eLife

Neuroscience

Reproducibility of *in vivo* electrophysiological measurements in mice

Reviewed Preprint
v1 • October 31, 2024
Not revised

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

University College London, London, UK • New York University, New York, USA • Columbia University, New York, USA • University of Washington, Seattle, USA • William James Center for Research, ISPA - Instituto Universitário, Lisbon, Portugal • Max-Planck-Institute, Tübingen, Germany • Sainsbury Wellcome Center, London, UK • University of Geneva, Geneva Switzerland • University of California, Los Angeles, Los Angeles USA • University of California, Berkeley, Berkeley, USA • Champalimaud Foundation, Lisbon, Portugal • Cold Spring Harbor Laboratory, Laurel Hollow, USA • Princeton University, Princeton, USA • Allen Institute for Neural Dynamics, Seattle, USA • Leiden University, The Netherlands

eLife Assessment

The authors have provided a **valuable** addition to the literature on large-scale electrophysiological experiments across many labs. The evidence that the authors provided was **incomplete** - while some comparisons with analyses outside of the manuscript's approaches were provided, a more complete manuscript would have compared with alternative standardized analyses. In particular, alternative spike sorting metrics and the alternative of GLM-based analysis of data would have made the interpretation of the results clearer.

https://doi.org/10.7554/eLife.100840.1.sa2

Abstract

Understanding brain function relies on the collective work of many labs generating reproducible results. However, reproducibility has not been systematically assessed within the context of electrophysiological recordings during cognitive behaviors. 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; this generated a total of 121 experimental replicates, a unique dataset for evaluating reproducibility of electrophysiology experiments. Despite

standardizing both behavioral and electrophysiological procedures, some experimental outcomes were highly variable. A closer analysis uncovered that variability in electrode targeting hindered reproducibility, as did the limited statistical power of some routinely used electrophysiological analyses, such as single-neuron tests of modulation by task parameters. Reproducibility was enhanced by histological and electrophysiological quality-control criteria. Our observations suggest that data from systems neuroscience is vulnerable to a lack of reproducibility, but that across-lab standardization, including metrics we propose, can serve to mitigate this.

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 (Crabbe et al., 1999; Sittig et al., 2016; 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 secondary visual areas, hippocampus, and thalamus) in mice performing an established decision-making task

(*The International Brain Laboratoryetal., 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 Guidelines for Optimizing 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 LFP power were reproducible across laboratories according to this definition. However, the proportions of cells modulated and the precise degree of modulation by decision-making variables (such as the sensory stimulus or the choice), while reproducible for many tests and brain regions, failed to reproduce in some regions. To interpret these potential lab-to-lab differences in reproducibility, we developed a multi-task neural network encoding model that allows nonlinear interactions between variables. We found that within-lab random effects captured by this model were comparable to between-lab random effects. Taken together, these results suggest that electrophysiology experiments are vulnerable to a lack of reproducibility, but that standardization of procedures and quality control metrics can help to mitigate this problem.

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 384 dual-band, low-noise recording channels providing the ability to sample many neurons in each of multiple brain regions simultaneously. 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 lateral posterior (LP) and posterior (PO) thalamic nuclei (Saalmann and Kastner, 2011; Roth et al., 2016).

Once data were acquired and brains were processed, we visualized probe trajectories using the Allen Institute Common Coordinate Frame (**Figure 1d**, more methodological information is below). This allowed us to associate each neuron with a particular region (**Figure 1e**). To evaluate whether our data were of comparable quality to previously published datasets (Steinmetz et al., 2019; Siegle et al., 2021), we compared the neuron yield. The yield of an insertion, defined as the number of quality control (QC)-passing units recovered per electrode site, is informative about the quality of both the raw recording and the spike sorting pipeline. We performed a comparative yield analysis of the repeated site insertions and the external electrophysiology datasets. Both external datasets were recorded from mice performing a visual task, and include insertions from diverse brain regions.

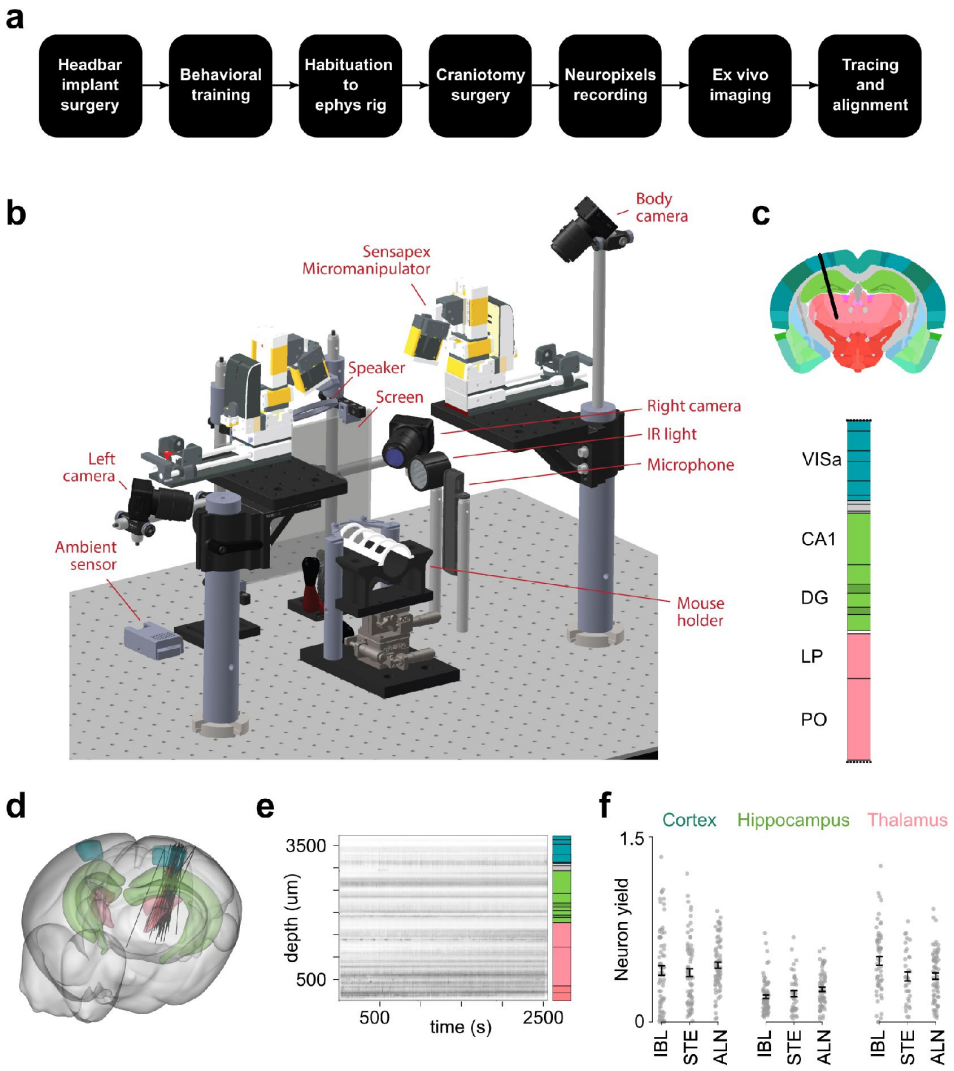


Figure 1.

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. Black lines: boundaries of sub regions within the five repeated site regions. **(d)**, Acquired repeated site trajectories shown within a 3D brain schematic. **(e)**, Raster plot of all measured neurons (including some that ultimately failed quality control) from one example session. **(f)**, A comparison of neuron yield (neurons/channel of the Neuropixels probe) for this dataset (IBL), *Steinmetz et al.* (2019) (STE) and *Siegle et al.* (2021) (ALN) in 3 neural structures. Bars: standard error of the mean; the center of each bar corresponds to the mean neuron yield for the corresponding study.

Different spike sorting algorithms detect distinct units as well as varying numbers of units, and yield is sensitive to the inclusion criteria for putative neurons. We therefore re-analyzed the two comparison datasets using IBL's spike sorting pipeline (*The International Brain Laboratory et al., 2022a*). When filtered using identical quality control metrics and split by brain region, we found the yield was comparable across datasets within each region (**Figure 1f**). A two-way ANOVA on the yield per insertion with dataset (IBL, Steinmetz, or Allen/Siegle) and region (Cortex, Thalamus, or Hippocampus) as categorical variables showed no significant effect of dataset origin (p=0.31), but a significant effect of brain region ($p < 10{-}17$). Systematic differences between our dataset and others were likewise absent for a number of other metrics (**Figure 1, supp. 4**).

Finally, in addition to the quantitative assessment of data quality, a qualitative assessment was performed. 100 total insertions were randomly selected from the three datasets and assigned random IDs to blind the raters during manual assessment. Three raters were asked to look at snippets of raw data traces with spikes overlaid and rate the overall quality of the recording and spike detection on a scale from 1-10, taking into account the apparent noise levels, drift, artifacts, and missed spikes. We found no evidence for systematic quality differences across the datasets (**Figure 1, supp. 5**).

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 probe trajectories (**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**; **Figure 2, 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 sizeable deviations from the planned insertion site (**Figure 2d** and **g**, total mean displacement = 104 µm).

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 = 356.0 µm). Assessing geometrical variability for all probes with permutation testing revealed a p-value of 0.19 across laboratories (**Figure 2h**), arguing 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 ($R2 < 0.03$), 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.

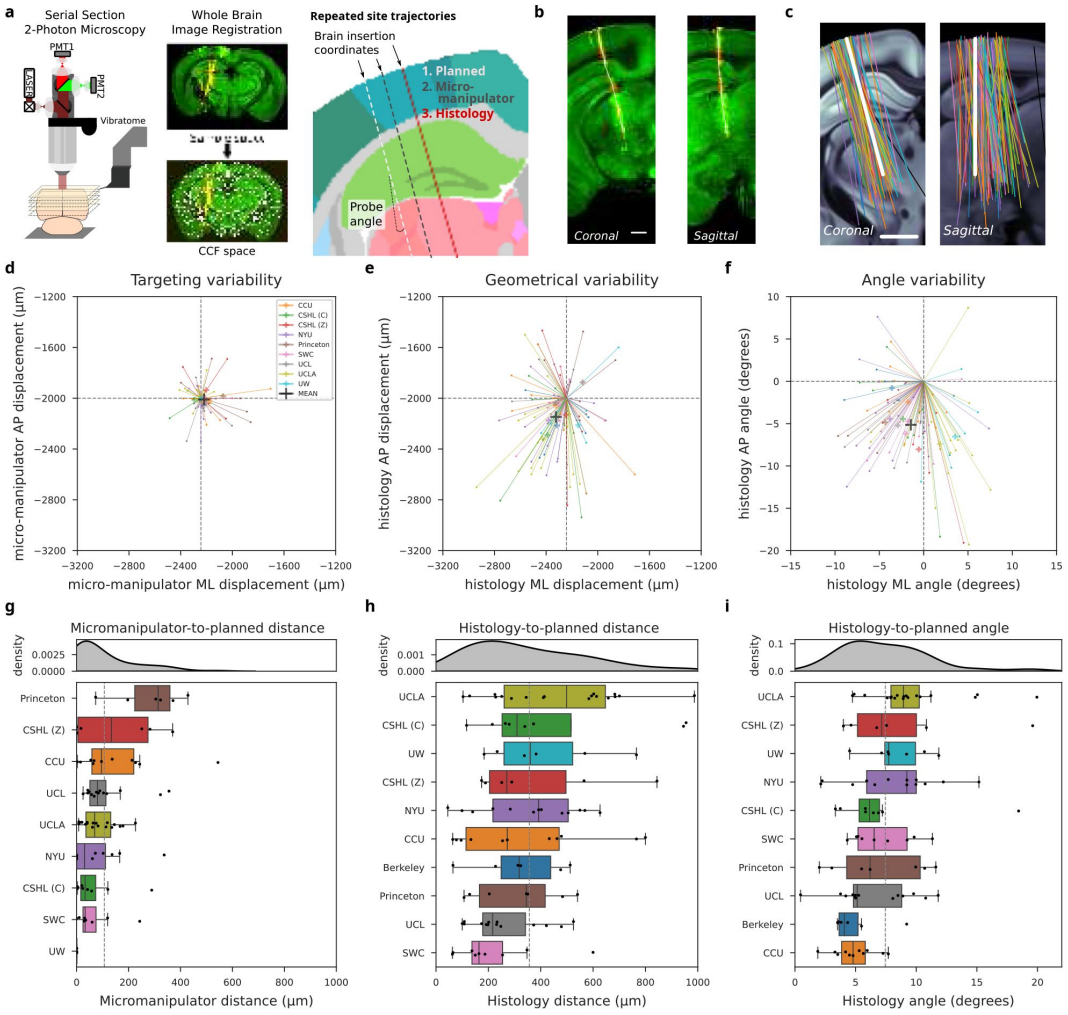


Figure 2.

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=88), histology brain surface insertion coordinate (e, geometrical variability, N=98), and histology probe angle (f, angle variability, N=99). 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, N=88), histology brain surface insertion coordinate (h, N=98), and histology probe angle (i, N=99). Kernel density estimate plots (top) are followed by boxplots (bottom) for probe displacement, ordered by descending median value. 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).

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: 7.5 degrees). Angle differences can be explained by the different orientations and geometries of the CCF and the stereotaxic coordinate systems, which we have recently resolved in a corrected atlas (Birman et al., 2023). The difference in histology angle to planned probe placement was assessed with permutation testing across labs, and shows a p-value of 0.45 (**Figure 2i**). This suggests that systematic lab-to-lab differences were minimal.

In conclusion, insertion variance, in particular geometrical variability, is sizeable enough to impact probe targeting to desired brain regions and poses a limit on the resolution of probe targeting. The histological reconstruction, fortunately, does provide an accurate reflection of the true probe trajectory (Liu et al., 2021), which is essential for interpretation of the Neuropixels recording data. We were unable to identify a prescriptive analysis to account for probe placement variance, 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 360 μm (based on the mean across labs, see **Figure 2h**), which
Reproducibility of electrophysiological features

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 twelve 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 of at least 3 sessions per lab for analyses that directly compared across labs (permutation tests; **Figure 3d**, **Figure 4e**, **Figure 5**). The remaining ten criteria, which we collectively refer to as the “Recording Inclusion Guidelines for Optimizing Reproducibility” (RIGOR; **Table 1**), are more general and can be applied widely to electrophysiology experiments: a yield criterion, a noise criterion, LFP power 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 spike amplitude). These metrics could serve as a benchmark for other studies to use when reporting results, acknowledging that a subset of the metrics will be adjusted for measurements made in different animals, regions or behavioral contexts.

We recorded a total of 121 sessions targeted at our planned repeated site (**Figure 3a**). Of these, 18 sessions were excluded due to incomplete data acquisition caused by a hardware failure during the experiment (10) or because we were unable to complete histology on the subject (8). 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, 10 recordings fail due to a high and unrecoverable count of noisy channels, 2 recordings fail due to artefacts, and 1 recording fail due to epileptiform activity. 1 recording failed our criterion for low yield, and 1 recording failed our criterion for noise level (both automated quality control criteria). Next, we applied criteria specific to our behavioral experiments, and found that 5 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 many 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 82 remaining sessions recorded in ten labs to determine the

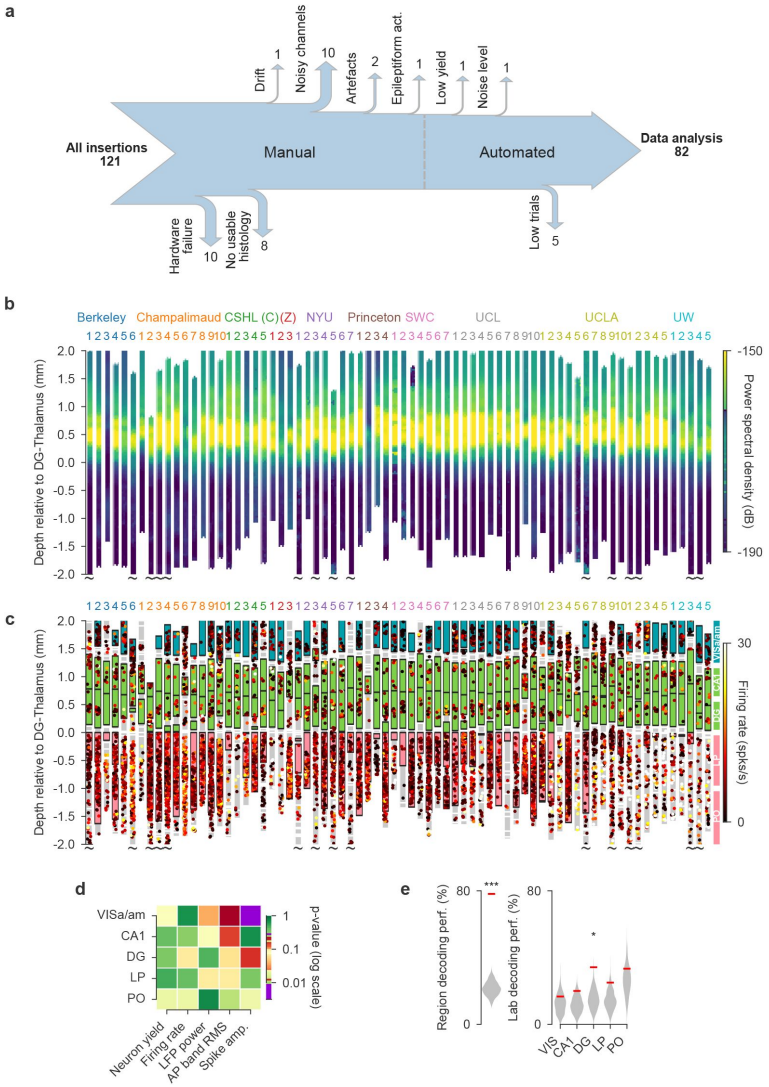


Figure 3.

Electrophysiological features are mostly reproducible across laboratories.

