Reproducibility of in-vivo electrophysiological measurements in mice

International Brain Laboratory[*], Kush Banga[7], Julius Benson[11], Niccolò Bonacchi[2], Sebastian A Bruijns[1], Rob Campbell[13], Gaëlle A Chapuis[5], Anne K Churchland[6], M Felicia Davatolhagh[6], Hyun Dong Lee[3], Mayo Faulkner[7], Fei Hu[9], Julia Hunterberg[2], Anup Khanal[6], Christopher Krasniak[10], Guido T Meijer[2], Nathaniel J Miska[7], Zeinab Mohammadi[12], Jean-Paul Noel[11], Liam Paninski[3], Alejandro Pan-Vazquez[12], Noam Roth[4], Michael Schartner[2], Karolina Socha[7], Nicholas A Steinmetz[4], Marsa Taheri[6], Anne E Urai[8], Miles Wells[7], Steven J West[7], Matthew R Whiteway[3], Olivier Winter[2]

***For correspondence:** anne.churchland@internationalbrainlab.org (AKC); liam.paninski@internationalbrainlab.org (LP); nicholas.steinmetz@internationalbrainlab.org (NAS)

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

May 9, 2022

Abstract

Understanding whole-brain-scale electrophysiological recordings will rely on the collective work of multiple labs. Because two labs recording from the same brain area often reach different conclusions, it is critical to quantify and control for features that decrease reproducibility. To address these issues, we formed a multi-lab collaboration using a shared, open-source behavioral task and experimental apparatus. We repeatedly inserted Neuropixels multi-electrode probes targeting the same brain locations (including posterior parietal cortex, hippocampus, and thalamus) in mice performing the behavioral task. We gathered data across 9 labs and developed a common histological and data processing pipeline to analyze the resulting large datasets. After applying stringent behavioral, histological, and electrophysiological quality-control criteria, we found that neuronal yield, firing rates, spike amplitudes, and task-modulated neuronal activity were reproducible across laboratories. To quantify variance in neural activity explained by task variables (e.g., stimulus onset time), behavioral variables (timing of licks/paw movements), and other variables (e.g., spatial location in the brain or the lab ID), we developed a multi-task neural network encoding model that extends common, simpler regression approaches by allowing nonlinear interactions between variables. We found that within-lab random effects captured by this model were comparable to between-lab random effects. Taken together, these results demonstrate that across-lab standardization of electrophysiological procedures can lead to reproducible results across labs. Moreover, our protocols to achieve reproducibility, along with our analyses to evaluate it are openly accessible to the scientific community, along with our extensive electrophysiological dataset with corresponding behavior and open-source analysis code.

Introduction

Reproducibility is a cornerstone of the scientific method: a given sequence of experimental methods should lead to comparable results if applied in different laboratories. In some areas of biological and psychological science, however, the reliable generation of reproducible results is a well-known challenge (***Baker, 2016***; ***Voelkl et al., 2020***; ***Li et al., 2021***; ***Errington et al., 2021***). In systems neuroscience at the level of single-cell-resolution recordings, evaluating reproducibility is difficult: experimental methods are sufficiently complex that replicating experiments is technically challenging, and many experimenters feel little incentive to do such experiments since negative results can be difficult to publish. Variability in experimental outcomes has nonetheless 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.

We have previously addressed the issue of reproducibility in the context of mouse psychophysical behavior, by training 140 mice in 7 laboratories and comparing their learning rates, speed, and accuracy in a simple binary visually-driven decision task. We demonstrated that standardized protocols can lead to highly reproducible behavior (***The International Brain Laboratory et al., 2021***). Here, we build on those results by measuring within- and across-lab variability in the context of intra-cerebral electrophysiological recordings. We repeatedly inserted Neuropixels multi-electrode probes (***Jun et al., 2017***) targeting the same brain regions (including posterior parietal cortex, hippocampus, and thalamus) in mice performing the behavioral task from (***The International Brain Laboratory et al., 2021***). We gathered data across 9 different labs and developed a common histological and data processing pipeline to analyze the resulting large datasets.

After applying stringent behavioral, histological, and electrophysiological quality-control criteria, features such as neuronal yield, firing rate, and normalized LFP power were reproducible across laboratories; their within-lab averages did not significantly deviate from the mean across labs. Similarly, the proportions of cells modulated by task events was largely reproducible across labs, as was the Fano Factor, a measure of neural variability. Finally, to quantify variance in neural activity explained by task variables (e.g., stimulus onset time), behavioral variables (timing of licks/paw movements), and other variables (e.g., spatial location in the brain or the lab ID), we developed a multi-task neural network encoding model that extends common, simpler regression approaches by allowing nonlinear interactions between variables. Again, we found that within-lab random effects captured by this model were comparable to between-lab random effects. Taken together, these results suggest that across-lab standardization of electrophysiological procedures can lead to reproducible results across laboratories.

Results

Repeated-site recordings in the same task across multiple labs

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 was successful. Nine 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. In each experiment, Neuropixels 1.0 probes were inserted, targeted at -2.0 mm AP, -2.24 mm ML, 4.0 mm DV relative to bregma; 15° angle (Figure 1c). This site was selected because it encompasses brain regions implicated in visual decision-making, including visual area A (***Najafi et al., 2020***; ***Harvey et al., 2012***), dentate gyrus, CA1, (***Turk-Browne, 2019***), and thalamic nuclei LP and PO (***Saalmann and Kastner, 2011***; ***Roth et al., 2016***).

