge, this has not been done in previous work either). Importantly, even in cortex/hippocampus where putative RS and FS neurons are distinguishable, there is still a large firing rate overlap between these two groups, especially for firing rates above 10-15 sp/s (the firing rate threshold from Figure 6 and supplemental figures). Hence, high-firing neurons do not seem to belong to a specific neuronal subtype.

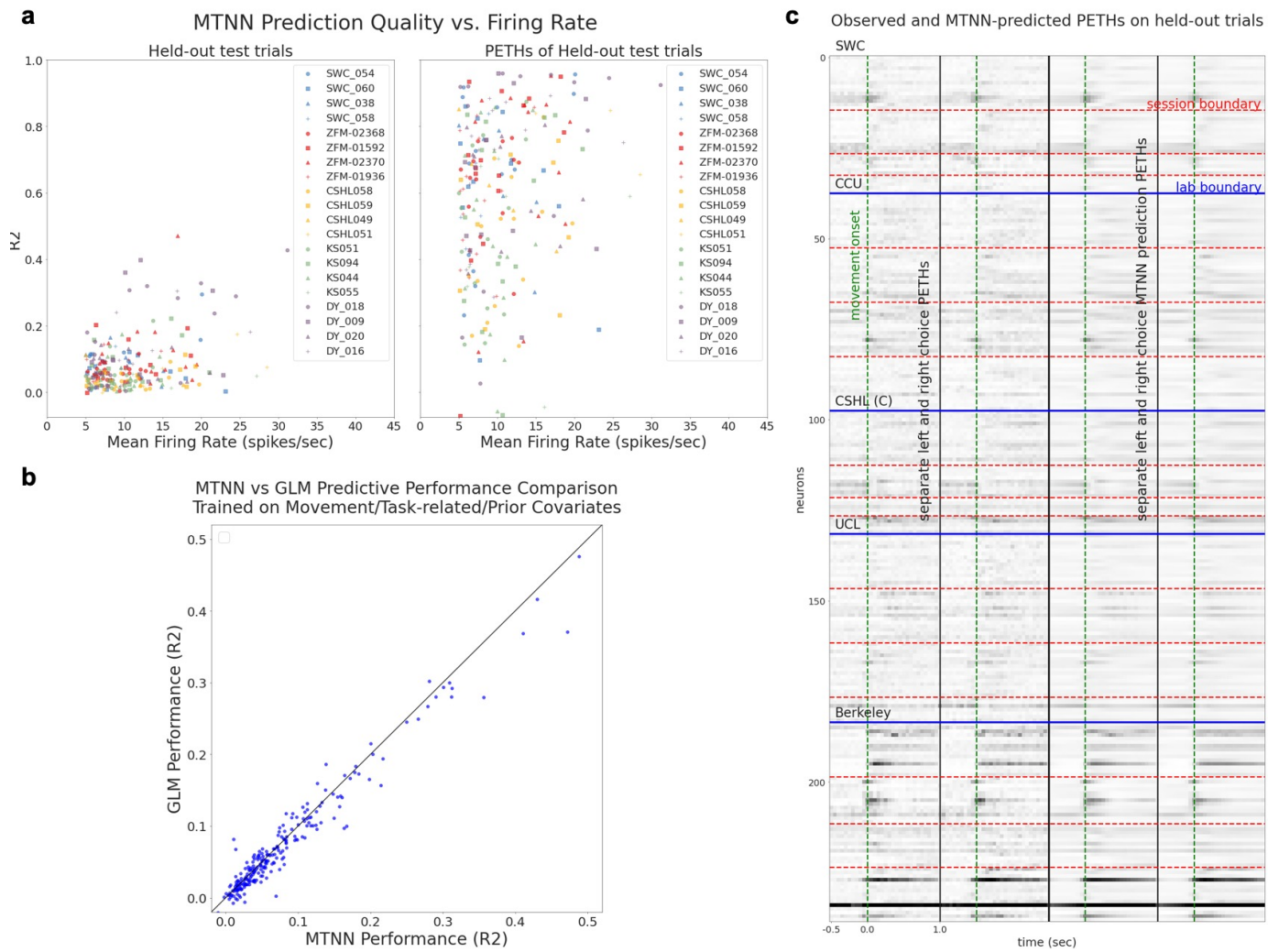

Figure 7–Figure supplement 1. (a) For each neuron in each session, we plot the MTNN prediction quality on held-out test trials against the firing rate of the neuron averaged over the test trials. Each lab/session is colored/shaped differently. R^2 values on concatenations of the held-out test trials are shown on the left, and those on PETHs of the held-out test trials on the right. **(b)** MTNN slightly outperforms GLMs on predicting the firing rates of held-out trials when trained on movement/task-related/prior covariates. **(c)** The left half shows for each neuron the trial averaged activity for left choice trials and next to it right choice trials. The vertical green lines show the first movement onset. The horizontal red lines separate recording sessions while the blue lines separate labs. The right half of each of these images shows the MTNN prediction of the left half. The trial-averaged MTNN predictions for held-out test trials captures visible modulations in the PETHs.