(a), Number of experimental sessions recorded; number of sessions used in analysis due to exclusion criteria. Up arrows: exclusions based on RIGOR criteria (**Table 1**); down arrows: exclusions based on IBL-specific criteria. For the rest of this figure an additional targeting criterion was used in which an insertion had to hit 2 of the target regions to be included. **(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 thalamus. Tildes indicate that the probe continued below -2.0mm. 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)**, 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. (* $p < 0.05$, *** $p < 0.001$)

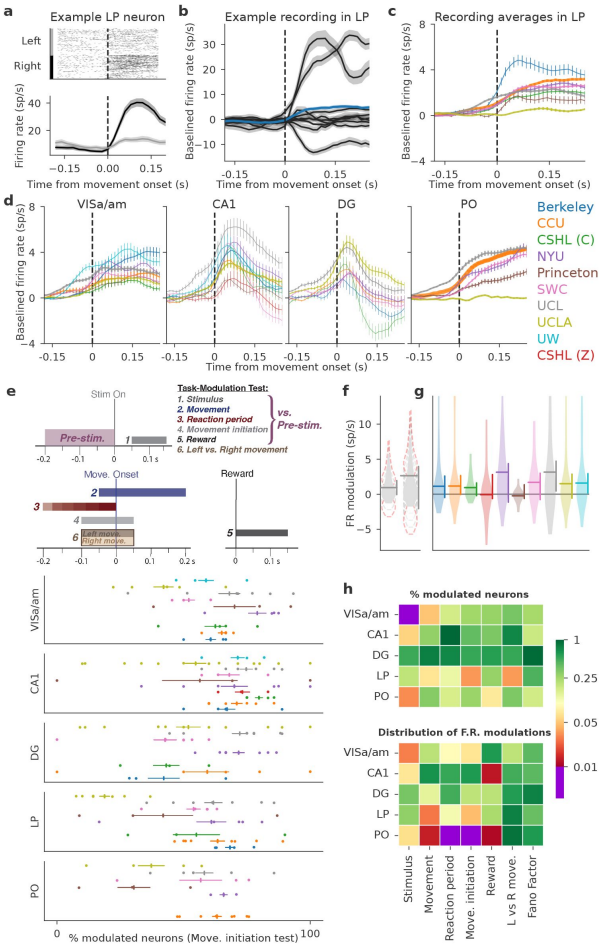


Figure 4.

Neural activity is modulated during decision-making in five neural structures and is variable between laboratories, particularly in the thalamus.

(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. 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 all neurons of each lab for LP (c) and the remaining four repeated site brain regions (d). Line thickness indicates the number of neurons in each lab (ranging from 18 to 554). **(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 row within a region correspond to a single lab (colors same as in (d), points are individual sessions). Horizontal lines around vertical marker: S.E.M. around mean across lab sessions **(f)**, Schematic to describe the power analysis. 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 lab; vertical line at right: how much the distribution can shift up- or downwards before the test becomes significant, while holding other labs constant.**(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 modulations. Comparisons performed for correct trials with non-zero contrast stimuli. (Figure analyses include collected data that pass our quality metrics and have at least 4 good units in the specified brain region and 3 recordings from the specified lab, to ensure that the data from a lab can be considered representative.)

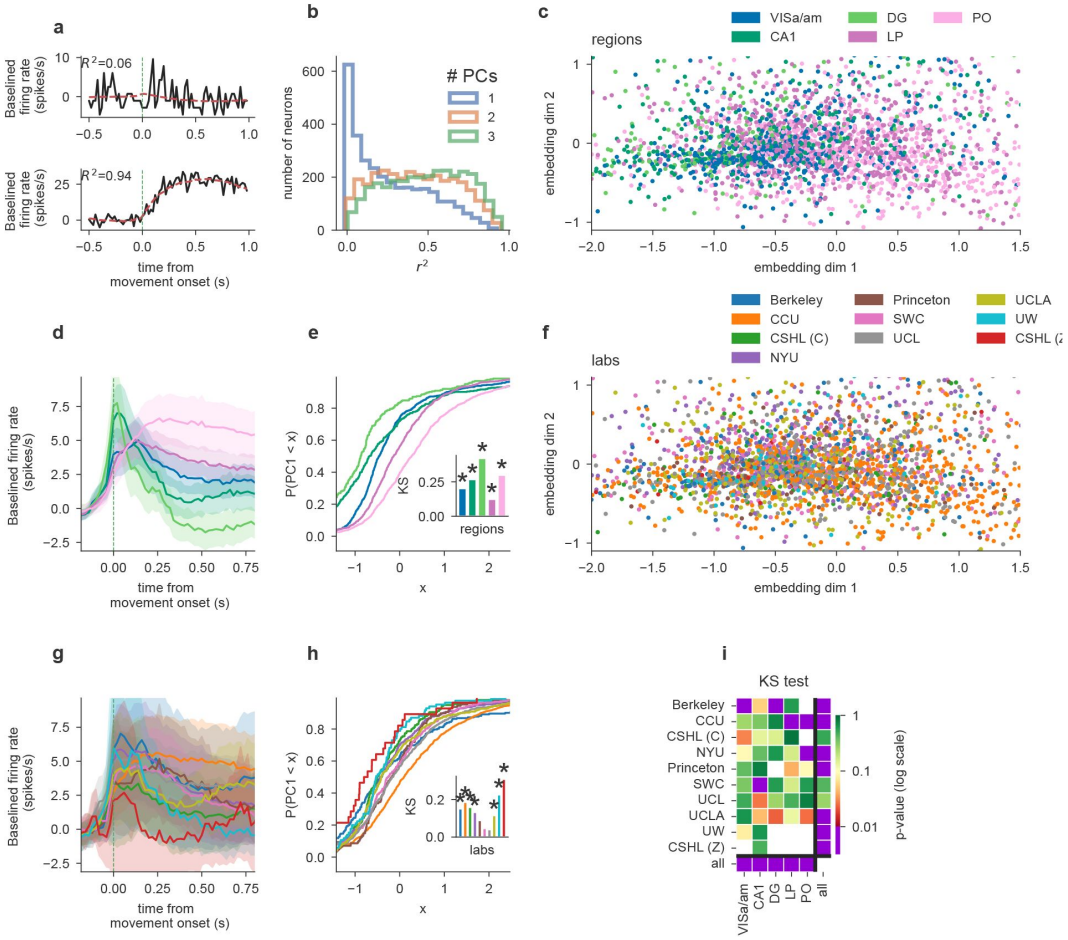


Figure 5.

Principal component embedding of trial-averaged activity uncovers differences that are clear region-to-region and more modest lab-to-lab.

(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 just 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 the scatter plot. 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 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 each lab, using cells from all regions (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)** p-values of all KS tests without sub-sampling; white squares indicate that there were too few cells for the corresponding region/lab pair. 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. Rightmost 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 are less so.

		Criterion	Definition
Whole recording	Computed	Yield	At least 0.1 neurons (that pass single unit criteria) per electrode channel in each brain region.
		Noise level	AP band: Median action-potential band RMS (AP RMS) less than 40 uV LF band: Median of the derivative of the LFP band (20 - 80Hz) less than 1 (may differ for other electrodes)
		LFP derivative	LFP power should smoothly vary over channels. If the derivative of the LFP power over all channels is > 1 this indicates the LFP power is noisy.
	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 channel 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; see Methods). 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 spike amplitude	Each neuron must have a median spike amplitude greater than 50 μV.

Table 1.

Recording Inclusion Guidelines for Optimizing Reproducibility (RIGOR).

reproducibility of our electrophysiological recordings. The responses of 5312 single neurons (all passing the metrics defined in (**Table 1**) are analyzed below; this total reflects an average of 105 ± 53 [mean ± std] per insertion.

We then evaluated whether electrophysiological features of these neurons, such as firing rates and LFP power, were re-producible 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; *Braginet al., 1995*; *Senzaiand 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. Several trajectories, most notably from UCLA and UW, missed their targets in deeper brain regions (LP, PO), as indicated by gray blocks. These trajectories are the ones that show a large displacement from the target insertion coordinates; note that insertions from UCLA and UW are consistently placed in one quadrant relative to the target (**Figure 2f**). These observations raise 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?

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.

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 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, neuronal yield was highly variable across animals, and less so within animals. By contrast, noise (assessed as action-potential band root mean square, or AP band RMS) was more variable within the same animal than across animals, presumably because probes have noise levels which are relatively stable over time.

To test whether recording-to-recording variability is lab dependent, we used a permutation testing approach (Methods). The tested features were neuronal yield, firing rate, spike amplitude, LFP power, and AP band RMS. These were calculated in each 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 tests, we were concerned that systematic corrections (such as a Bonferroni correction) might be too lenient and could lead us to overlook failures (or at least threats) to reproducibility. We therefore opted to use a more stringent a of 0.01 when establishing significance. The

permutation test revealed a significant effect for the spike amplitude in VISa/am. Otherwise, we found that all electrophysiological features were reproducible across laboratories for all regions studied (**Figure 3d**).

The permutation testing approach evaluated the reproducibility of each electrophysiological feature separately. It could 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 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 than the original labels. For validation, we first verified that the decoder could successfully identify brain region (instead of lab identity) from the electrophysiological features; the decoder was able to identify brain region with high accuracy (**Figure 3e**, left). When decoding lab identity, the classifier had above-chance performance only on the dentate gyrus and not from any of the other regions, indicating that most electrophysiological features were reproducible across laboratories for these regions (**Figure 3e**, right).

Reproducibility of functional activity depends on brain region and analysis method

Concerns about reproducibility extend not only to electrophysiological properties, but also functional properties. To address this, we analyzed the reproducibility of neural activity driven by perceptual decisions about the spatial location of a visual grating stimulus (The International Brain Laboratory et al., 2021). In total, we compared the activity of 4400 neurons across all labs and regions. We focused on whether the targeted brain regions have comparable neural responses to stimuli, movements and rewards. An inspection of individual neurons revealed clear modulation by, for instance, the onset of movement to report a particular choice (**Figure 4a**). Despite considerable neuron-to-neuron variability (**Figure 4b**), the temporal dynamics of session-averaged neural activity were similar across labs in some regions (see VISa/am, CA1 and DG in **Figure 4d**). Elsewhere, the session-averaged neural response was more variable (see LP and PO in **Figure 4c** and **d**).

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 neurons that were modulated by stimuli, movements or rewards (**Figure 4e**). We used six comparisons of task-related time windows (using Wilcoxon signed-rank tests and Wilcoxon rank-sum tests, *Steinmetz et al.* (*2019*)) to identify neurons with significantly modulated firing rates (for this, we use $\alpha = 0.05$) during the task (**Figure 4e**, top and **Figure 4-supplemental 1** specify which time windows we compared for which test on each trial, since leftwards versus rightwards choices are not paired we use Wilcoxon rank-sum tests to determine modulation for them). For most tests, the proportions of modulated neurons across sessions and across brain regions were quite variable (**Figure 4e**, bottom and **Figure 4-supplemental 1**).

After determining that our measurements afforded sufficient power to detect differences across labs (**Figure 4f,g**, **Figure 4-supplemental 2**, see Methods), we evaluated reproducibility of these proportions using permutation tests (**Figure 4h**). Our tests uncovered only one region and period for which there were significant differences across labs ($\alpha = 0.01$, no multiple comparisons correction, as described above): VISa/am during the stimulus onset period (**Figure 4h**, top). In addition to examining the proportion of responsive neurons (a quantity often used to determine that a particular area subserves a particular function), we also compared the full distribution of

firing rate modulations. Here, reproducibility was tenuous and failed in thalamic regions for some tests (**Figure 4h**, bottom). Taken together, these tests uncover that some traditional metrics of neuron modulation are vulnerable to a lack of reproducibility.

These failures in reproducibility were driven by outlier labs; for instance, for movement-driven activity in PO, UCLA reported an average change of 0 spikes/s, while CCU reported a large and consistent change (**Figure 4d**, right most panel, compare orange vs. yellow traces). Similarly, the differences in modulation in VISa/am were driven by one lab, in which all four mice had a higher percentage of modulated neurons than mice in other labs (**Figure 4-supplemental 1a**, purple dots at right).

Looking closely at the individual sessions and neuronal responses, we were unable to find a simple explanation for these outliers: they were not driven by differences in mouse behavior, wheel kinematics when reporting choices, or spatial position of neurons (examined in detail in the next section). This high variability across individual neurons meant that differences that are clear in across-lab aggregate data can be unclear within a single lab. For instance, a difference between the proportion of neurons modulated by movement initiation vs. left/right choices is clear in the aggregate data (compare panels d and f in **Figure 4-supplemental 1**), but much less clear in the data from any single lab (e.g., compare values for individual labs in panels d vs. f of **Figure 4-supplemental 1**). This is an example of a finding that might fail to reproduce across individual labs, and underscores that single-neuron analyses are susceptible to lack of reproducibility, in part due to the high variability of single neuron responses. Very large sample sizes of neurons could serve to mitigate this problem.

Having discovered vulnerabilities in the reproducibility of standard single-neuron tests, we tested reproducibility using an alternative approach: comparing low-dimensional summaries of activity across labs and regions. To start, we 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 (0.15 s was the mean reaction time when considering the distribution of reaction times across all sessions). We concatenated the resulting vectors to obtain a more thorough summary of each cell's average activity during decision-making. 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 exposes differences in brain regions (**Figure 5c**; e.g., PO and DG are segregated in the embedding space), verifying that the approach is sufficiently powerful to detect expected 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, differences are less obvious when coloring the same embedded activity by labs (**Figure 5f, g, h**), however some labs, such as CCU and CSHL(Z) are somewhat segregated. In comparison to regions, the activity point clouds overlap somewhat more homogeneously across most labs (**Figure 5-supplemental 1** for scatter plots, PETHs, cumulative distributions for each region separately, colored by lab).

We quantified this observation via two tests. First, we applied a permutation test using the first 2 PCs of all cells, computing each region's 2D 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. Second, 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 2D distance permutation test results, hence we report only the KS test results below.

This analysis confirmed that the neural activity in each region differs significantly (as defined above, $\alpha = 0.01$, no multiple comparison correction) from the neural activity measured in the other regions (**Figure 5i**, bottom row), demonstrating that neural dynamics differ from region-to-region, as expected and as suggested by **Figure 5c-e**. We also found that the neural activity from 7/10 labs differed significantly from the neural activity of all remaining labs when considering all neurons (**Figure 5i**, right column). When restricting this test to single regions, significant differences were observed for 6/40 lab-region pairs (**Figure 5i**, purple squares in the left columns). Note that in panels i of **Figure 5**, the row "all" indicates that all cells were used when testing if a region had a different distribution of first PCs than the remaining regions and analogously the column "all" for lab-targeted tests. Overall, these observations are in keeping with **Figure 4** and suggest, as above, that outlier responses during decision-making can be present despite careful standardization (**Figure 5i**).

The spatial position and spike waveforms of neurons explain minimal variability

In addition to lab-to-lab variability, which was low for electrophysiological features and higher for functional modulation during decision-making (**Figures 4**, **5**), we observed considerable variability between recording sessions and mice. Here, we examined whether this variability could be explained by either the within-region location of the Neuropixels probe or the single-unit spike waveform characteristics of the sampled neurons.

To investigate variability in session-averaged firing rates, we identified neurons that had firing rates different from the majority of neurons within each region (absolute deviation from the median firing rate being >15% of the firing rate range). These outlier neurons, which all turned out to be high-firing, 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 DG, the highest firing neurons were not entirely uniformly distributed in space. For instance, in LP, high firing neurons tended to be positioned more laterally and centered on the anterior-posterior axis (**Figure 6b**). In VISa/am and CA1, only the spatial position of neurons, but not differences in spike characteristics, contributed to differences in session-averaged firing rates (**Figure 6-supplemental 1b** and **2b**). In contrast, high-firing neurons in LP, PO, and DG had different spike characteristics compared to other neurons in their respective regions (**Figures 6b** and **6-supplemental 3a,c**). It does not appear that high-firing neurons in any brain region belong clearly to a specific neuronal subtype (see **Figure 6-supplemental 5**).

To quantify 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, features that had significant weights were mostly consistent with the results reported above: In VISa/am and CA1, spatial position explained part of the variance; in DG, the variance could not be explained by either spatial position or spike characteristics; in LP and PO, both spatial position and

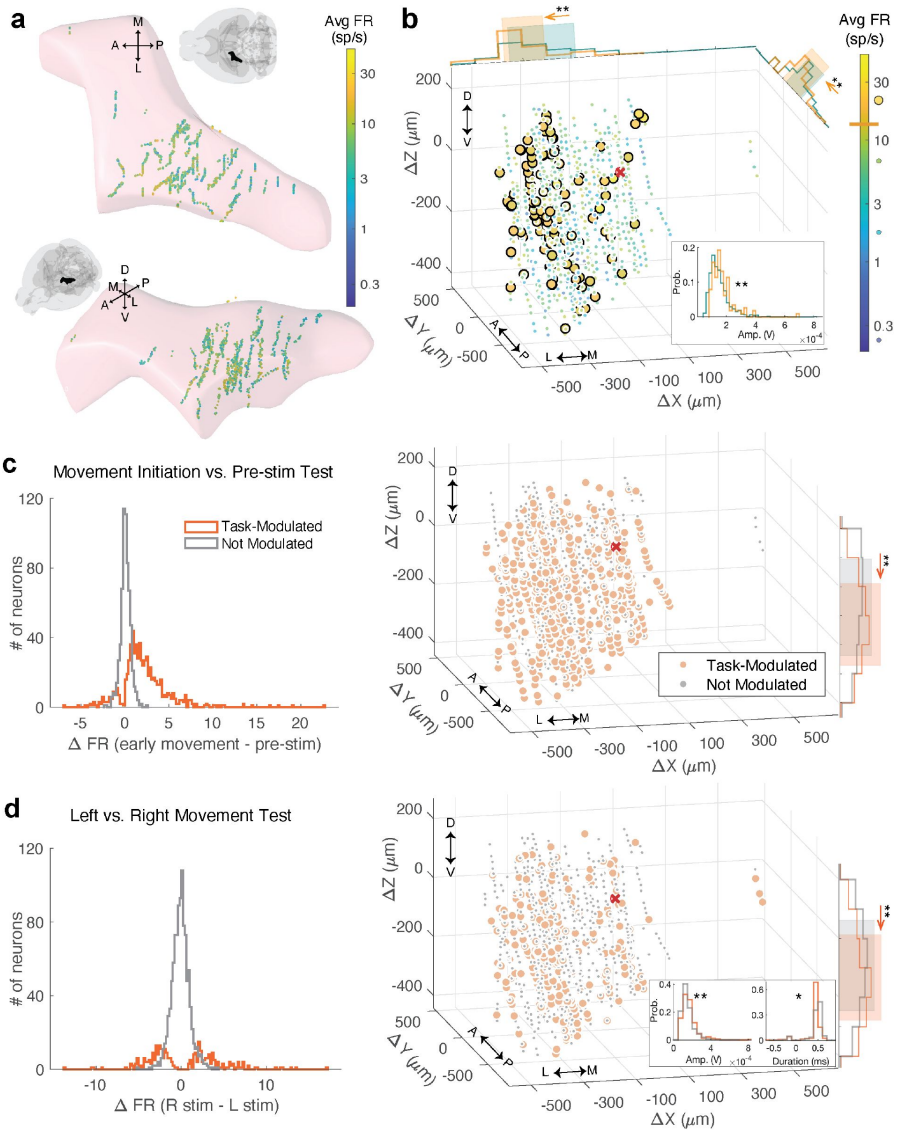


Figure 6.

High-firing and task-modulated LP neurons have slightly different spatial positions and spike waveform features than other LP neurons, possibly contributing only marginally to variability between sessions.