Probe placement contributes to experimental variability

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

Variability in brain insertions can be assessed by probe placement at the brain surface, and by probe angle. Probe placement at the brain surface comprises two components. The first, 'targeting variability,' was obtained by calculating the difference between the planned and actual probe placement, measured with the micro-manipulator at the time of recording (Figure 2d). Targeting variability is expected to be non-zero because experimenters sometimes move probes slightly from the planned location to avoid blood vessels or irregularities (Figure 2d, top, total mean displacement = 115 μm, exclusion criteria passed mean displacement = 72μm). Reproducibility of targeting variability across labs was evaluated via a permutation test: values were shuffled between the lab identities 10,000 times, and the original targeting variability mean per lab distribution was compared to all permuted distributions to compute a p-value. Targeting variability shows no significant effect across laboratories across all probes (Figure 2d, bottom), permutation test p-value for all probes p=0.2118). When applying our exclusion criteria, including the anatomical requirement that the probe must record from three of our five repeated site brain regions, the computed p-value increased (Figure 2d, bottom), permutation test p-value for exclusion criteria passed probes p=0.2295), indicating the data are more likely from the same distribution. Thus, targeting reproducibility is enhanced with appropriate anatomical exclusion criteria.

The second component of probe placement variability in brain insertions is 'geometrical variability.' Geometrical variability was obtained by calculating the difference between our planned position and the final identified probe position obtained from the reconstructed histology. This encompasses the targeting variance above, plus anatomical differences and errors in defining the stereotaxic coordinate system, including residual errors from a mismatch in skull landmarks and underlying brain structure. Geometrical variability was likewise non-zero (Figure 2e, top, total mean displacement = 392 μm, exclusion criteria passed mean displacement = 253 μm) with some individual insertion locations up to 1500 μm from the planned coordinate. Assessing geometrical variability for all probes with permutation testing revealed no significant effect across laboratories

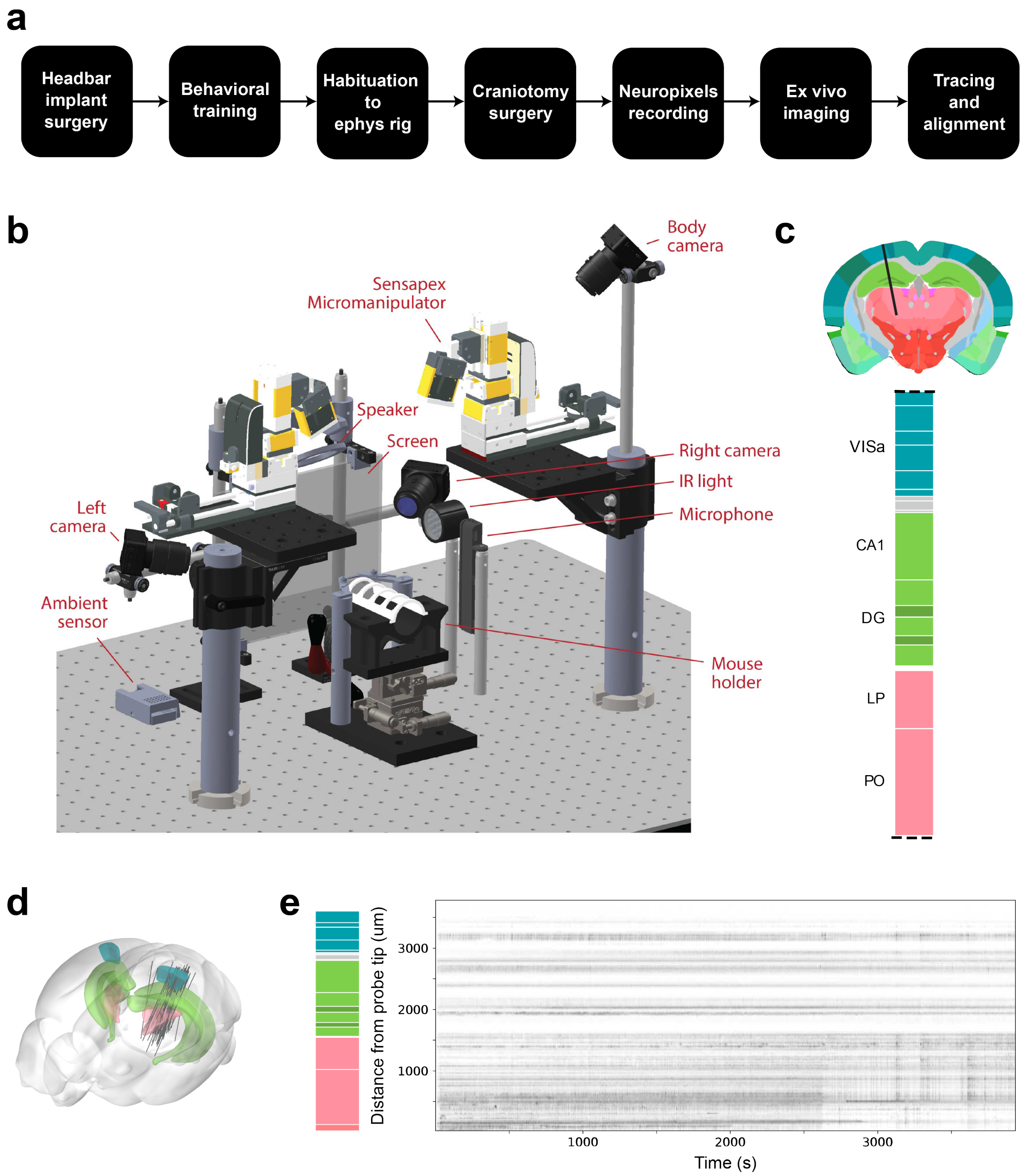


Figure 1. Standardized experimental pipeline and apparatus; location of the repeated site. a, The pipeline for electrophysiology experiments. **b**, Drawing of the experimental apparatus. **c**, Location and brain regions of the repeated site. VISa: Visual Area A; CA1: Hippocampal Field CA1; DG: Dentate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