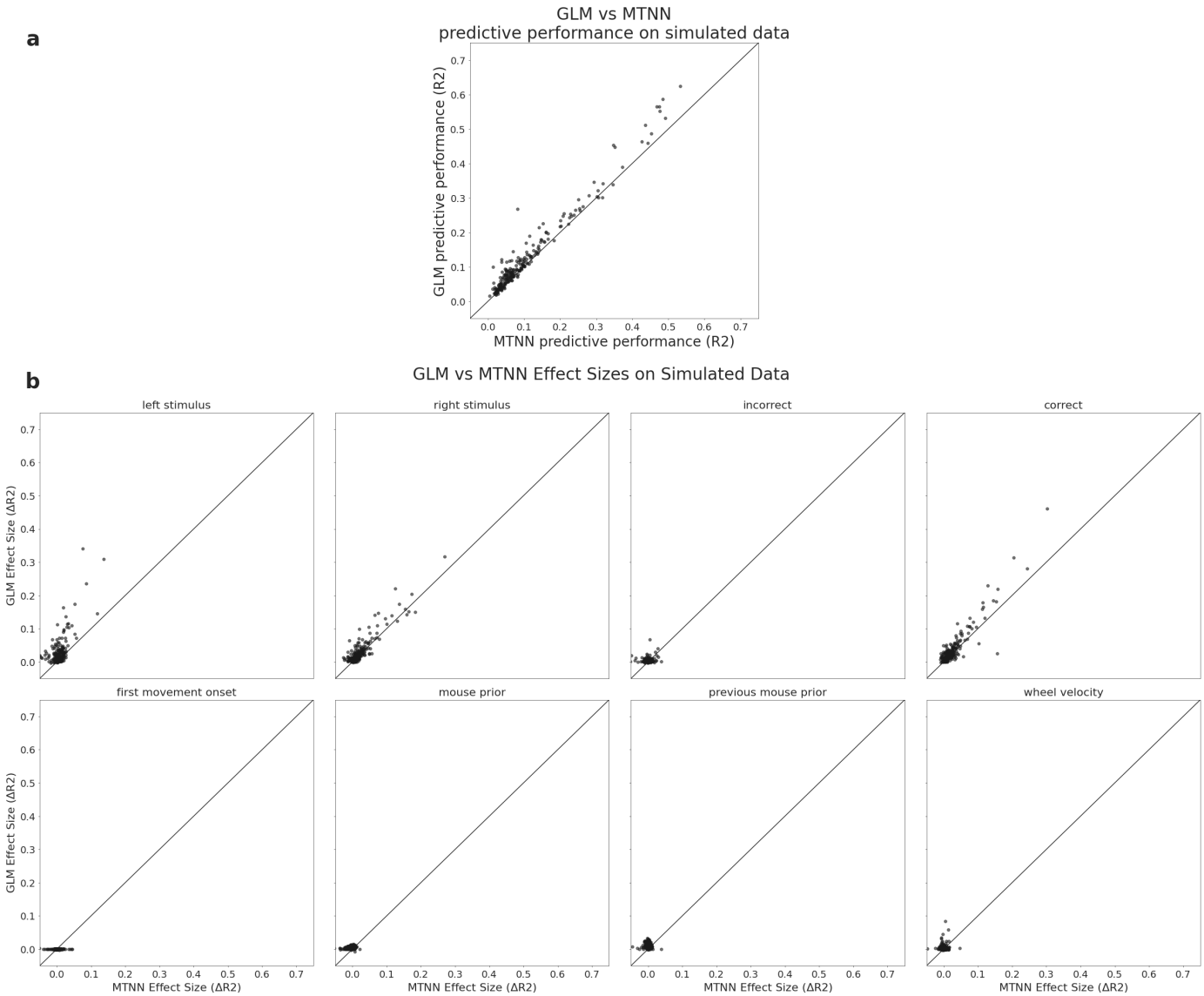

Figure 7–Figure supplement 2. To verify that the MTNN leave-one-out analysis is sensitive enough to capture effect sizes, we simulate data from GLMs and compare the effect sizes estimated by the MTNN and GLM leave-one-out analyses. We first fit GLMs to the same set of sessions that are used for the MTNN effect size analysis and then use the inferred GLM kernels to simulate data. **(a)** We show the scatterplot of the GLM and MTNN predictive performance on held-out test data, where each dot represents the predictive performance for one neuron. The MTNN prediction quality is comparable to that of GLMs. **(b)** We run GLM and MTNN leave-one-out analyses and compare the estimated effect sizes for eight covariates. The effect sizes estimated by the MTNN and GLM leave-one-out analyses are comparable.