(a) Spatial positions of recorded neurons in LP. Colors: session-averaged firing rates on a log scale. To enable visualization of overlapping data points, small jitter was added to the unit locations. **(b)** Spatial positions of LP neurons plotted as distance from the planned target center of mass, indicated with red x. Colors: session-averaged firing rates on a log scale. Larger circles: outlier neurons (above the firing rate threshold of 13.5 sp/s shown on the colorbar). In LP, 114 out of 1337 neurons were outliers. 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 movement initiation (**Figure 4e**, **Figure 4-supplemental 1d**) for task-modulated (orange) and non-modulated (gray) neurons. 697 out of 1337 LP neurons were modulated during movement initiation. (Right) Spatial positions of task-modulated and non-modulated LP neurons, with histograms of significant features (here, z position) 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; histogram is bimodal because both left- and right-preferring neurons are included.292 out of 1337 LP neurons were modulated differently for leftward vs. rightward movements. (Figure analyses include all collected data that pass our quality metrics, regardless of the number of recordings per lab or number of repeated site brain areas that the probes pass through.)

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 $R2$ values), these five features still only accounted for a total of,. 5% of the firing rate variability. In VISa/am, CA1, and DG, they accounted for approximately 2%, 4%, and 2% 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 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 time-period comparisons in **Figure 4e** and **Figure 4-supplemental 1**). For instance, LP neurons modulated according to the movement initiation test and the left versus right movement test tended to be more ventral (**Figure 6c-d**). Other brain regions had weaker spatial differences than LP (**Figure 6-supplemental 1**, **2**, **3**). Spike characteristics were significantly different between task-modulated and non-modulated neurons only for one modulation test in LP (**Figure 6d**) and two tests in DG, but not in other brain regions. 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 (**Figure 6-supplemental 4**). Spike characteristics of neurons with lower vs. higher Fano Factors were only slightly different in VISa/am, CA1, and PO. 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 ($R2$., 5%) about the embedded PETHs of cells.

In summary, our results suggest that neither spatial position within a region nor waveform characteristics account for very much variability in neural activity. We examine other sources of variability in the next section.

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. 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 identity 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; Steinmetz et al., 2019) can be seen as a special case of the multi-task neural network, with a single hidden layer and linear weights in each layer.

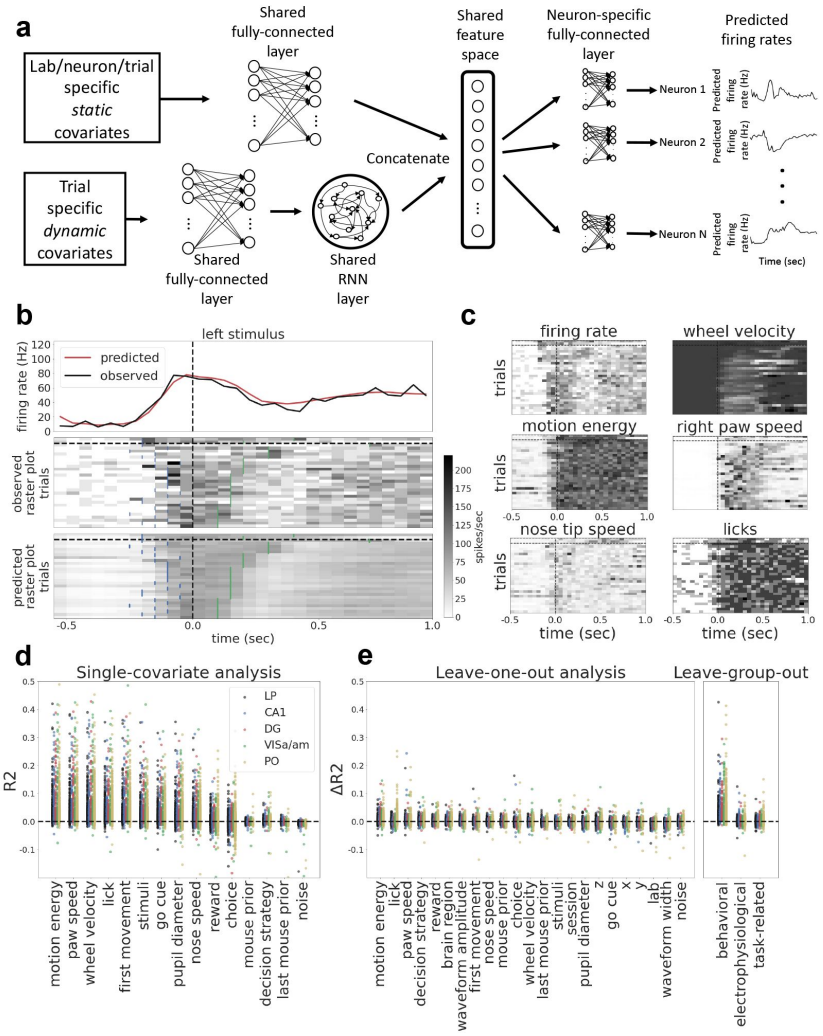


Figure 7.

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 $R2$ is 0.45 on held-out test trials and 0.94 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. **(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 1133 responsive neurons across 32 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.

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 Lightning Pose
Nose speed	Real / Dynamic	Movement	Inferred from DLC
Pupil diameter	Real / Dynamic	Movement	Inferred from Lightning Pose
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.

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 for left trials. 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 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 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**, with a simulated positive control analysis in **Figure 7-supplemental 3**. According to the single-covariate analysis (**Figure 7d**), face motion energy (derived from behavioral video), paw speed, wheel velocity, and first movement onset timing can individually explain about 5% of variance of the neurons on average, and these single-covariate effect sizes are highly correlated (**Figure 7-supplemental 4**). 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 10 geographically separated laboratories. We observed notable variability in the position of the electrodes in the brain despite efforts to target the same location (**Figure 2**). Fortunately, after applying stringent quality-control criteria (including the RIGOR standards, **Table 1**), we found that electrophysiological features such as neuronal yield, firing rate, and LFP power were reproducible across laboratories (**Figure 3d**). The proportion of neurons with responses that were modulated, and the ways in which they were modulated by behaviorally-relevant task events was more variable: lab-to-lab differences were evident in the thalamic region PO and in the cortical region VISa/am (**Figures 4**, **5**) and these were not explainable by, for instance, systematic variation in the subregions that were targeted (**Figure 6**). When we trained a multi-task neural network to predict neural activity, lab identity accounted for little neural variance (**Figure 7**), arguing that the lab-to-lab differences we observed are driven by outlier neurons/sessions and nonlinear interactions between variables, rather than, for instance, systematic biases. This is reassuring, but points to the need for large sample sizes of neurons to

overcome the inherent variability of single neuron recording. Altogether, our results suggest that standardization and quality control standards can enhance reproducibility, and that caution should be taken with regard to electrode placement and the interpretation of some standard single-neuron metrics.

The absence of systematic differences across labs for some metrics argues that standardization of procedures is a helpful step in generating reproducible results. Our experimental design precludes an analysis of 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 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 (retaining 82 / 121 experimental sessions). Quality control was enforced for diverse aspects of the experiments, including histology, behavior, targeting, neuronal yield, and the total number of completed sessions. A number of recordings with high noise and/or low neuronal yield were excluded. Exclusions were driven by 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). The autophagy community has a similar set of guidelines for experiments (Klionsky, 2016). Our work here suggests the creation of a similar set of standards for electrophysiology and behavioral experiments would be beneficial, at least on an “opt in” basis: for instance, manuscript authors, especially those new to electrophysiology, could state which of the metrics were adhered to, and which were not needed. The metrics might also serve as the starting point for training students how to impose inclusion criteria for electrophysiological studies. Importantly, the use of RIGOR need not prevent flexibility: studies focusing on multi-unit activity would naturally omit single unit metrics, and studies on other animals or contexts would understandably need to adjust some metrics.

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 some electrophysiological measures such as LFP power (**Fig 3d**) or functional responses in some regions but found considerable variability across experiments in other regions (**Figure 4h**, **Figure 5i**).

We found probe targeting to be a large source of variability. One possibility is that some aspect of our approach was vulnerable to targeting errors. Another possibility is that our approach uncovered a weakness in the standard targeting methods. Our ability to detect targeting error benefited from 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 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. We envision three solutions to this problem: first, probe angles must be carefully computed from the CCF, as the CCF and stereotaxic coordinate systems do not define the same coronal plane angle. Second, collecting structural MRI data in advance of implantation (Browning et al., 2023) could reduce targeting error, although this is infeasible for most labs. A third solution is to rely on stereotaxic coordinates and account for the inevitable off-target measurements by increasing cohort sizes. Experimenters are often reluctant to exclude subjects, but our analysis demonstrates that targeting error can lead to missed targets with non-negligible frequency. Indeed, 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. (Harvey et al., 2012; Raposo et al., 2014; Erlich et al., 2015; Goard et al., 2016; Najafi et al., 2020). 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 (see Data Availability) and suite of analysis tools for quantifying reproducibility (see Code Availability) 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 et al.* (*2022b*,a) for full details). 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 experimental findings. The current study demonstrates that reproducibility is attainable for many features measured during a standardized perceptual decision task. We offer several recommendations to enhance reproducibility, including (1) standardized protocols for data collection and processing, (2) methods to account for variability

in electrode targeting and (3) rigorous data quality metrics, such as RIGOR. These recommendations are urgently needed as neuroscience continues to move towards increasingly large and complex datasets.

Resources

Data availability

All data are freely available. 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 availability

All code is freely available. 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 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=78 adult mice (C57BL/6, male and female, obtained from either Jackson Laboratory or Charles River) were used in this study. Mice were aged 111 - 442 days and weighed 16.4-34.5 g on the day of their first electrophysiology recording session in the repeated site.

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 (640×512), 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

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 planned 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) must have bias shifts < 5%, and lapse rates measured on asymmetric blocks must be 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. For the analyses in this study, only electrophysiology sessions where the mouse completed at least 400 trials were used.

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 a scalpel blade. The 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 did not dry during the recording. Subsequently, a grounding pin was cemented to the skull using Metabond. 1-2 craniotomies (1 × 1 mm) were made over the marked locations using 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 relocated to 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. CM-Dil was strored in the freezer -at 20C until ready for use. On the day of recording, we thawed 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 PyKilosort 1.4.7, a Python port of the Kilosort 2.5 algorithm that includes modifications to preprocessing (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, 2021*b*). See *The International Brain Laboratory et al.* (*2022a*) for full details.

Single cluster quality metrics

To determine whether a single cluster will be used in downstream analysis, we used three metrics (listed as part of RIGOR in **Table 1**): 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 was applied to rephase each channel according to the time-sampling difference due to sequential sampling of the hardware. Then faulty bad channels were automatically identified, removed and interpolated. Finally, the median reference is subtracted at each time sample. See *The International Brain Laboratory et al.* (*2022a*) 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. The toe-pinch test was performed as confirmation that the mouse was deeply anesthetized before proceeding with the surgical procedure. Thoracic cavity was opened, the atrium was cut, and PBS followed by 4% formaldehyde solution (Thermofisher 28908) in 0.1M PB pH 7.4 were perfused through the left ventricle. The 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 ±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 independent registrations were performed, samples were registered to the CCF template image and the CCF template was registered to the sample.

Probe track tracing and alignment

Tracing of Neuropixels electrode tracks was 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. Neu-ropixels 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 (**Figure 4, Supp. Fig 2**), the goal was to find how much each value (firing rate modulations or Fano factors) would need to shift within the individual labs 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 a 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, ensures that our permutation test is sensitive enough to detect deviations of a given 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 **Figure 4, Supp. 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

The analyses in **Figure 5** rely on principal component analysis (PCA) to embed PETHs into a two-dimensional feature space. 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 a summary of each cell's activity. Next we stack these double PETHs from all labs into a single matrix. Finally, we used 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, 0.15sec being the mean reaction time when considering the distribution of reaction times across all sessions) 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, subtracting baseline activity and then 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.* (*2019*).

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 $UV \cong Y$. 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 5a** shows the responses of two cells of Y (black traces) and the corresponding PCA approximation from UV (red traces).

The scatter plots in **Figure 5c,f** 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,f**.

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 ($R2$) 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 (1280×1024) and 60 Hz filming the mouse from one side, one called 'right' at half resolution (640×512) 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, 2021*a*).

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, Lightning Pose, for these body parts (Biderman et al., 2023). Lightning Pose models were trained on the same final labeled dataset used to train DeepLabCut. We then applied the Ensemble Kalman Smoother (EKS) post-processor introduced in *Biderman et al.* (*2023*), which incorporates information across multiple networks trained on different train/validation splits of the data. For the paw data, we used a version of EKS that further incorporates information across the left and right camera views in order to resolve occlusions. See *Biderman et al.* (*2023*) for full details.

Multi-task neural network model to quantify sources of variability

Data preprocessing

For the multi-task neural network (MTNN) analysis, we used data from 32 sessions recorded in SWC, CCU, CSHL (C), UCL, Berkeley, NYU, UCLA, and UW. 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 selected sessions with more recorded neurons across all repeated sites. We included various covariates in our feature set (e.g. go-cue signals, stimulus/reward type, DeepLabCut/Lightning Pose 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.

The lab IDs and session IDs were each encoded in a "one-hot" format (i.e., each lab is encoded as a length 8 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 $x_{dynamic} \in$ that include dynamic covariates, such as Deep Lab Cut (DLC) outputs, and wheel velocity, and a feature vector $x$static $\in \mathbb{R}D$static that includes static covariates, such as the lab ID, neuron's 3-D location, are input to the MTNN to compute the prediction $ypred \in \mathbb{R}T$, where Dstatic is the number of static features, Ddynamic 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 *ypred* by:

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

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

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

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

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

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 four-layer bidirectional RNNs for dynamic covariates, with the embedding size set to 128.

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 1024.

The dataset consists of 32 sessions, 1133 neurons and 11490 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 1133 responsive neurons from 32 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 500 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.

Acknowledgements

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 Collaboration on the Global Brain. We thank R. Poldrack, T. Zador, P. Dayan, S. Hofer, N. Ghani, 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, AKC, MFD, CL, HDL, MF, GM, LP, NR, MS, NS, MT, and SJW.

Additional information

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.

Conflict of interest

Anne Churchland receives an honorarium for her role on the executive committee for the Simons Collaboration on the Global brain, one of the funders of this work.

Supplementary figures

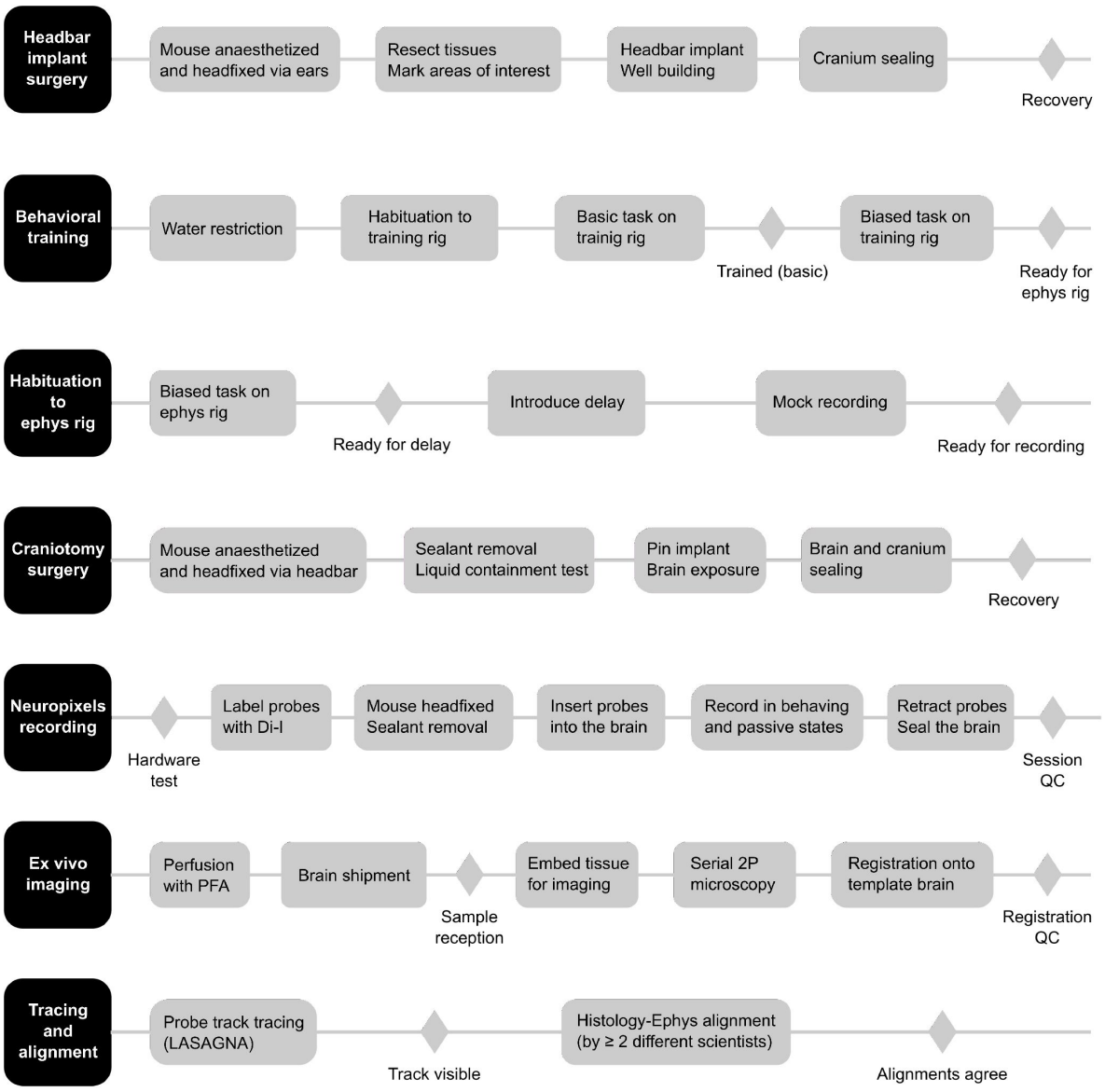


Figure 1–Figure supplement 1.

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.