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

(Figure 2e, bottom, permutation test p-value for all probes p=0.1974), which produced a higher p-value after the application of our exclusion criteria (Figure 2e, bottom, permutation test p-value for exclusion criteria passed probes p=0.0.5499). This demonstrates that after histology reconstruction, the reproducibility of probe placement is enhanced across labs for the brain insertion coordinate with the application of anatomical exclusion criteria.

The final way to assess variability in brain insertions is via 'angle variability,' also calculated from the histological reconstructions. We observed a consistent mean displacement from the planned angle in both medio-lateral (ML) and anterior-posterior (AP) angles (mean difference in angle from planned: 7 degrees, Figure 2f, top). AP angle differences can be explained by the different orientation of the CCF and the stereotaxic coordinate system; ML differences may result from the histological asmples being compressed in the DV direction compared to the CCF. The difference in histology angle to planned probe placement was assessed with permutation testing across labs, and shows a significant difference with our exclusion criteria applied (Figure 2f, bottom, permutation test p-value for all probes p=0.1993; permutation test p-value for exclusion criteria passed probes p=0.0491). This significant result can be explained by the repeated use of the same rig and micromanipulator angle within each laboratory, resulting in reduced variability in probe angle within labs versus across labs.

To determine the extent that anatomical differences drive geometrical variability, we used the micro-manipulator to histology distance at the brain surface and regressed this measurement against animal weight. This easily measured parameter should correlate with mouse brain size and provide a quantifiable predictor of anatomical differences. No such correlation was identified ($R^2 < 0.01$), indicating differences between CCF and mouse brain sizes are not the major cause of variance. We therefore surmise 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.

In conclusion, targeting, geometrical and angle variability revealed lab-to-lab differences that can hinder reproducibility. To control this variability we applied a “targeting" exclusion criterion, which discarded insertions from further analysis when they failed to include sites from at least 3 of the 5 selected areas. This exclusion criterion improved the reproducibility of probe placement at the brain surface, and was used in all subsequent analyses. Probe angle reproducibility was not improved with the exclusion criterion, and this appears to be driven by variance between recording rigs repeatedly used for probe placement within labs. We were unable to identify a prescriptive analysis to predict probe placement accuracy, which may reflect that the major driver of probe placement variance derives from differences in skull landmarks used for establishing the coordinate system, and the underlying brain structures.