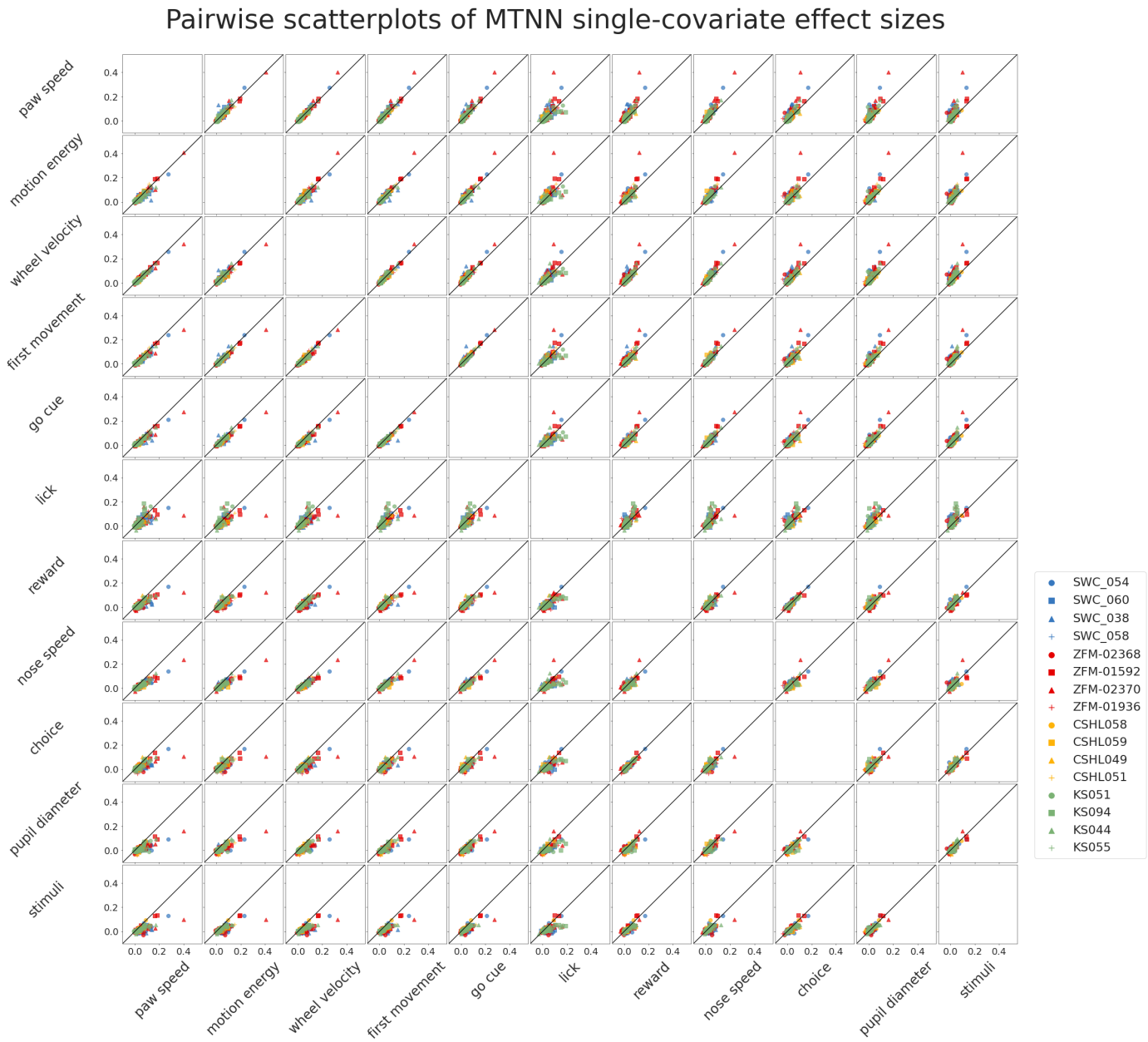

Figure 7–Figure supplement 3. We plot pairwise scatterplots of MTNN single-covariate effect sizes. Each dot represents the effect sizes of one neuron and is colored by lab. There is no outlier lab. The effect sizes are highly correlated.