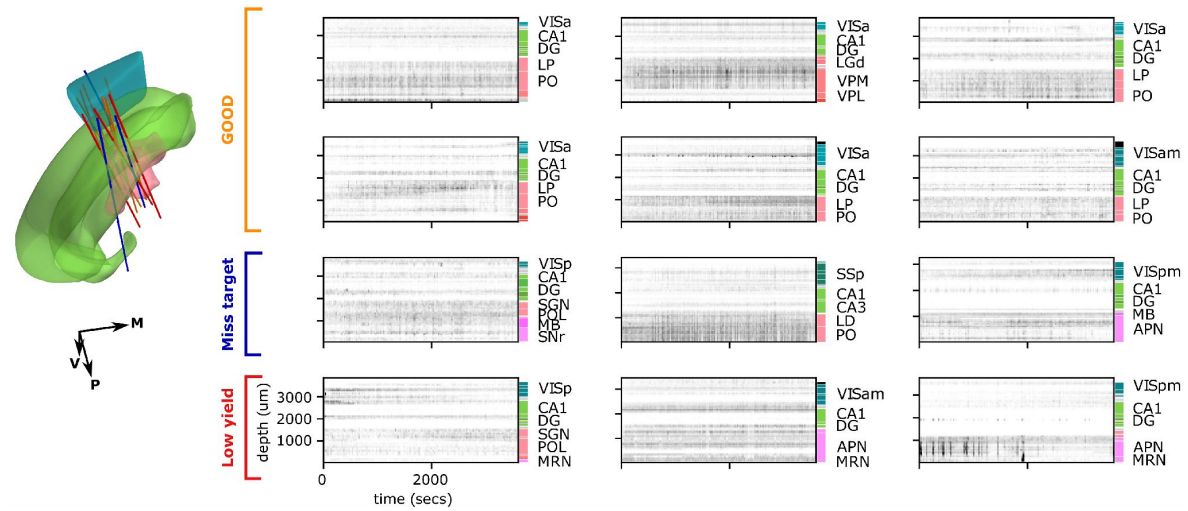


Figure 1–Figure supplement 2.

Spiking activity qualitatively appears heterogeneous across recordings.

(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. All measured neurons are shown, including some that ultimately failed quality control. 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).

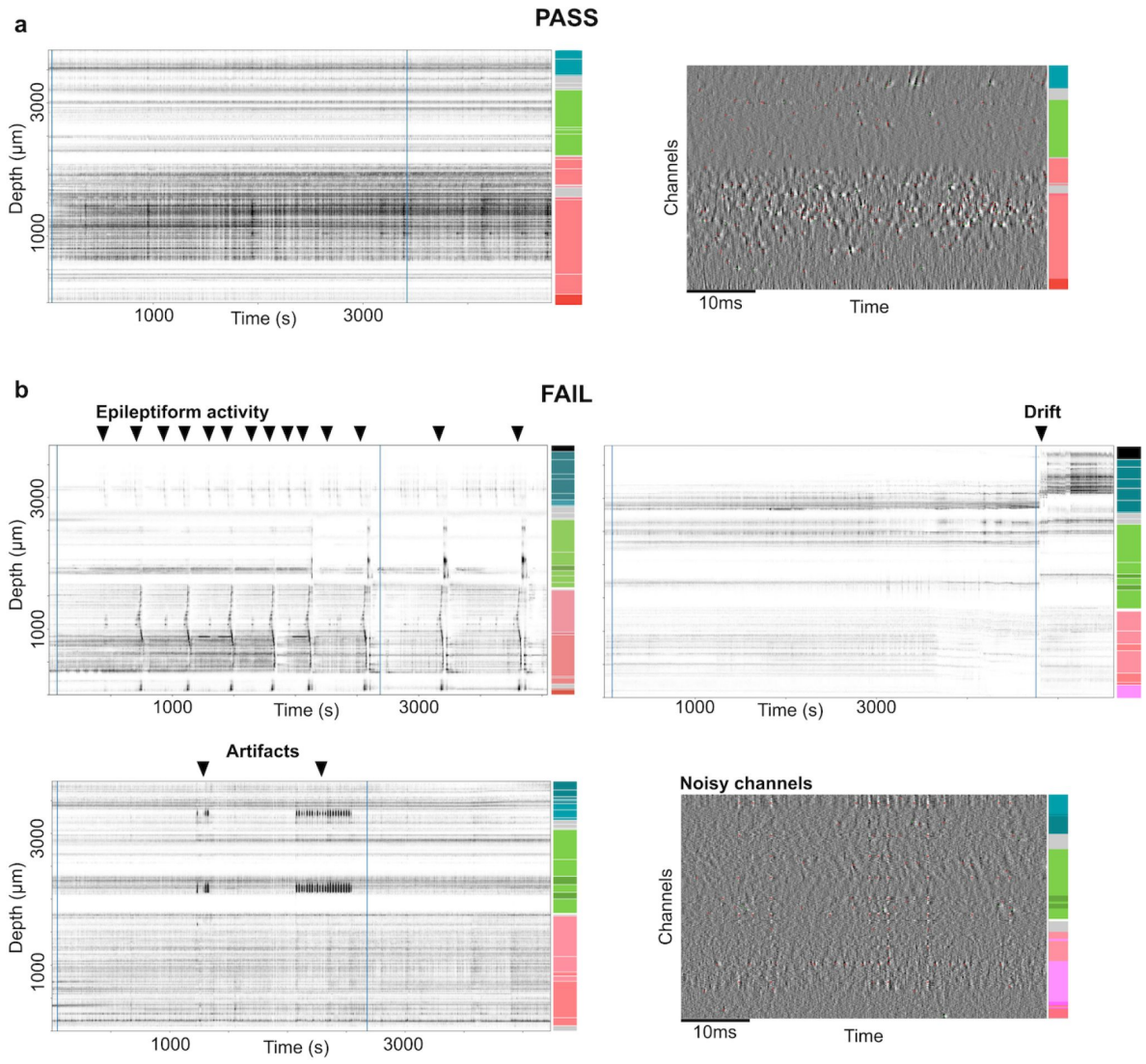


Figure 1–Figure supplement 3.

Electrophysiology data quality examples.

(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. Red dots on the raw data snippets indicate spikes from multi-unit activity; green dots indicate spikes from "good" units. **(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).

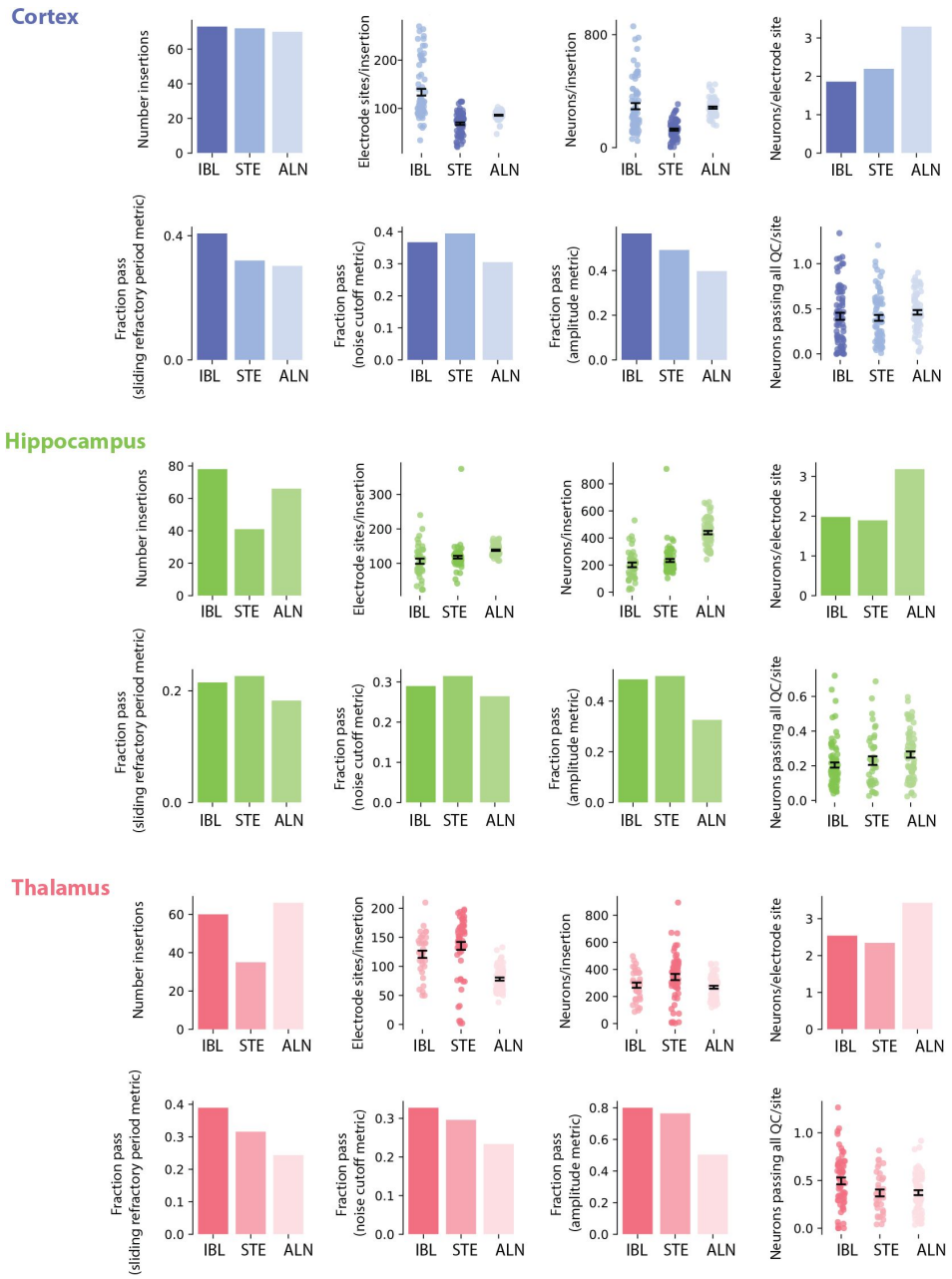


Figure 1–Figure supplement 4.

A comparison of our metrics with other studies.

Comparison of multiple metrics (columns) for the current dataset (IBL, left), Steinmetz et al. 2019 (STE, middle), and Siegle et al. 2021 (ALN, right) for measurements within cortex (Top; blue graphs), hippocampus (Middle; green graphs), and thalamus (Bottom; pink graphs).

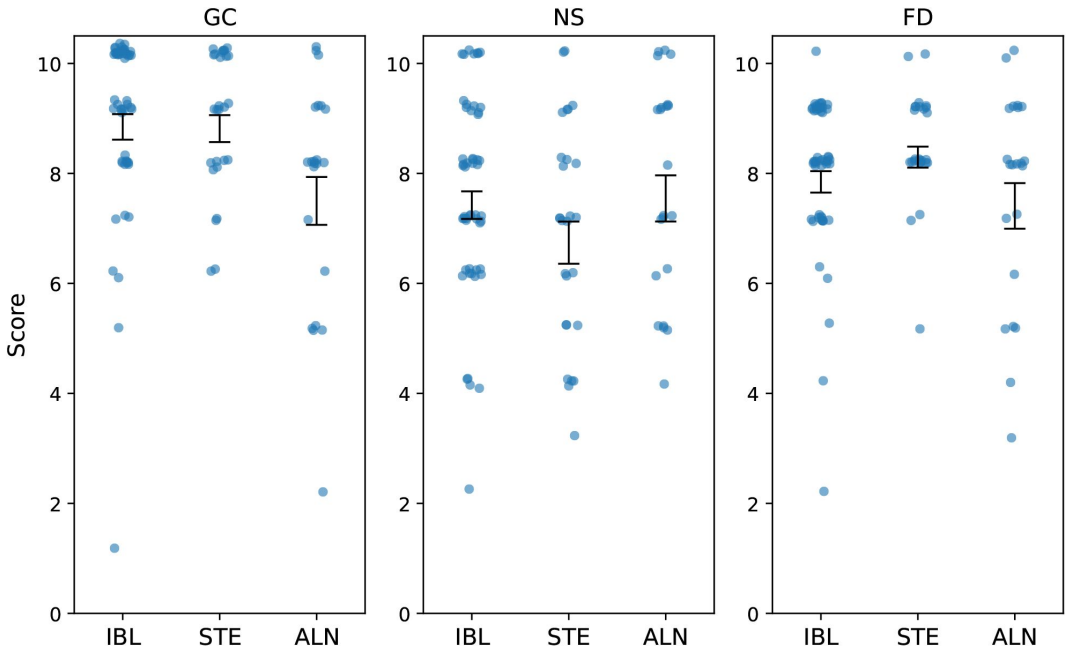


Figure 1–Figure supplement 5.

Visual inspection of datasets by 3 observers blinded to data identity yielded similar metrics for all 3 studies.

Each plot reflects manually-determined scores from a single observer (letters at top indicate their initials), who was blinded to the origin of the dataset at the time of scoring. Labels on the horizontal axis indicate the dataset: the current dataset (IBL, left), Steinmetz et al., 2019 (STE, middle), and Siegle et al., 2021 (ALN, right). Each point is the score assigned to a single snippet. Error bars represent standard error of the mean. A two way ANOVA was performed with rater identity and source dataset as categorical variables. We found no significant effect of rater ID ($p < 10–6$), a small effect of dataset ($p = 0.08$), and no interaction between rater ID and dataset ($p < 0.05$)

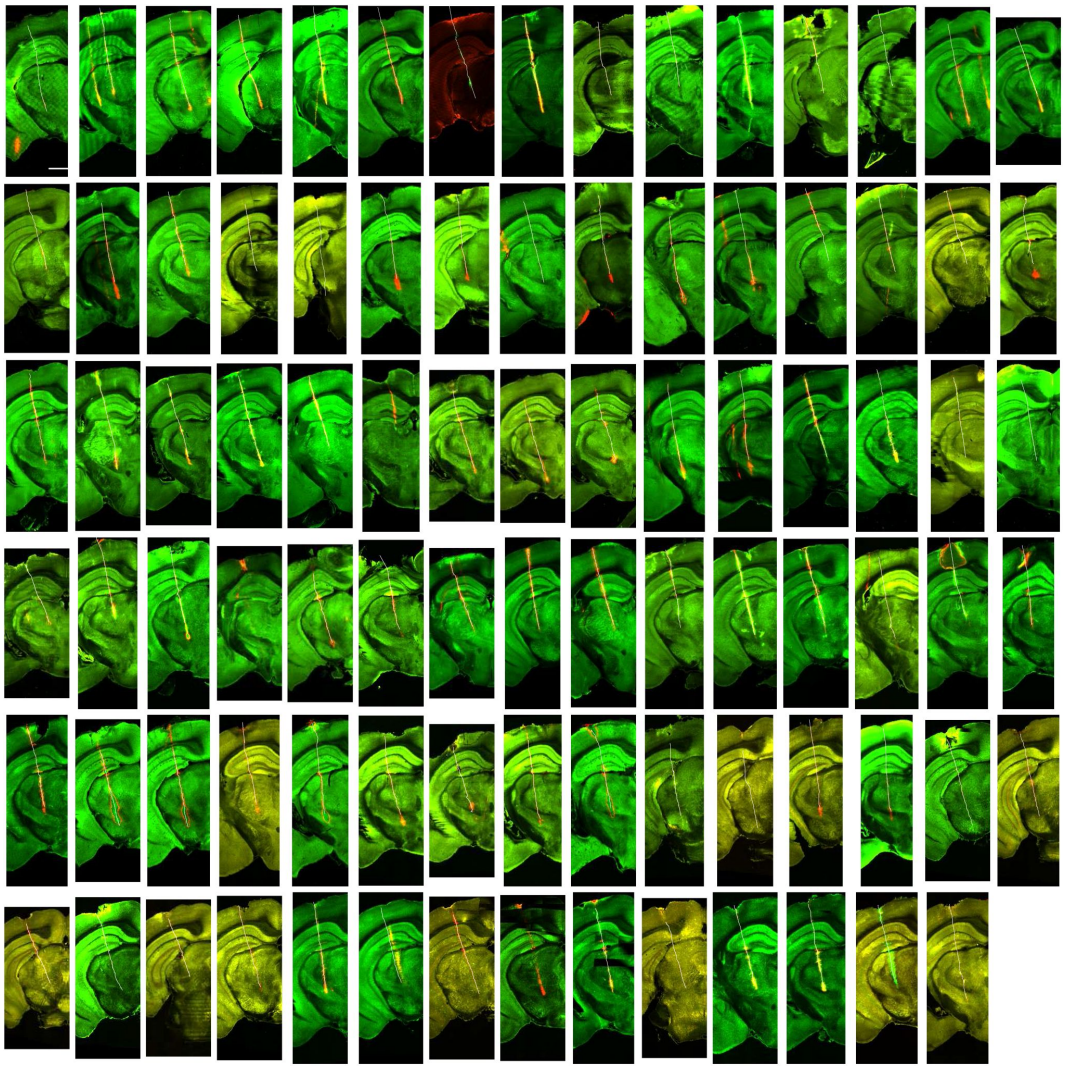


Figure 2–Figure supplement 1.

Tilted slices along the histology insertion for all insertions used in assessing probe placement.

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.

Recordings that did not pass QC can be visually accessed as outliers. (a, b)

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

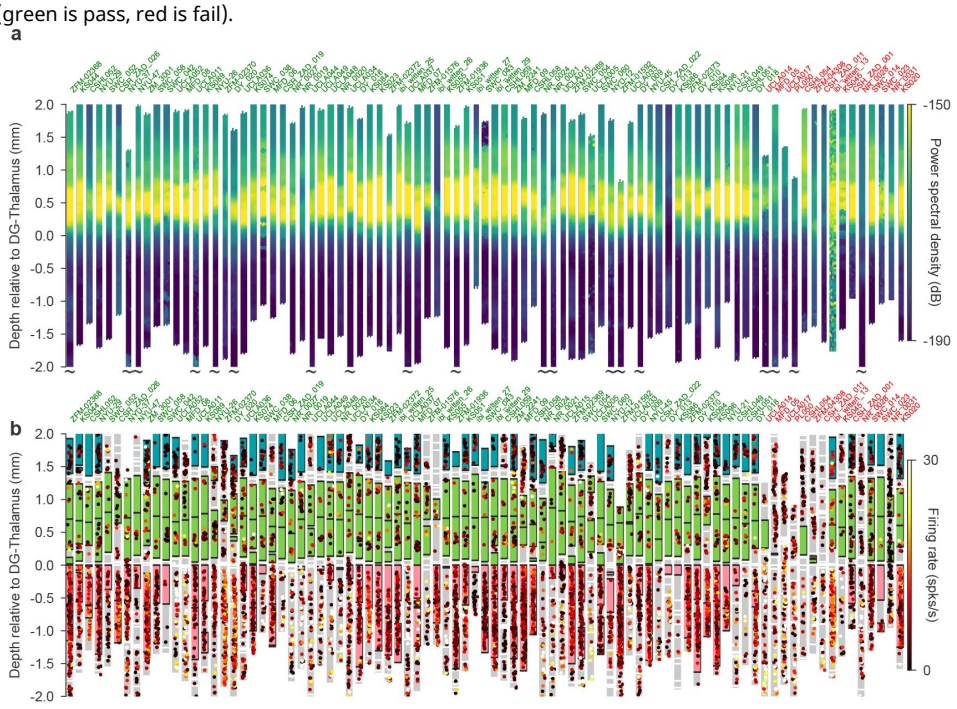
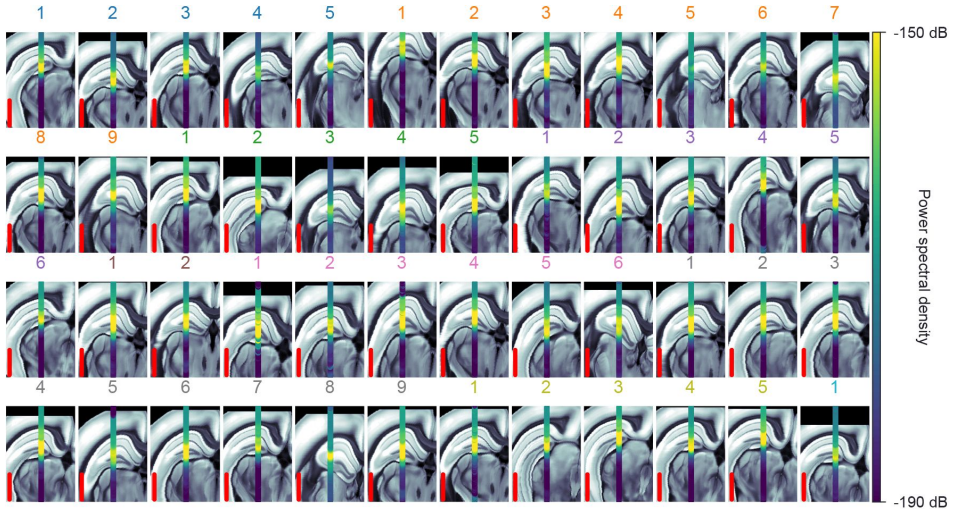


Figure 3–Figure supplement 2.