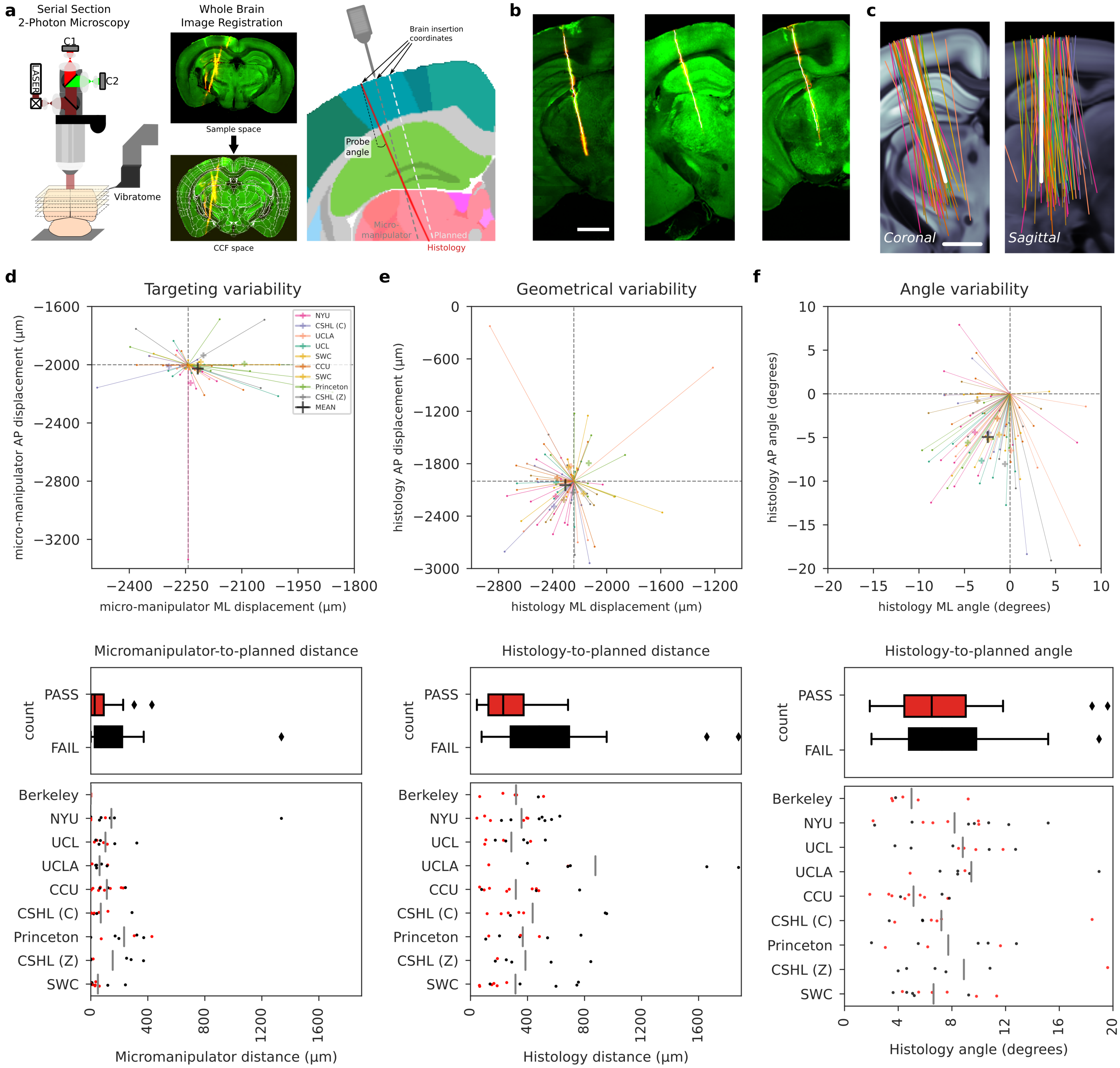


Figure 2. Probe placement shows variance that is reduced with exclusion criteria. a, The histology pipeline for electrode probe track reconstruction and its assessment, consisting of serial section 2-photon microscopy, and manual probe tracing. Three separate trajectories can be defined per probe: the planned trajectory; the micro-manipulator trajectory, based on the experimenter's stereotaxic coordinates; and the histology trajectory, interpolated from tracks traced in the histology data. **b**, Examples of 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 are interpolated from traced probe tracks and plotted as 2D projections in coronal and sagittal planes, tilted along the repeated site trajectory over the allen CCF, color coded by laboratory. Scale bar: 1mm. **d**, Targeting variability of probe placement on the brain surface: scatterplot showing the planned insertion coordinate on the brain surface in ML-AP dimensions, with the position of each subjects' insertion plotted according to the experimenter's stereotaxic coordinates of the probe, color coded by laboratory. Below, boxplots of the distances from planned to stereotaxic coordinates grouped by exclusion criteria, and dotplots by laboratory of stereotaxic-to-planned distances, colour coded by passing our exclusion criteria. **e**, Geometrical variability of probe placement on the brain surface: scatterplot of the planned insertion coordinate on the brain surface in ML-AP dimensions, with the position of each subjects' insertion plotted according to the histology-derived coordinates of the probe, color coded by laboratory. Below, boxplots of the distances from planned to histology coordinates grouped by exclusion criteria, and dotplots by laboratory of histology-to-planned distances, colour coded by passing our exclusion criteria. **f**, Angle variability of probe insertion angle: scatterplot showing the magnitude and direction of the probe angle in ML-AP dimensions, derived from histological reconstructions. Below, boxplots of the relative angles from histology to planned trajectories grouped by exclusion criteria, and dotplots by laboratory of histology-to-planned angle, colour coded by passing our exclusion criteria.

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

Criterion	Definition
Targeting criterion	At least 4 electrode channels in at least 3 of the 5 target brain regions
Behavior criterion	Mouse completed at least 400 trials
Yield criterion	At least 0.1 neurons (that pass single unit criteria*) per electrode channel in each region
Noise criterion	Median action-potential band RMS (AP RMS) less than 40 uV and Median LFP power less than -140 dB
Session number criterion	For analyses that directly compared between labs (permutation tests: Fig 3d-f, Fig 4c, Fig 6), only labs with at least 3 passing sessions per brain region were included.
*Single unit metrics	Each neuron was defined as passing single unit QC if it passed three metrics: a refractory period violation metric, a noise cutoff metric, and a median amplitude threshold. Described further in (***The International Brain Laboratory et al., 2022a***).

Table 1. Quality control criteria for sessions and neurons

Electrophysiological features are reproducible across laboratories

In addition to the "targeting" exclusion criterion, we implemented four other exclusion criteria (see Table 1). We recorded a total of 74 sessions targeted at our planned repeated site (Figure 3a). Of these, 13 were excluded due to unsuccessful data acquisition that could occur from session interruptions (e.g. power outage). Three recordings did not pass our targeting criterion (at least 5 electrode channels in at least 3 of the target brain regions). Six did not pass our behavior criterion (at least 400 trials completed). Nine did not pass our criteria for low yield recordings. Finally, three recordings did not pass our criterion for noise or other electrical artifacts. In subsequent figures, only recordings that passed these quality control criteria were included. In analyses that directly compared across labs (permutation tests; Fig 3d-f, 4c, 5d, 6), only labs which performed three or more successful sessions were included. Furthermore, single units had to pass three quality control metrics to be included in single unit analyses (***The International Brain Laboratory et al., 2022a***). When plotting all recordings, including those that failed to meet quality control criteria, one can observe that discarded sessions were often clear outliers (Figure 3b-c, supp. 1). Overall, we analyzed data recorded from the 40 remaining sessions recorded in 9 labs to determine the reproducibility of our electrophysiological recordings.