High LFP power in dentate gyrus was used to align probe locations in the brain.

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 denotes 1 mm scalebar.



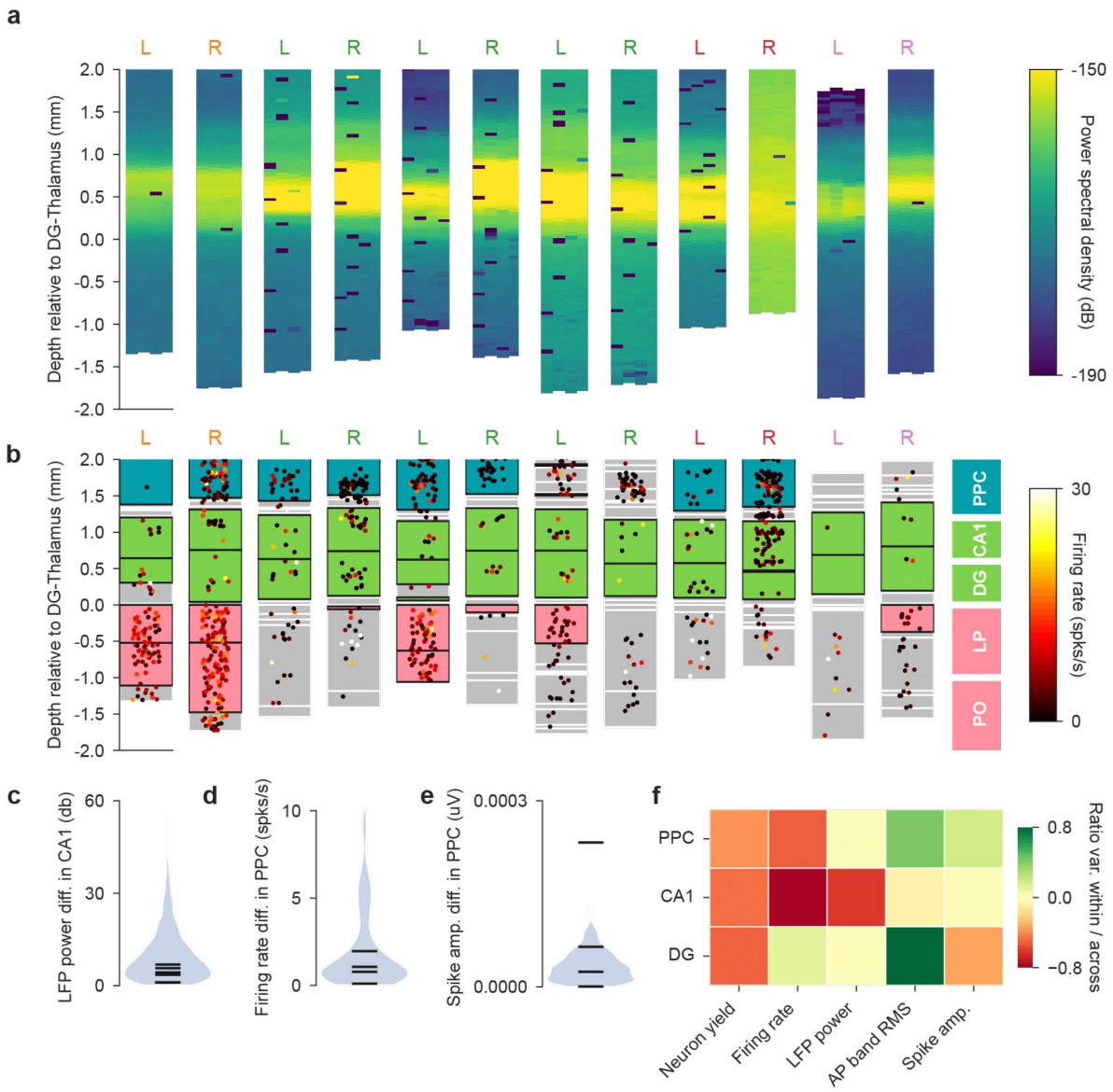


Figure 3–Figure supplement 3.

Bilateral recordings assess within-vs across-animal variance.

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.

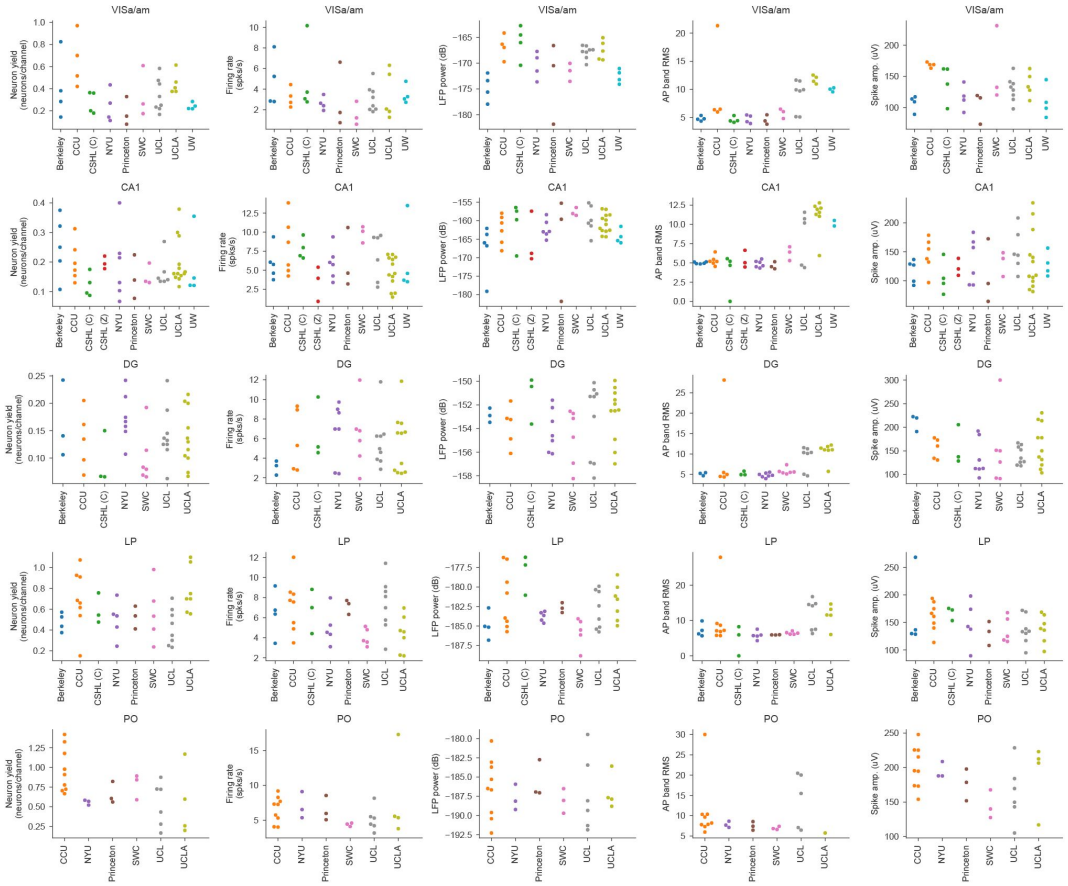


Figure 3–Figure supplement 4.

Values used in the decoding analysis, per metric and per brain region.

All electrophysio-logical 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.

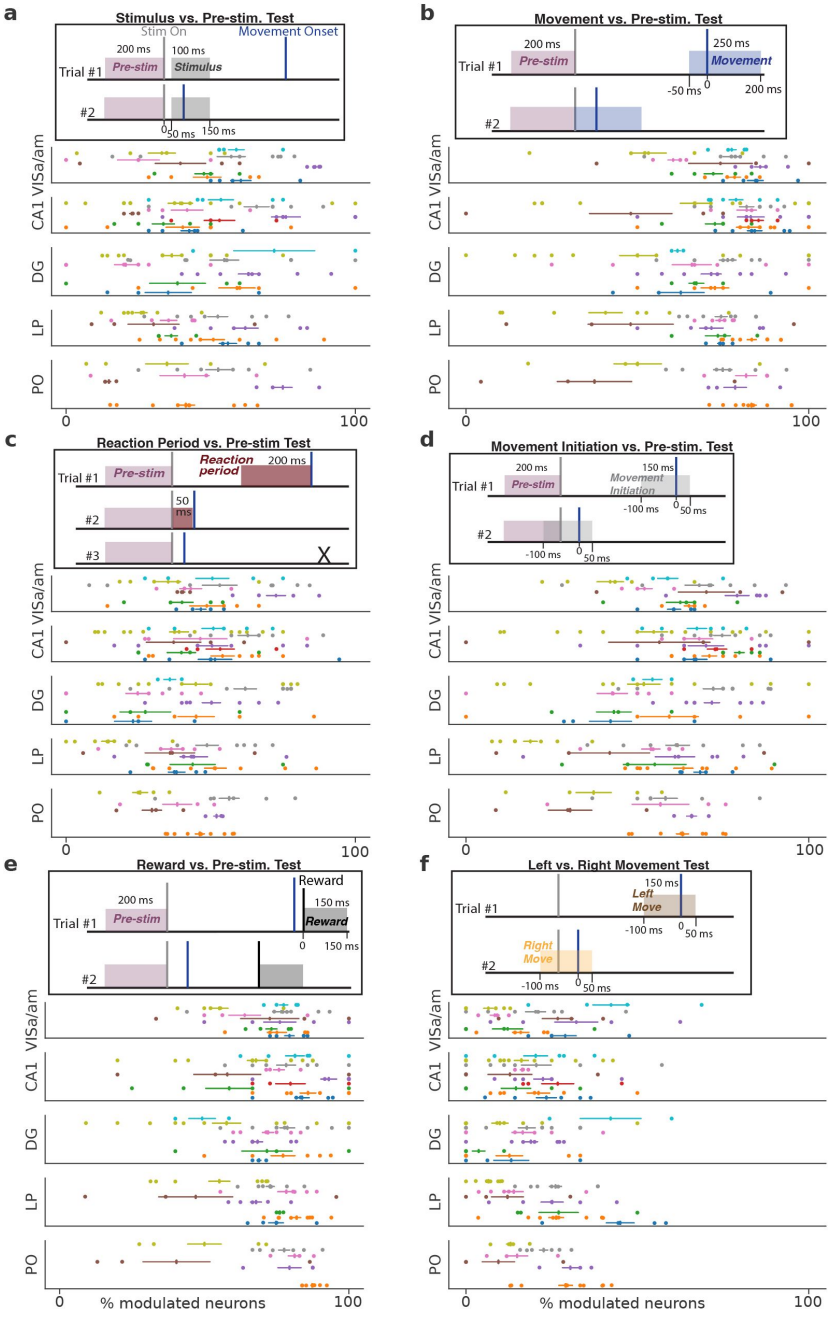


Figure 4–Figure supplement 1.

Proportion of task-modulated neurons, defined by six time-period comparisons, across mice, labs, and brain regions.

(a)-(f), Schematics of six time-period comparisons of task-related time windows 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), to avoid any overlap between pre-stimulus and reaction periods, we calculated the reaction period only in trials with >50 ms between the stimulus and movement onset. We also limited the reaction period to a maximum of 200 ms before movement onset. Below each test schematic, the corresponding proportion of task-modulated neurons across brain regions is shown, colored by lab ID (points are individual sessions; Horizontal lines around vertical marker: S.E.M. around mean across lab sessions).

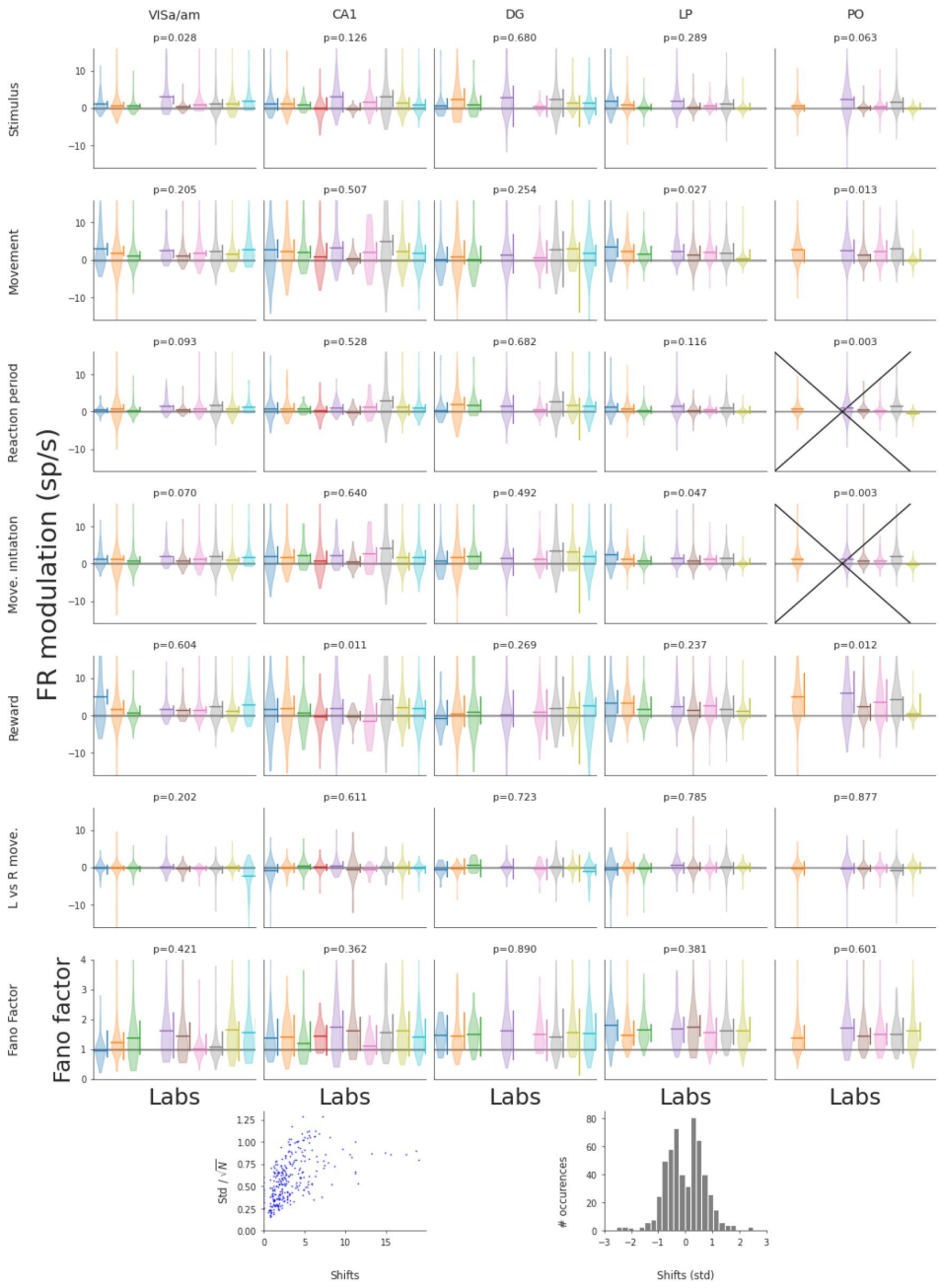


Figure 4–Figure supplement 2.

Power analysis of permutation tests.

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 the 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), and 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 deviation of the corresponding lab distribution.

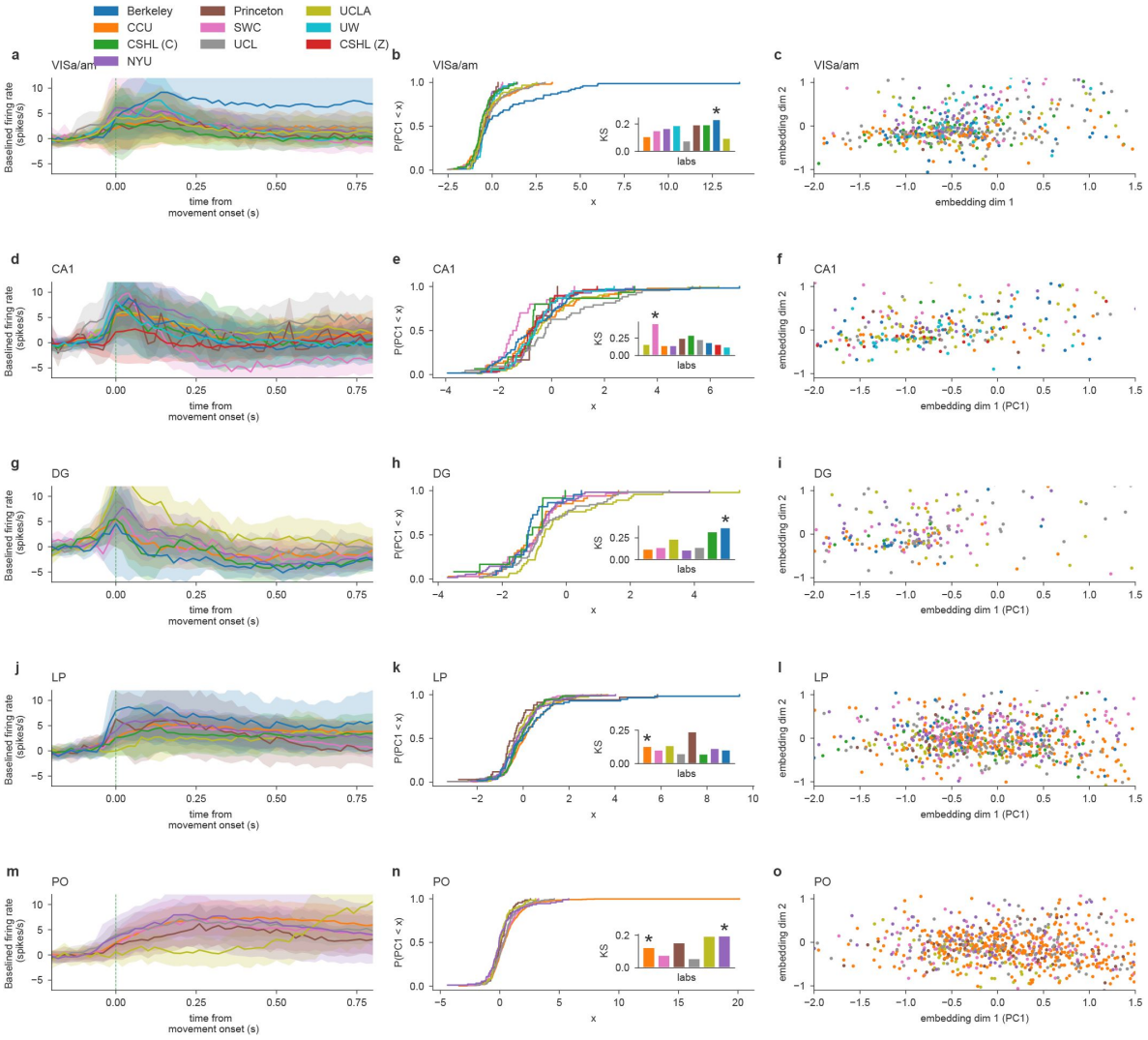


Figure 5–Figure supplement 1.

Lab-grouped average PETH, CDF of the first PC and 2-PC embedding, separate per brain region.

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 Kolmogorov-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. The third column of panels (c, f, i, l, o) displays the embedded activity of neurons from VISa/am, CA1, DG, LP, and PO. For 6 region/lab pairs, dynamics were significantly different from the mean of all remaining labs using the KS test.

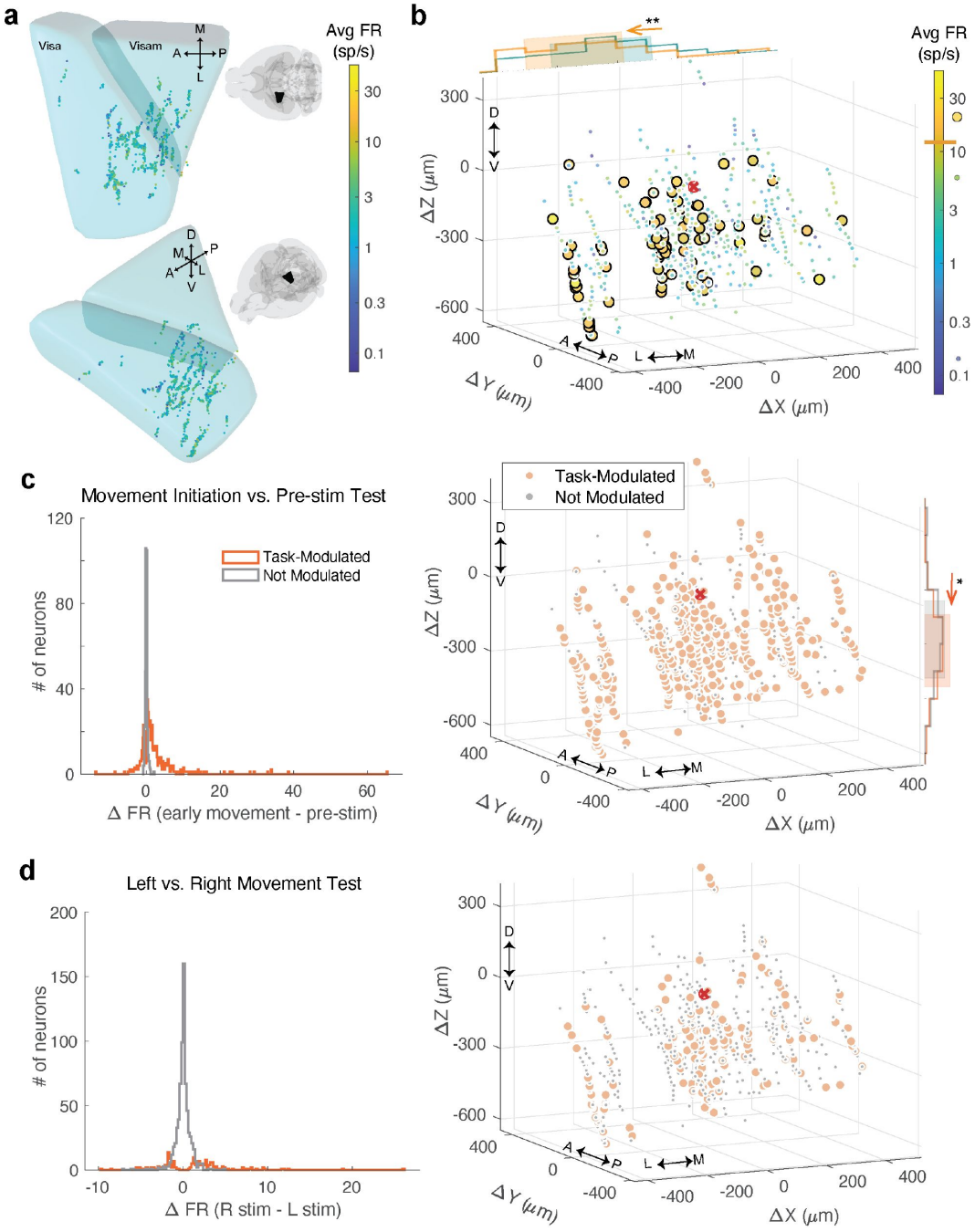


Figure 6–Figure supplement 1.

High-firing and task-modulated VISa/am neurons. (a-d)

Similar to **Figure 6** but for VISa/am. High-firing and some task-modulated VISa/am neurons had slightly different spatial positions than other VISa/am neurons, as shown, but had similar spike waveform features. Out of 877 VISa/am neurons, 550 were modulated during movement initiation (in (c)) and 170 were modulated differently for leftward vs. rightward movements (in (d)).

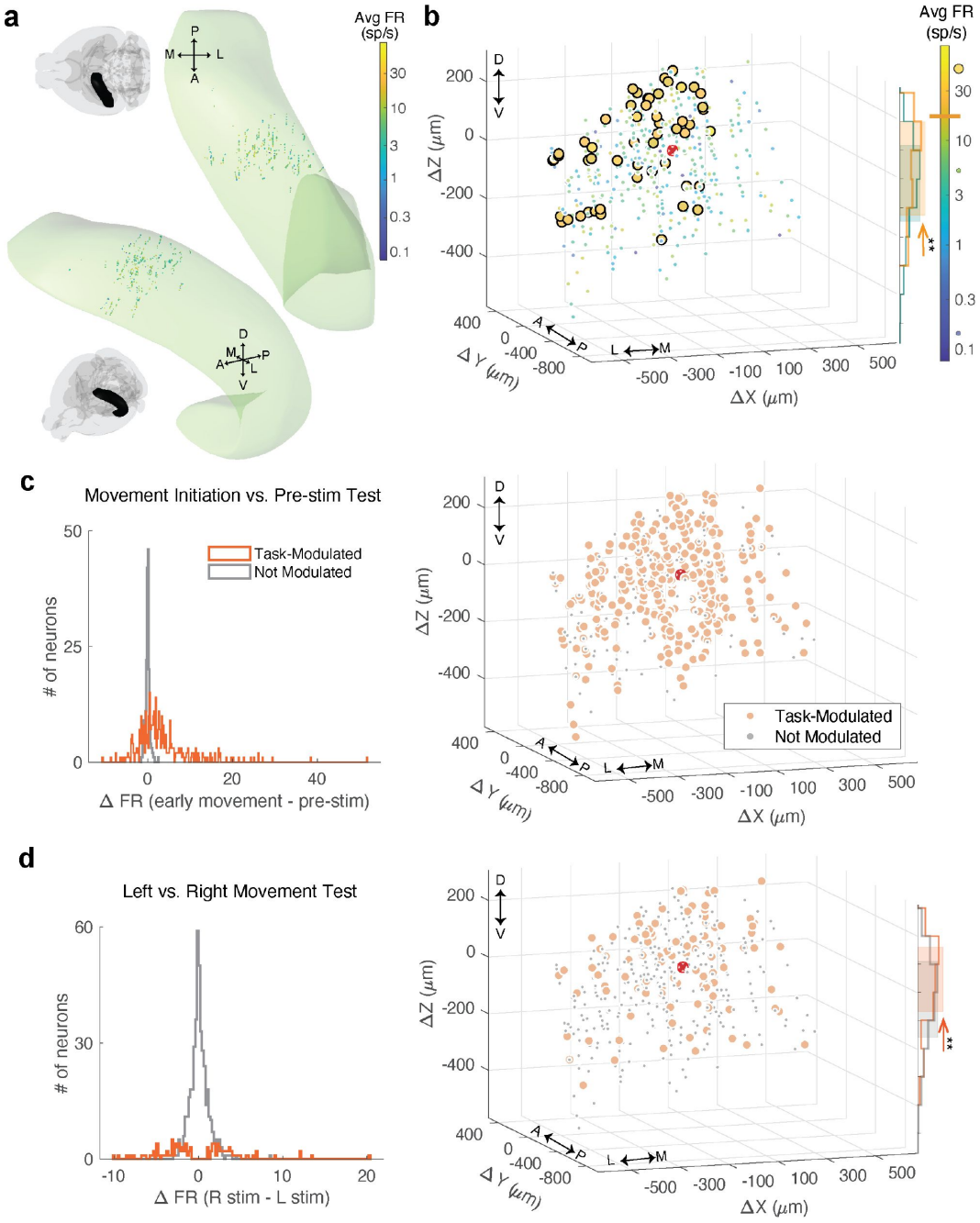


Figure 6–Figure supplement 2.

High-firing and task-modulated CA1 neurons. (a-d)

Similar to **Figure 6** but for CA1. High-firing and some task-modulated CA1 neurons had slightly different spatial positions than other CA1 neurons, as shown, but had similar spike waveform features. Out of 548 CA1 neurons, 366 were modulated during movement initiation (in (c)) and 109 were modulated differently for leftward vs. rightward movements (in (d)).

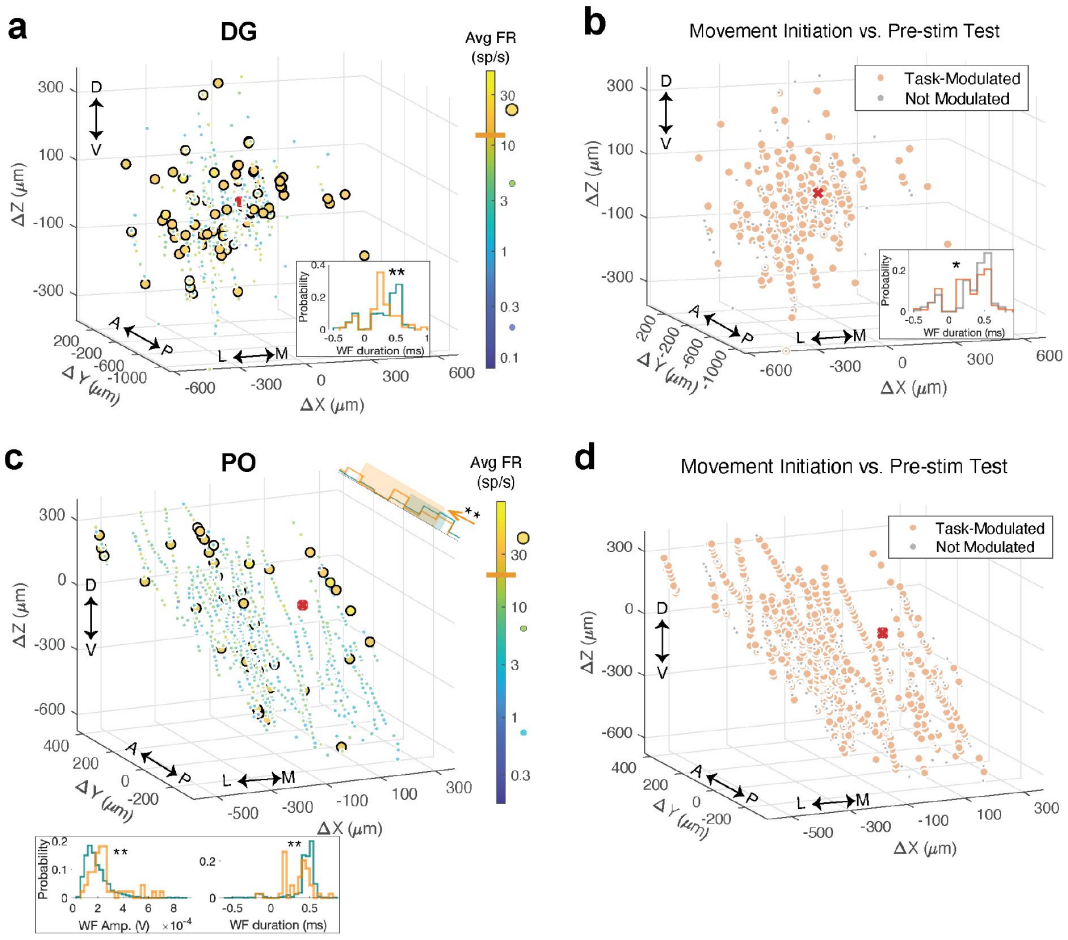


Figure 6–Figure supplement 3.

High-firing and task-modulated DG and PO neurons.

Spatial positions of high-firing and some task-modulated neurons in PO, but not DG, were different from other neurons. **(a-b)**, Spatial positions of DG neurons plotted as distance from the planned target center of mass, indicated with the red x. Spatial positions were not significantly different between high-firing neurons (yellow) and the general population of neurons (blue hues) in (a), nor between task-modulated (orange) and non-modulated (gray) neurons in (b). High-firing neurons and task-modulated neurons (for only some tests) tended to have lower spike durations, possibly related to cell subtype differences. 262 out of 448 DG neurons were modulated during movement initiation (in (b)). **(c-d)**, Same as (a-b) but for PO neurons. Spatial positions and spike waveform features were significantly different between outliers and the general population of neurons (in (c)). Task-modulated and non-modulated PO neurons had small differences in spatial position for only some tests (not the test shown in (d)) and no difference in spike waveform features. 833 out of 1529 PO neurons were modulated during movement initiation (in (d)).

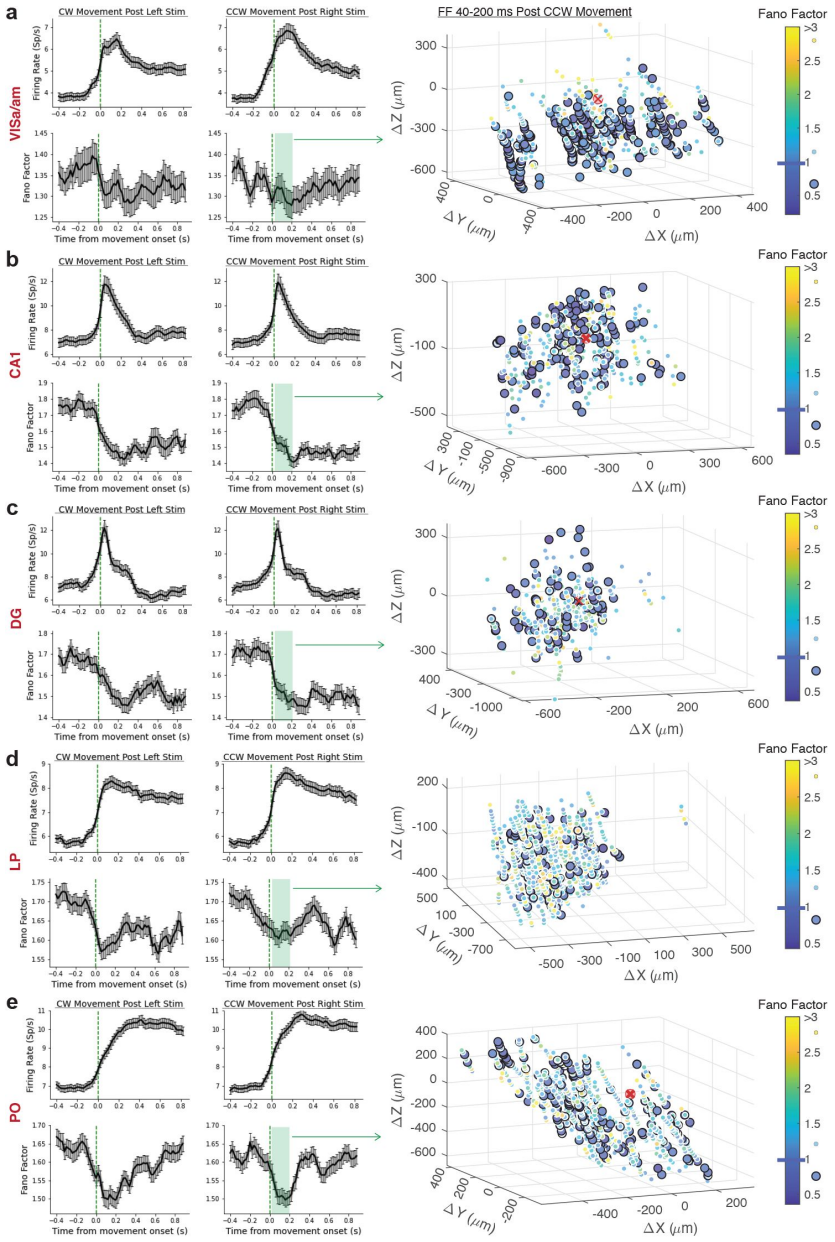


Figure 6–Figure supplement 4.

Time-course and spatial position of neuronal Fano Factors.

(a) *Left column*: 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). In VISa/am, spike duration between high and low Fano Factor neurons was also significantly different, possibly due to cell subtype differences (neurons with shorter spike durations tended to have higher Fano Factors; histograms not shown). Lastly, in CA1 and PO, neurons with larger spike amplitudes had slightly higher Fano Factors.