We set out to answer the question whether electrophysiological features, such as firing rates and LFP power, were reproducible across laboratories. In other words, is there consistent variation across laboratories in these features that is larger than expected by chance? We first visualized LFP power, a feature used by experimenters to guide the alignment of the probe position to brain regions, for all the repeated site recordings (Figure 3b). The dentate gyrus (DG) is characterized by high power spectral density of the LFP (***Penttonen et al., 1997***; ***Bragin et al., 1995***; ***Senzai and Buzsáki, 2017***) and this feature was used to guide physiology-to-histology alignment of probe positions (Figure 3 supplementary 2). By plotting the LFP power of all recordings along the length of the probe side-by-side, aligned to the boundary between the DG and thalamus, we confirmed that this band of elevated LFP power was clearly visible in all recordings at the same depth. 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 3b).

The reproducibility of electrophysiological features over laboratories was investigated using permutation testing. The tested features included neuronal yield, firing rate, spike amplitude, LFP power, and action-potential band RMS (AP RMS). For each feature and each brain region, the within-lab and across-lab means were calculated (example in Figure 3c). If the electrophysiological feature is reproducible across laboratories, there should be a small deviation between the mean over animals within a lab and the mean over all the lab means. To investigate whether the deviation was significantly larger than expected by chance, we performed permutation testing in which the lab labels were shuffled and a p-value was calculated by comparing the actual deviation from the shuffled null-distribution. Because a test is performed per region-metric pair, the p-values were corrected for multiple testing using the Benjamini-Hochberg procedure (***Seabold and Perktold, 2010***; ***Benjamini and Hochberg, 1995***). We found that all electrophysiological features were reproducible across laboratories for all regions studied.

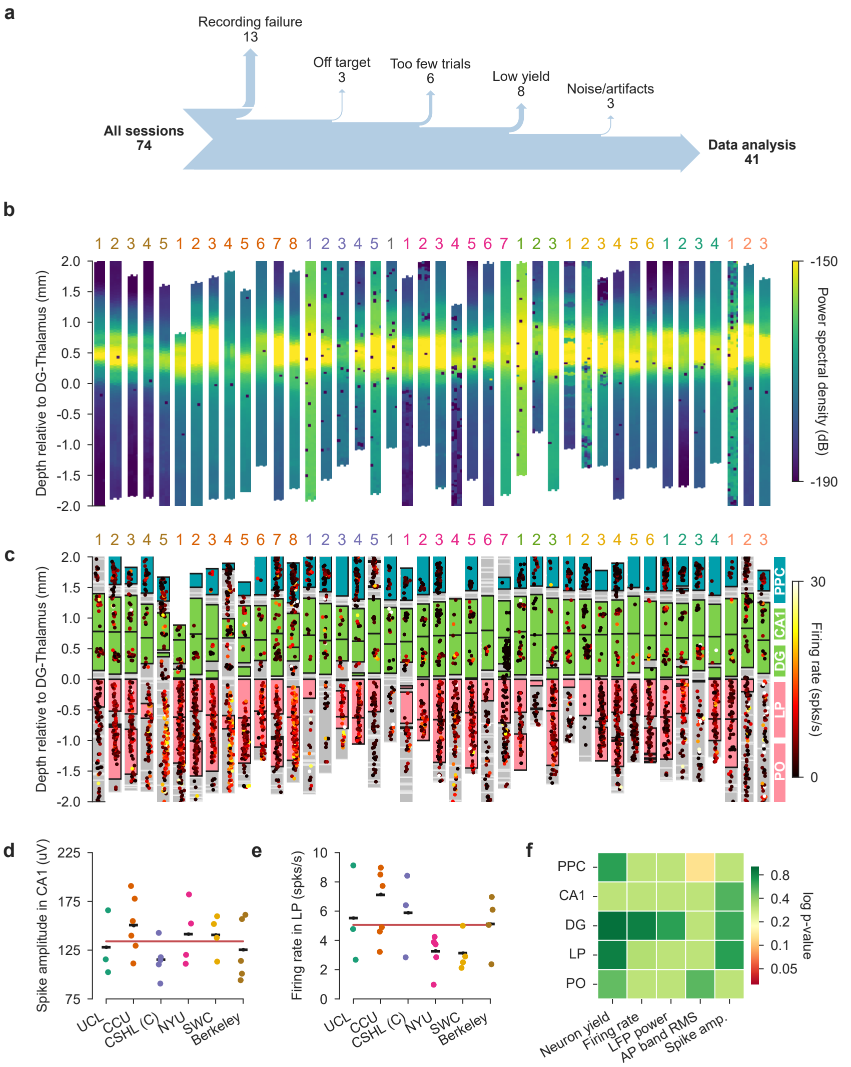


Figure 3. Electrophysiological features are reproducible across laboratories. **a**, Number of experimental sessions recorded; number of sessions used in analysis due to exclusion criteria. **b**, Power spectral density between 20 and 80 Hz of each channel of each probe insertion (vertical columns) shows reproducible alignment of electrophysiological features to histology. Insertions are aligned to the boundary between the dentate gyrus and the thalamus. **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; if no block is plotted the neurons are in a region that is not one of the targets. Dots are neurons, colors indicate firing rate, displacement along the x-axis indicates spike amplitude. **d**,**e**, Examples of permutation testing to determine whether the deviation of lab means (short black lines) from the mean across labs (red line) was larger than expected by chance. For each region, only laboratories that had three or more recordings in that region were included in the permutation testing. Here the median spike amplitude in CA1 and median firing rate in LP is plotted per lab. A p-value was determined by shuffling the lab labels 10,000 times. 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.**f**, P-values for five electrophysiological metrics, computed separately for all target regions. P-values are plotted on a log-scale to visually emphasize values close to significance.

Figure 3–Figure supplement 1. Electrophysiological features of *all* recordings, including recordings that failed quality criteria.

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

Task-driven activity of brain regions is reproducible across laboratories

Concerns about reproducibility include not only basic electrophysiological properties, but also modulation of firing rates by task variables. To address this, we analysed the reproducibility of the relationship between neural activity and task variables across laboratories. In particular, we were interested in whether the brain regions targeted here have comparable neural responses to task events, such as stimulus onset, movement onset, and reward delivery. An inspection of individual neurons revealed clear modulation by, for instance, the onset of movement (Fig. 4**a**). When considering all neurons within a single region of a given session however, it becomes clear that, while a number of neurons are modulated, there is also a proportion of neurons that do not change their firing in relation to task events (Fig. 4**b**) (***Urai et al., 2022***). Plotting the session-averaged response for each experiment in a given area reveals that despite variability, many key features are reproduced, such as the general response time course and timing (Fig. 4**c**; also Fig. 6**d**).

Having observed that many individual neurons are modulated during the task, we then wanted to compare how the proportion of modulated neurons differed across labs. This is especially important, as we are often interested in determining which regions are involved in the neural computations underlying task performance. Therefore, within each brain region, we compared the proportion of the neural population that was sensitive to specific elements of the task. Using Wilcoxon sign-rank tests and Wilcoxon rank-sum tests (***Steinmetz et al., 2019***), we used seven tests to identify neurons with significantly modulated firing rates during specific time-periods of the task. The general logic of these tests is displayed in Fig. 5**b** and Fig. 5-supplemental 1. The neurons that were found by these tests showed a clear modulation to the tested events, as expected (Fig. 5**a-b**). For most tests, the proportions of modulated neurons across sessions and across brain regions were quite variable (Fig. 5**c** and Fig. 5-supplemental 1). However, when applying a permutation test as used in our previous analyses, we found no significant differences across labs regarding the proportion of task-modulated units (Fig. 5**d**). We can therefore conclude that task-modulated activity is reproducible across labs.

To further investigate neuronal task-modulation, we also measured the Fano Factor of single units. The Fano Factor is a useful measurement of firing rate variability and is defined as the spike count variance over trials divided by spike count mean. The Fano Factor enables the comparison of the fidelity of signals across neurons and regions, despite differences in firing rates (***Tolhurst et al., 1983***). Further, the temporal dynamics of the Fano Factor can be informative about underlying neural computations (***Churchland et al., 2010***, ***2011***). We calculated the Fano Factor using a sliding window over each trial. In most brain regions, the Fano Factor, averaged over all neurons, decreased around the time of movement onset (Fig. 7-supplemental 4, left column). Based on the Fano Factor time course, we selected the period between 40-200 ms after movement onset (for correct trials with full-contrast stimuli on the right side) to calculate an average Fano Factor per neuron and quantify differences in Fano Factor across labs. While Fano Factor values varied between neurons and across sessions, we found no difference across labs after applying a permutation test (Fig. 5**d**). This argues that the decrease in neural variability around the time of movements is reproducible and is present not only in cortical structures, as previously reported (***Churchland et al., 2010***), but is also reliably present in the hippocampus and thalamus.