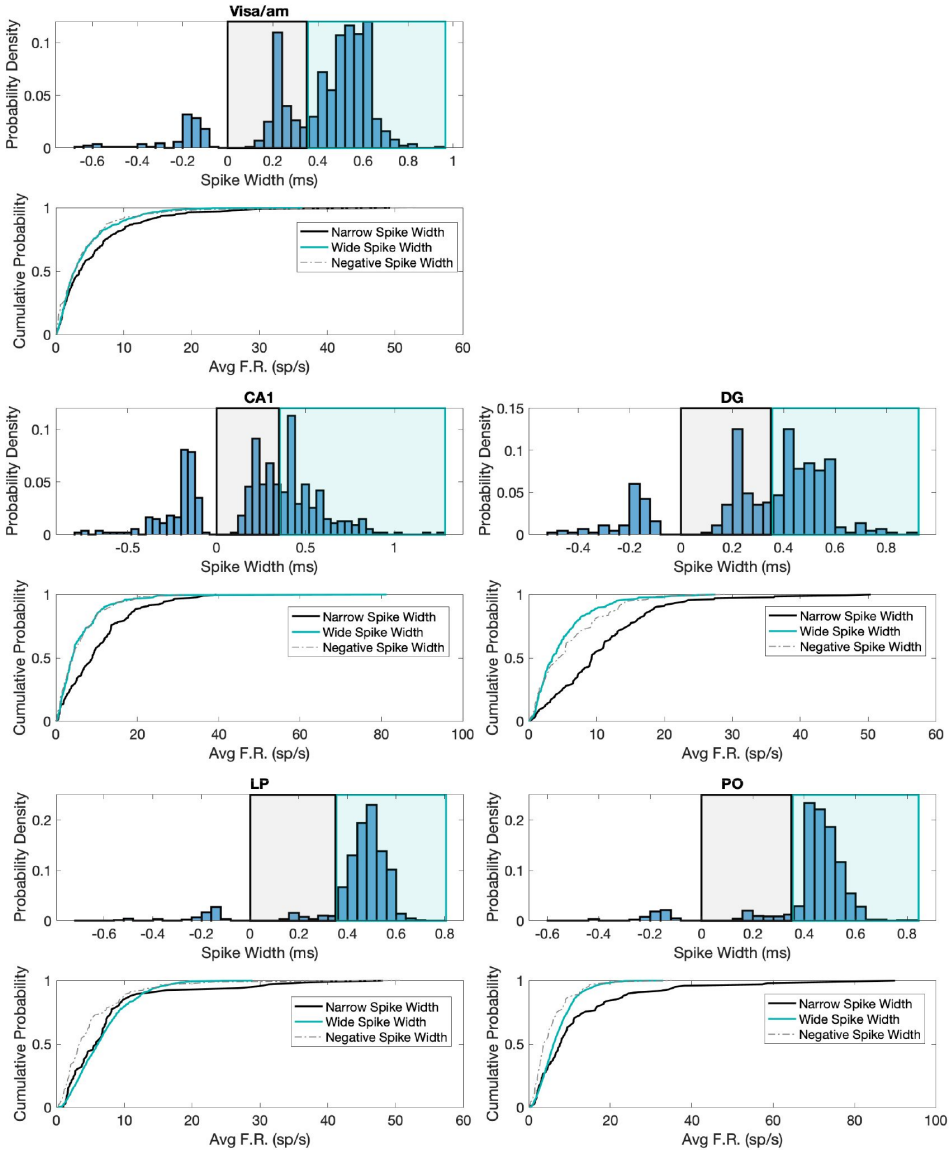


Figure 6–Figure supplement 5.

Neuronal subtypes and firing rates.

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 11-19 sp/s (the firing rate thresholds from **Figure 6** and supplemental figures). Hence, high-firing neurons do not seem to belong to only a specific neuronal subtype.

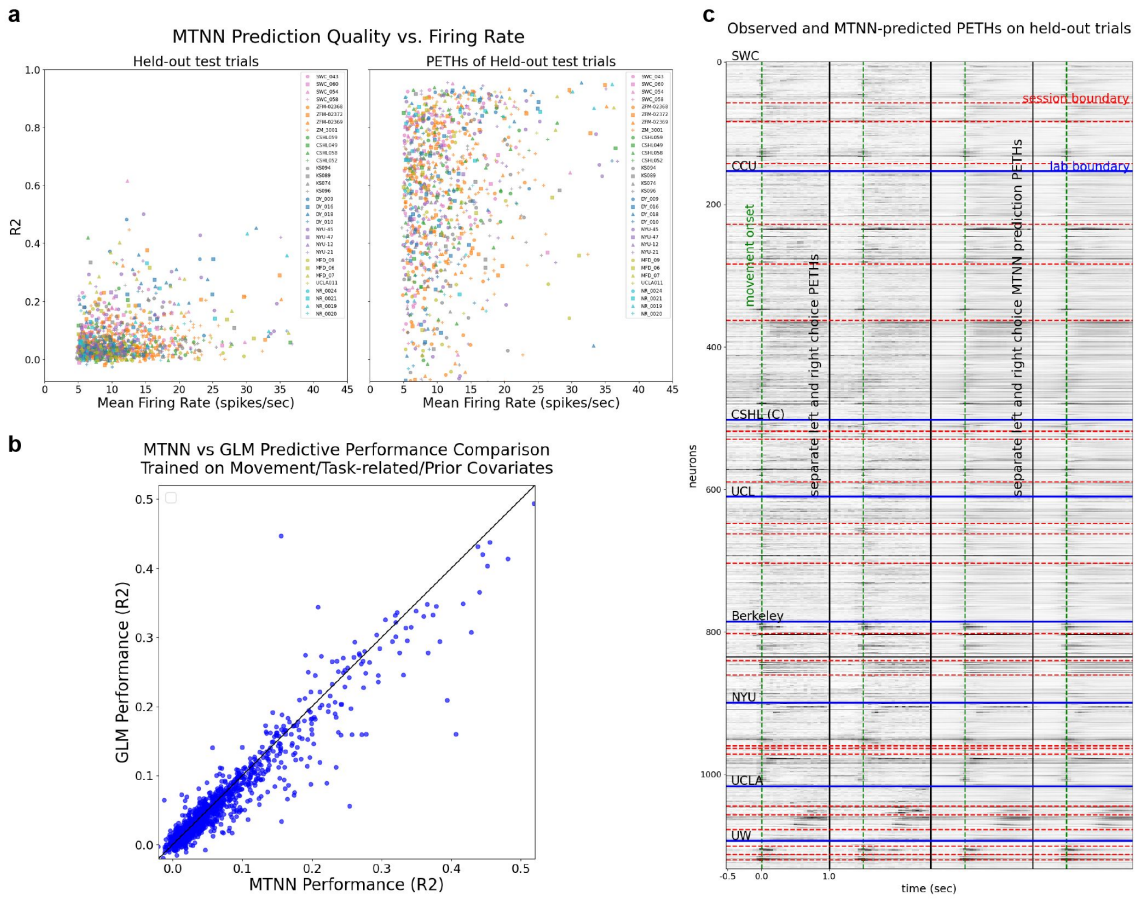


Figure 7–Figure supplement 1.

MTNN prediction quality

(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. *R2* 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 capture visible modulations in the PETHs.

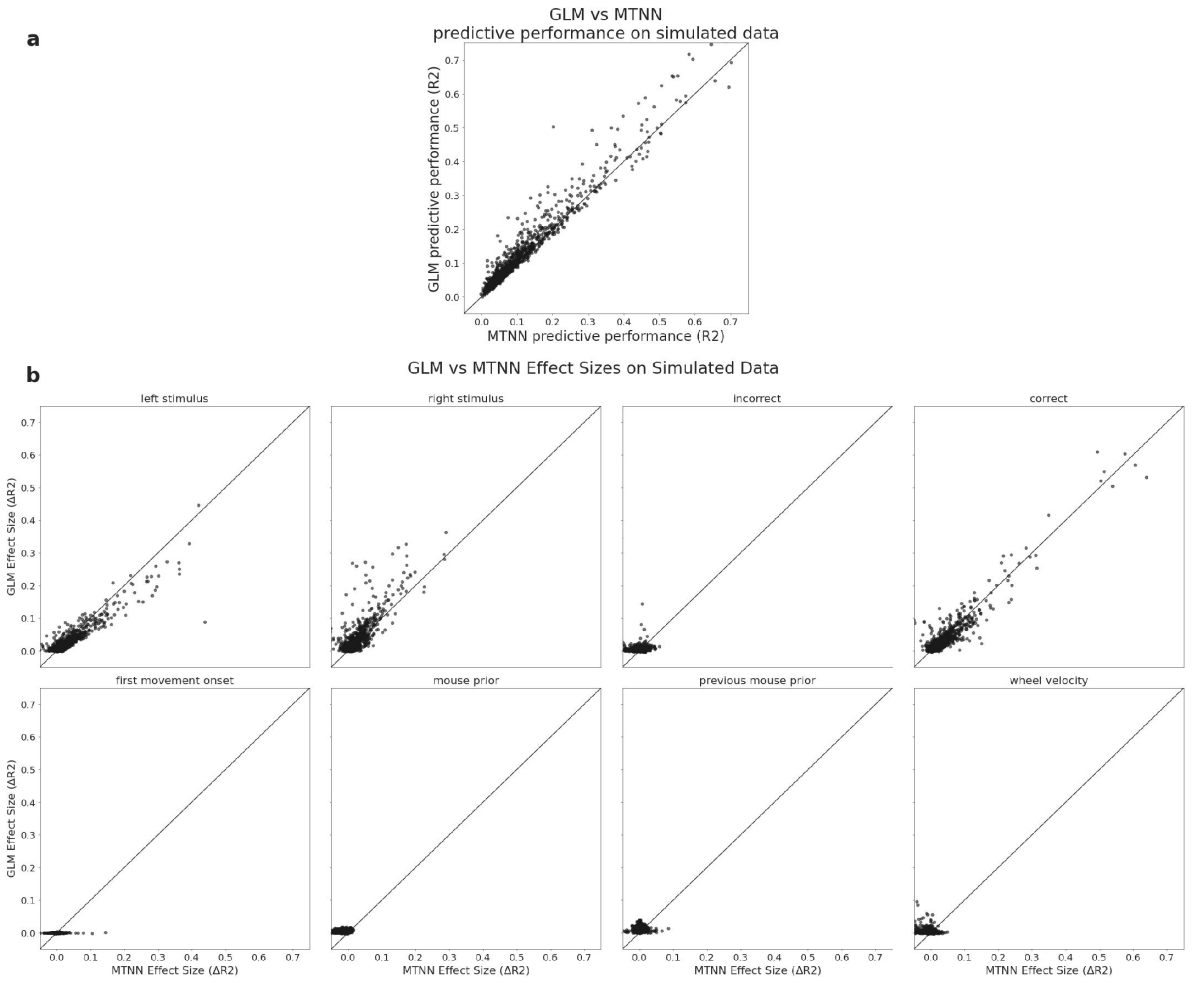


Figure 7–Figure supplement 2.

MTNN prediction quality on the data simulated from GLMs is comparable to the GLMs' prediction quality.

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.

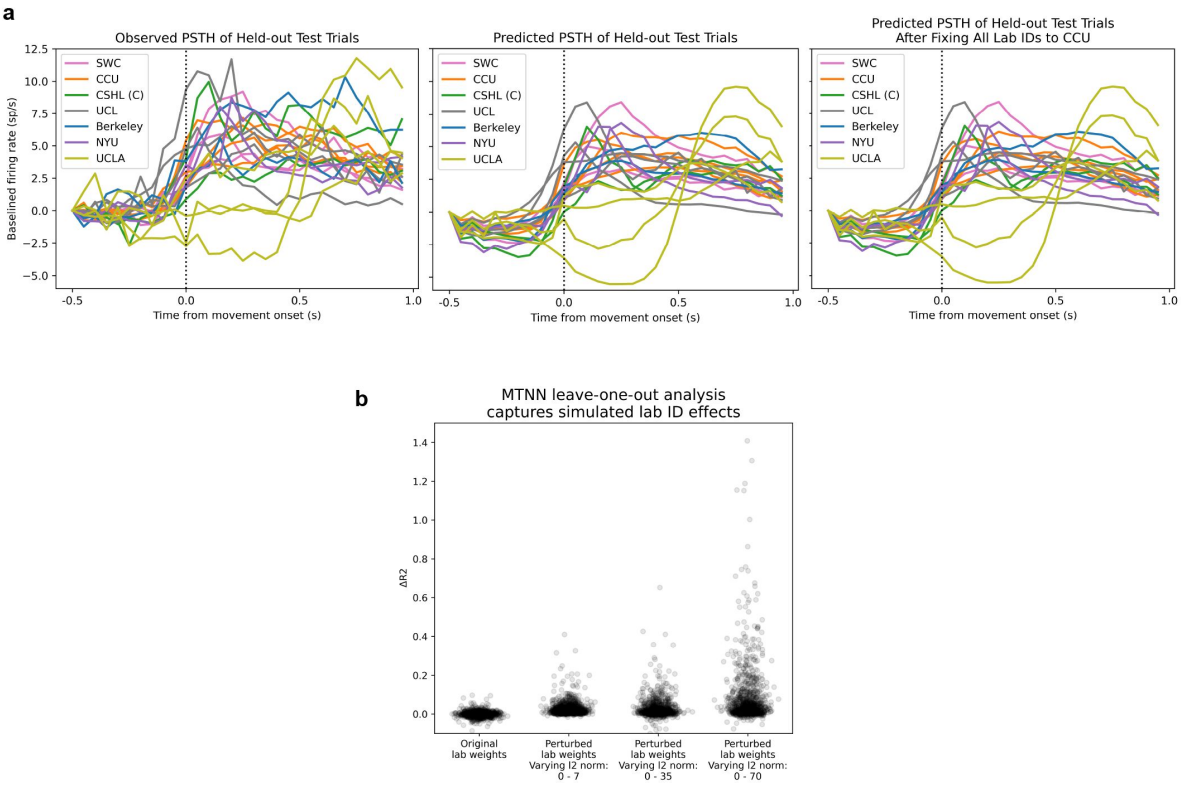


Figure 7–Figure supplement 3.

Lab IDs have a negligible effect on the MTNN prediction. MTNN leave-one-out analysis captures artificially inflated lab ID effects.

(a) (Left) Observed per-lab PETH of held-out test trials for PO. (Center) Per-lab PETH of held-out test trials with MTNN predictions for PO. (Right) Per-lab PETH of held-out test trials with MTNN predictions for PO, after fixing the lab IDs of the input to CCU. **(b)** Here we verify that the MTNN leave-one-out analysis is sensitive enough to capture variability induced by lab IDs. First, given an MTNN model trained with the full set of covariates listed in **Table 2**, we created four different MTNN models by perturbing the weights for the lab IDs. One is a model with the originally learned weights for the lab IDs ("Original lab weights"). The other three models have (randomly sampled) orthogonal weights for the lab IDs, where the l2 norms of the weights for the first and last labs (i.e., the labs with lab IDs 1 and 8) are set to *a* and *b*, respectively ("Perturbed lab weights, Varying l2 norm: *a*-*b*"). The l2 norms of the weights for the labs in between (i.e., the labs with lab IDs from 2 to 7) are set to values equally spaced between *a* and *b*. We then generated four different sets of datasets by running the models forward with the original features that are used to train, validate and test the model. Finally, we reran the MTNN leave-one-out analysis for the lab IDs with each simulated dataset. Each dot in each column represents the change in prediction quality (measured by *R*2) for each neuron in the dataset, when the lab IDs are left out from the set of covariates.

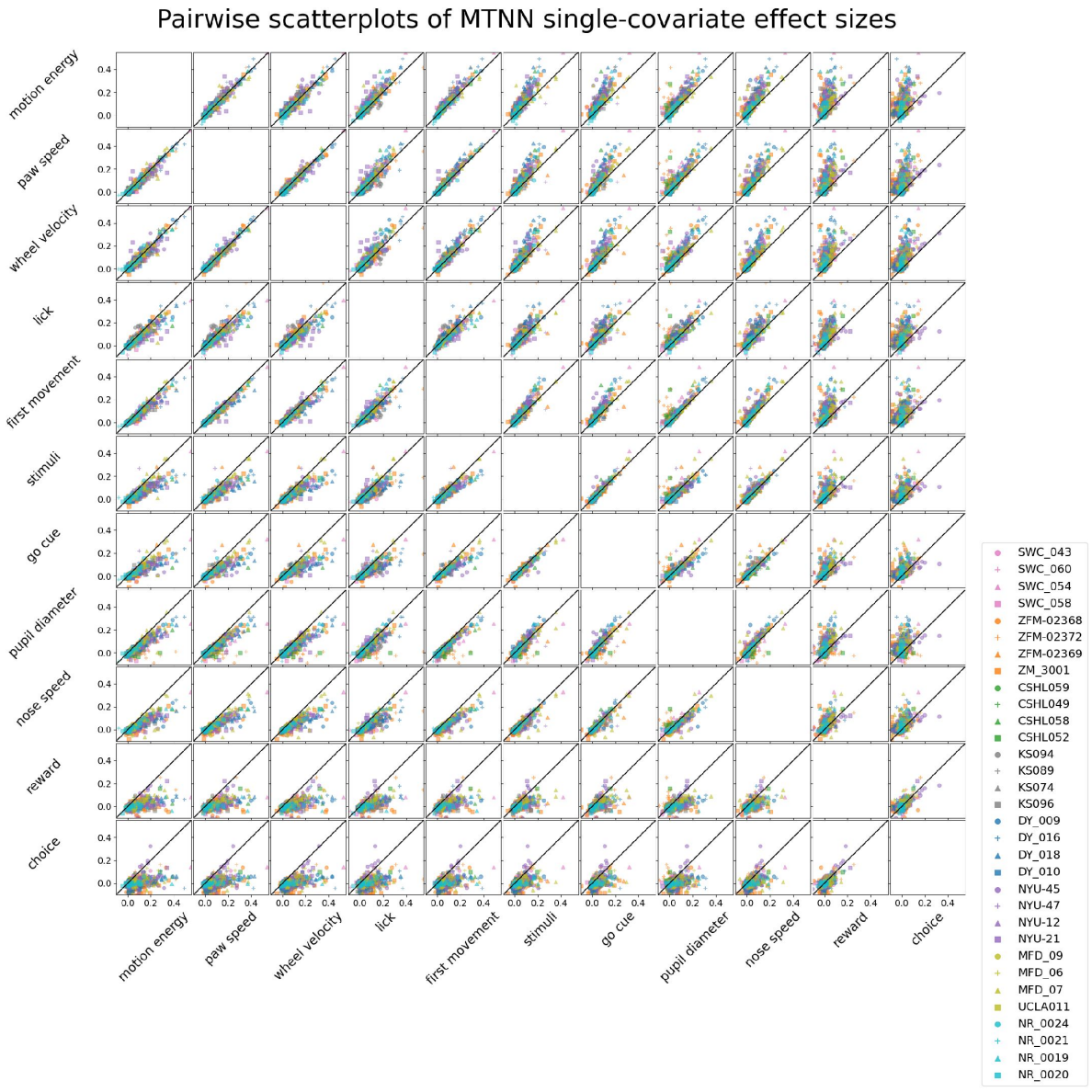


Figure 7–Figure supplement 4.

We plot pairwise scatterplots of MTNN single-covariate effect sizes. Each dot represents the effect sizes of one neuron and is colored by lab. Many of the effect sizes are highly correlated across sessions and labs.