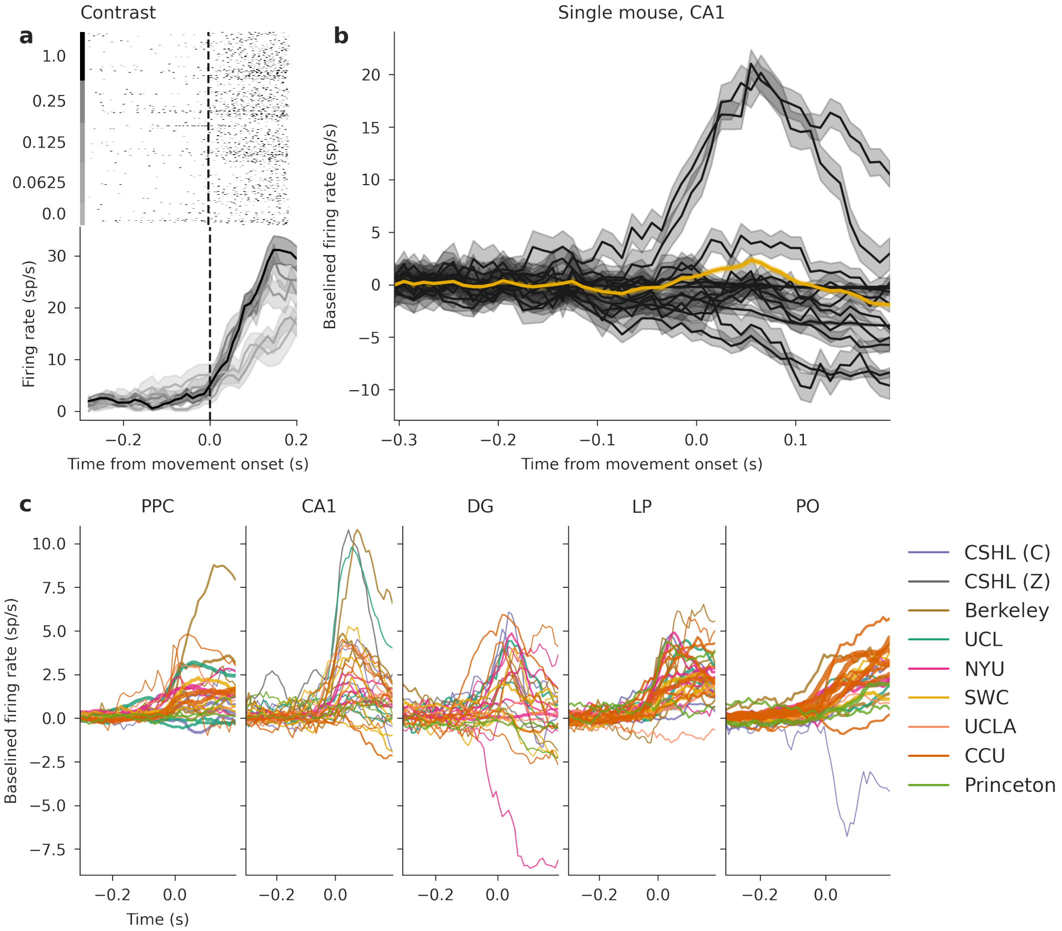


Figure 4. Neural activity is modulated during decision-making in 5 neural structures. **(a)** Neural activity in relation to movement onset towards the left for different contrasts, raster plot (top), peristimulus time histogram (bottom). **(b)** Peri-event time histograms (PETHs) for correct left choices of all neurons from CA1 of a single mouse, aligned to movement onset. 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, coloured by the lab from which the recording originates. **(c)** Average PETHs for correct left choices across regions within individual mice (same as thick line in (b)). Line thickness indicates how many neurons went into the average (min=4, max=86). (As we do not compare across labs, we do not subset to labs with sufficient recordings here).