References

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

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

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

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

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

Biderman D *et al.* (2023) **Lightning Pose: improved animal pose estimation via semi-supervised learning, Bayesian ensembling, and cloud-native open-source tools** *bioRxiv*

Birman D, Yang KJ, West SJ, Karsh B, Browning Y, Siegle JH, Steinmetz NA (2023) **Pinpoint: trajectory planning for multi-probe electrophysiology and injections in an interactive web-based 3D environment** *eLife* **12** https://doi.org/10.1101/2023.07.14.548952

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

Brazma A *et al.* (2001) **Minimum information about a microarray experiment (MIAME)—toward standards for microarray data** *Nature genetics* **29**:365–371

Browning Y, Lynch GF, Totten S, Lee D, Svoboda K, Siegle JH (2023) **Brain-wide, MRI-guided electrophysiology** *Society for Neuroscience Abstracts*

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

Campbell R. (2020) **BakingTray** *GitHub*

Campbell R. (2021) **StitchIt** *GitHub*

Campbell R, Blot A, Rousseau C, Winter O. (2020) **Lasagna** *GitHub*

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

Crabbe JC, Wahlsten D, Dudek BC (1999) **Genetics of mouse behavior: interactions with laboratory environment** *Science* **284**:1670–1672

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

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

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

Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA (2021) **Investigating the replicability of preclinical cancer biology** *eLife* eLife Sciences Publications, Ltd https://doi.org/10.7554/eLife.71601

Faulkner M. (2020) **Ephys Atlas GUI** *GitHub*

Goard MJ, Pho GN, Woodson J, Sur M (2016) **Distinct roles of visual, parietal, and frontal motor cortices in memory-guided sensorimotor decisions** *elife* **5**

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

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

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

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

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

Jun JJ *et al.* (2017) **Fully integrated silicon probes for high-density recording of neural activity** *Nature* **551**:232–236

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

Klionsky DJ (2016) **Developing a set of guidelines for your research field: a practical approach** *Mol Biol Cell* **27**:733–8 https://doi.org/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. (2016) **Demixed principal component analysis of neural population data** *Elife* **5**

Li X *et al.* (2021) **Moving Beyond Processing and Analysis-Related Variation in Neuroscience** *bioRxiv* https://doi.org/10.1101/2021.12.01.470790

Lithgow GJ, Driscoll M, Phillips P (2017) **A long journey to reproducible results** *Nature* **548**:387–388 https://doi.org/10.1038/548387a

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

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

Lopes G *et al.* (2015) **Bonsai: an event-based framework for processing and controlling data streams** *Frontiers in neuroinformatics* **9**

Lopes G *et al.* (2021) **Creating and controlling visual environments using BonVision** *Elife* **10**

Lucanic M *et al.* (2017) **Impact of genetic background and experimental reproducibility on identifying chemical compounds with robust longevity effects** *Nature communications* **8**:1–13

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

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

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

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

Nichols TE *et al.* (2017) **Best practices in data analysis and sharing in neuroimaging using MRI** *Nature neuroscience* **20**:299–303

Ólafsdóttir HF, Barry C, Saleem AB, Hassabis D, Spiers HJ (2015) **Hippocampal place cells construct reward related sequences through unexplored space** *Elife* **4**

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

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

Rajasethupathy P *et al.* (2015) **Projections from neocortex mediate top-down control of memory retrieval** *Nature* **526**:653–659

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

Rossant C, Winter O, Hunter M, Huntenburg J, Faulkner M, Wells M, Steinmetz N, Harris K, Bonacchi N. (2021) **Alyx** *GitHub*

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

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

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

Siegle JH *et al.* (2021) **Survey of spiking in the mouse visual system reveals functional hierarchy** *Nature* **592**:86–92

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

Sittig LJ, Carbonetto P, Engel KA, Krauss KS, Barrios-Camacho CM, Palmer AA (2016) **Genetic background limits generalizability of genotypephenotype relationships** *Neuron* **91**:1253–1259

Steinmetz NA *et al.* (2021) **Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings** *Science* **372**

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

The International Brain Laboratory (2021) **iblvideo** *GitHub*

The International Brain Laboratory (2021) **pykilosort** *GitHub*

The International Brain Laboratory *et al.* (2021) **Standardized and reproducible measurement of decision-making in mice** *eLife* **10** https://doi.org/10.7554/eLife.63711

The International Brain Laboratory *et al.* (2022) **Spike sorting pipeline for the International Brain Laboratory** *figshare* https://doi.org/10.6084/m9.figshare.19705522.v3

The International Brain Laboratory *et al.* (2022) **Video hardware and software for the International Brain Laboratory** *figshare* https://doi.org/10.6084/m9.figshare.19694452

Tomar S (2006) **Converting video formats with FFmpeg** *Linux Journal* **146**

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

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

Voelkl B *et al.* (2020) **Reproducibility of animal research in light of biological variation** *Nature Reviews Neuroscience* **21**:384–393 https://doi.org/10.1038/s41583-020-0313-3

de Vries SEJ *et al.* (2020) **A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex** *Nature Neuroscience* **23**:138–151 https://doi.org/10.1038/s41593-019-0550-9

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

Wang Q *et al.* (2020) **The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas** *Cell* **181**:936–953 https://doi.org/10.1016/j.cell.2020.04.007

West SJ (2021) **BrainRegister** *GitHub*

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

Editors

Reviewing Editor
Caleb Kemere
Rice University, Houston, United States of America

Senior Editor
Michael Frank
Brown University, Providence, United States of America

Reviewer #1 (Public review):

Summary:

The authors explore a large-scale electrophysiological dataset collected in 10 labs while mice performed the same behavioral task, and aim to establish guidelines to aid reproducibility of results collected across labs. They introduce a series of metrics for quality control of electrophysiological data and show that histological verification of recording sites is important for interpreting findings across labs and should be reported in addition to planned coordinates. Furthermore, the authors suggest that although basic electrophysiology features were comparable across labs, task modulation of single neurons can be variable, particularly for some brain regions. The authors then use a multi-task neural network model to examine how neural dynamics relate to multiple interacting task- and experimenter-related variables, and find that lab-specific differences contribute little to the variance observed. Therefore, analysis approaches that account for correlated behavioral variables are important for establishing reproducible results when working with electrophysiological data from animals performing decision-making tasks. This paper is very well-motivated and needed. However, what is missing is a direct comparison of task modulation of neurons across labs using standard analysis practice in the fields, such as generalized linear model (GLM). This can potentially clarify how much behavioral variance contributes to the neural variance across labs; and more accurately estimate the scale of the issues of reproducibility in behavioral systems neuroscience, where conclusions often depend on these standard analysis methods.

Strength:

(1) This is a well-motivated paper that addresses the critical question of reproducibility in behavioural systems neuroscience. The authors should be commended for their efforts.

(2) A key strength of this study comes from the large dataset collected in collaboration across ten labs. This allows the authors to assess lab-to-lab reproducibility of electrophysiological data in mice performing the same decision-making task.

(3) The authors' attempt to streamline preprocessing pipelines and quality metrics is highly relevant in a field that is collecting increasingly large-scale datasets where automation of these steps is increasingly needed.

(4) Another major strength is the release of code repositories to streamline preprocessing pipelines across labs collecting electrophysiological data.

(5) Finally, the application of MTNN for characterizing functional modulation of neurons, although not yet widely used in systems neuroscience, seems to have several advantages over traditional methods.

Weaknesses:

(1) In several places the assumptions about standard practices in the field, including preprocessing and analyses of electrophysiology data, seem to be inaccurately presented:

a) The estimation of how much the histologically verified recording location differs from the intended recording location is valuable information. Importantly, this paper provides citable evidence for why that is important. However, histological verification of recording sites is standard practice in the field, even if not all studies report them. Although we appreciate the authors' effort to further motivate this practice, the current description in the paper may give readers outside the field a false impression of the level of rigor in the field.

b) When identifying which and how neurons encode particular aspects of stimuli or behaviour in behaving animals (when variables are correlated by the nature of the animals behaviour), it has become the standard in behavioral systems neuroscience to use GLMs - indeed many labs participating in the IBL also has a long history of doing this (e.g., Steinmetz et al., 2019; Musall et al., 2023; Orsolic et al., 2021; Park et al., 2014). The reproducibility of results when using GLMs is never explicitly shown, but the supplementary figures to Figure 7 indicate that results may be reproducible across labs when using GLMs (as it has similar prediction performance to the MTNN). This should be introduced as the first analysis method used in a new dedicated figure (i.e., following Figure 3 and showing results of analyses similar to what was shown for the MTNN in Figure 7). This will help put into perspective the degree of reproducibility issues the field is facing when analyzing with appropriate and common methods. The authors can then go on to show how simpler approaches (currently in Figures 4 and 5) - not accounting for a lot of uncontrolled variabilities when working with behaving animals - may cause reproducibility issues.

When the authors introduce a neural network approach (i.e. MTNN) as an alternative to the analyses in Figures 4 and 5, they suggest: 'generalized linear models (GLMs) are likely too inflexible to capture the nonlinear contributions that many of these variables, including lab identity and spatial positions of neurons, might make to neural activity'). This is despite the comparison between MTNN and GLM prediction performance (Supplement 1 to Figure 7) showing that the MTNN is only slightly better at predicting neural activity compared to standard GLMs. The introduction of new models to capture neural variability is always welcome, but the conclusion that standard analyses in the field are not reproducible can be unfair unless directly compared to GLMs.

In essence, it is really useful to demonstrate how different analysis methods and preprocessing approaches affect reproducibility. But the authors should highlight what is actually standard in the field, and then provide suggestions to improve from there.

(2) The authors attempt to establish a series of new quality control metrics for the inclusion of recordings and single units. This is much needed, with the goal to standardize unit inclusion across labs that bypasses the manual process while keeping the nuances from manual curation. However, the authors should benchmark these metrics to other automated metrics and to manual curation, which is still a gold standard in the field. The authors did this for whole-session assessment but not for individual clusters. If the authors can find metrics that capture agreed-upon manual cluster labels, without the need for manual intervention, that would be extremely helpful for the field.

(3) With the goal of improving reproducibility and providing new guidelines for standard practice for data analysis, the authors should report of n of cells, sessions, and animals used in plots and analyses throughout the paper to aid both understanding of the variability in the plots - but also to set a good example.

Other general comments:

(1) In the discussion (line 383) the authors conclude: 'This is reassuring, but points to the need for large sample sizes of neurons to overcome the inherent variability of single neuron recording'. - Based on what is presented in this paper we would rather say that their results suggest that appropriate analytical choices are needed to ensure reproducibility, rather than large datasets - and they need to show whether using standard GLMs actually allows for reproducible results.

(2) A general assumption in the across-lab reproducibility questions in the paper relies on intralab variability vs across-lab variability. An alternative measure that may better reflect experimental noise is across-researcher variability, as well as the amount of experimenter experience (if the latter is a factor, it could suggest researchers may need more training before collecting data for publication). The authors state in the discussion that this is not possible. But maybe certain measures can be used to assess this (e.g. years of conducting surgeries/ephys recordings etc)?

(3) Figure 3b and c: Are these plots before or after the probe depth has been adjusted based on physiological features such as the LFP power? In other words, is the IBL electrophysiological alignment toolbox used here and is the reliability of location before using physiological criteria or after? Beyond clarification, showing both before and after would help the readers to understand how much the additional alignment based on electrophysiological features adjusts probe location. It would also be informative if they sorted these penetrations by which penetrations were closest to the planned trajectory after histological verification.

(4) In Figures 4 and 6: If the authors use a 0.05 threshold (alpha) and a cell simply has to be significant on 1/6 tests to be considered task modulated, that means that they have a false positive rate of ~30% (0.05*6=0.3). We ran a simple simulation looking for significant units (from random null distribution) from these criteria which shows that out of 100.000 units, 26500 units would come out significant (false error rate: 26.5%). That is very high (and unlikely to be accepted in most papers), and therefore not surprising that the fraction of task-modulated units across labs is highly variable. This high false error rate may also have implications for the investigation of the spatial position of task-modulated units (as effects of the spatial position may drown in falsely labelled 'task-modulated' cells).

(5) The authors state from Figure 5b that the majority of cells could be well described by 2 PCs. The distribution of R2 across neurons is almost uniform, so depending on what R2 value one considers a 'good' description, that is the fraction of 'good' cells. Furthermore, movement onset has now been well-established to be affecting cells widely and in large fractions, so while this analysis may work for something with global influence - like movement - more sparsely encoded variables (as many are in the brain) may not be well approximated with

this suggestion. The authors could expand this analysis into other epochs like activity around stimulus presentation, to better understand how this type of analysis reproduces across labs for features that have a less global influence.

(6) Additionally, in Figure 5i: could the finding that one can only distinguish labs when taking cells from all regions, simply be a result of a different number of cells recorded in each region for each lab? It makes more sense to focus on the lab/area pairing as the authors also do, but not to make their main conclusion from it. If the authors wish to do the comparison across regions, they will need to correct for the number of cells recorded in each region for each lab. In general, it was a struggle to fully understand the purpose of Figure 5. While population analysis and dimensionality reduction are commonplace, this seems to be a very unusual use of it.

(7) In the discussion the authors state: "This approach, which exceeds what is done in many experimental labs". Indeed this approach is a more effective and streamlined way of doing it, but it is questionable whether it 'exceeds' what is done in many labs. Classically, scientists trace each probe manually with light microscopy and designate each area based on anatomical landmarks identified with nissl or dapi stains together with gross landmarks. When not automated with 2-PI serial tomography and anatomically aligned to a standard atlas, this is a less effective process, but it is not clear that it is less precise, especially in studies before neuropixels where active electrodes were located in a much smaller area. While more effective, transforming into a common atlas does make additional assumptions about warping the brain into the standard atlas - especially in cases where the brain has been damaged/lesioned. Readers can appreciate the effectiveness and streamlining provided by these new tools without the need to invalidate previous approaches.

(8) What about across-lab population-level representation of task variables, such as in the coding direction for stimulus or choice? Is the general decodability of task variables from the population comparable across labs?

https://doi.org/10.7554/eLife.100840.1.sa1

Reviewer #2 (Public review):

Summary:

The authors sought to evaluate whether observations made in separate individual laboratories are reproducible when they use standardized procedures and quality control measures. This is a key question for the field. If ten systems neuroscience labs try very hard to do the exact same experiment and analyses, do they get the same core results? If the answer is no, this is very bad news for everyone else! Fortunately, they were able to reproduce most of their experimental findings across all labs. Despite attempting to target the same brain areas in each recording, variability in electrode targeting was a source of some differences between datasets.

Major Comments:

The paper had two principal goals:
(1) to assess reproducibility between labs on a carefully coordinated experiment
(2) distill the knowledge learned into a set of standards that can be applied across the field.
The manuscript made progress towards both of these goals but leaves room for improvement.

(1) The first goal of the study was to perform exactly the same experiment and analyses across 10 different labs and see if you got the same results. The rationale for doing this was to test how reproducible large-scale rodent systems neuroscience experiments really are. In

this, the study did a great job showing that when a consortium of labs went to great lengths to do everything the same, even decoding algorithms could not discern laboratory identity was not clearly from looking at the raw data. However, the amount of coordination between the labs was so great that these findings are hard to generalize to the situation where similar (or conflicting!) results are generated by two labs working independently.

Importantly, the study found that electrode placement (and thus likely also errors inherent to the electrode placement reconstruction pipeline) was a key source of variability between datasets. To remedy this, they implemented a very sophisticated electrode reconstruction pipeline (involving two-photon tomography and multiple blinded data validators) in just one lab-and all brains were sliced and reconstructed in this one location. This is a fantastic approach for ensuring similar results within the IBL collaboration, but makes it unclear how much variance would have been observed if each lab had attempted to reconstruct their probe trajectories themselves using a mix of histology techniques from conventional brain slicing, to light sheet microscopy, to MRI imaging.

This approach also raises a few questions. The use of standard procedures, pipelines, etc. is a great goal, but most labs are trying to do something unique with their setup. Bigger picture, shouldn't highly "significant" biological findings akin to the discovery of place cells or grid cells, be so clear and robust that they can be identified with different recording modalities and analysis pipelines?

Related to this, how many labs outside of the IBL collaboration have implemented the IBL pipeline for their own purposes? In what aspects do these other labs find it challenging to reproduce the approaches presented in the paper? If labs were supposed to perform this same experiment, but without coordinating directly, how much more variance between labs would have been seen? Obviously investigating these topics is beyond the scope of this paper. The current manuscript is well-written and clear as is, and I think it is a valuable contribution to the field. However, some additional discussion of these issues would be helpful.

(2) The second goal of the study was to present a set of data curation standards (RIGOR) that could be applied widely across the field. This is a great idea, but its implementation needs to be improved if adoption outside of the IBL is to be expected. Here are three issues:

(a) The GitHub repo for this project (https://github.com/int-brain-lab/paper-reproducible-ephys/) is nicely documented if the reader's goal is to reproduce the figures in the manuscript. Consequently, the code for producing the RIGOR statistics seems mostly designed for re-computing statistics on the existing IBL-formatted datasets. There doesn't appear to be any clear documentation about how to run it on arbitrary outputs from a spike sorter (i.e. the inputs to Phy).

(b) Other sets of spike sorting metrics that are more easily computed for labs that are not using the IBL pipeline already exist (e.g. "quality_metrics" from the Allen Institute ecephys pipeline [https://github.com/AllenInstitute/ecephys_spike_sorting/blob/main/ecephys_spike_sorting/modules/quality_metrics/README.md] and the similar module in the Spike Interface package [https://spikeinterface.readthedocs.io/en/latest/modules/qualitymetrics.html]). The manuscript does not compare these approaches to those proposed here, but some of the same statistics already exist (amplitude cutoff, median spike amplitude, refractory period violation).

(c) Some of the RIGOR criteria are qualitative and must be visually assessed manually. Conceptually, these features make sense to include as metrics to examine, but would ideally be applied in a standardized way across the field. The manuscript doesn't appear to contain a detailed protocol for how to assess these features. A procedure for how to apply these criteria for curating non-IBL data (or for implementing an automated classifier) would be helpful.

Other Comments:

(1) How did the authors select the metrics they would use to evaluate reproducibility? Was this selection made before doing the study?

(2) Was reproducibility within-lab dependent on experimenter identity?

(3) They note that UCLA and UW datasets tended to miss deeper brain region targets (lines 185-188) - they do not speculate why these labs show systematic differences. Were they not following standardized procedures?

(4) The authors suggest that geometrical variance (difference between planned and final identified probe position acquired from reconstructed histology) 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. In this case, the use of skull landmarks (e.g. bregma) to determine locations of brain structures might be unreliable and provide an error of ~360 microns. While it is known that there is indeed variance in the position between skull landmarks and brain areas in different animals, the quantification of this error is a useful value for the field.

(5) Why are the thalamic recording results particularly hard to reproduce? Does the anatomy of the thalamus simply make it more sensitive to small errors in probe positioning relative to the other recorded areas?

https://doi.org/10.7554/eLife.100840.1.sa0