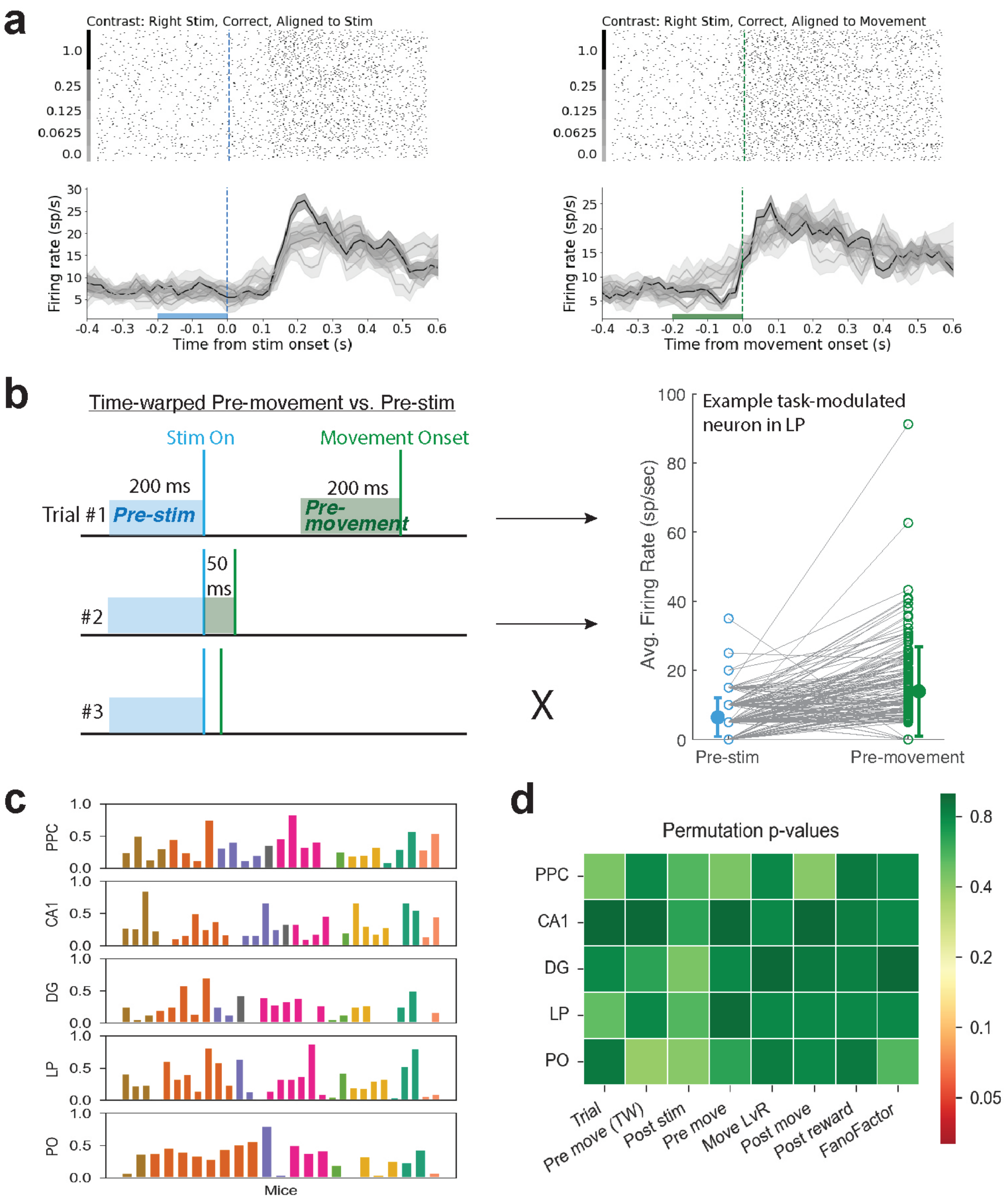


Figure 5. Task-modulated neurons are not significantly different between laboratories. (a) Raster plots and firing rate time courses of an example neuron in LP, aligned to either stimulus onset or movement onset; plotted only for right visual stimuli and correct movements. (The firing rates are calculated using a sliding window and are causal, such that each time point includes a 40 ms window prior to the indicated point.) **(b)** Schematic of the time-warped (TW) pre-movement vs. pre-stimulus test for finding task-modulated neurons (*left*), where the firing rate prior to movement onset is compared against the firing rate during 200 ms before the stimulus. This is only calculated for trials where the time between pre-movement time and stimulus is at least 50 ms (third example trial is excluded). Also, the pre-movement time is considered only up to 200 ms prior to the movement onset, i.e., the pre-movement period can range anywhere from 50 ms to 200 ms prior to the onset of the stimulus (resulting in continuous firing rates in the right panel), unlike the pre-stimulus period which is always set to 200 ms (thus, firing rates in the right panel change in increments of 5 sp/sec). (*right*) The change in firing rate of the example neuron in **a**, which is considered a task-modulated neuron using the TW pre-movement test; each gray line indicates one trial. Mean pre-stimulus and pre-movement firing rates across all trials are shown with filled circles (error bar: standard deviation). **(c)** Proportion of task-modulated neurons for each mouse in each of the five brain regions using the TW pre-movement test. Each column or color indicate, in order, a different recording session or lab. (Note that there is no correspondence here between columns across different brain regions.) **(d)** Permutation test results comparing across-lab variation in the proportion of task-modulated neurons found using each of the seven tests examined (the TW pre-movement test in **b** and **c** and six other tests described in Fig 5-Figure Supplement 1), as well as variation in the neuronal Fano Factors. All task-modulated comparisons were performed for correct trials with non-zero contrast stimuli.

Figure 5–Figure supplement 1. Proportion of task-modulated neurons, defined by six additional tests, across mice, labs, and brain regions.

Principal component embedding analysis reveals little functional separation between labs

In the previous section, we tested specific hypotheses about modulations in task-driven activity at different times within the behavioral trial. We wondered if our conclusions about reproducibility would remain consistent if we perform comparisons across labs and brain regions at the level of the trial-averaged firing rate vectors computed over the entire trial.

The first step is to choose a summary of each cell's neural activity that can be directly compared across experimental sessions and labs. The peri-event time histogram (PETH) is one such summary that is commonly used. The PETH depends on the event used to align trials, and also discards information about behavioral variability across trials. To retain more of this information, we coarsely split trials into two sets, one with fast reaction times ($< 0.15\,\mathrm{s}$) and one with slower reaction times ($> 0.15\,\mathrm{s}$). Then we computed PETHs within each of these subsets and concatenated the resulting vectors to obtain a more informative summary of each cell's average activity within these different types of behavioral trials. (The results described below did not depend strongly on the details of the trial-splitting we chose; for example, splitting trials by "left" vs "right" behavioral choice led to similar results.) See Figure 6**a** for two example cells' PETHs, showing only the PETH obtained by averaging fast reaction time trials.

Next, we project these high-dimensional summary vectors into a low-dimensional "embedding" space that captures the variability of the neuronal population but at the same time allows for easy visualization and further analysis. We found that a simple principal component analysis (PCA) provided a useful embedding. 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). Figure 6**a** shows two cells and the corresponding two-dimensional PCA approximation, with one high-accuracy reconstruction example and one low-accuracy example shown here. Figure 6**b** displays the goodness of this PCA approximation over the full population as a function of the number of PCs used, showing that the PETHs of the majority of cells can be well reconstructed even with just 2 PCs.

Now we have obtained a simple two-dimensional summary of each cell's activity that we can visualize easily; see Figure 6**c**. This simple embedding is already sufficiently powerful to distinguish different brain regions: in Figure 6**c** we have colored cells by region, and we see that e.g. regions PO and CA1 show displaced clusters, illustrating clear regional differences in cell activities. These per-region differences are also visible in the region-averaged PETHs (Figure 6**d**). We quantified this separation via a permutation test, computing the sum across each region's distance between its mean embedded activity and the mean across all regions and comparing that to the null distribution of values obtained in the same way after shuffling the region labels. The p-value is < 0.0001, indicating a significant difference between regional PCA-reduced PETHs.

To test for activity differences between labs, we subdivided the embedded point clouds (Figure 6**c**) by lab (Figure 6**e** and supp. Figure 1). The standard deviation of these activity point clouds show large overlap across most labs, indicating similar activity. For each region separately, we determined whether the sum across each lab's distance between its mean embedded activity and the mean across all labs is significantly different, using the same permutation test as described in the previous paragraph, this time shuffling lab labels. We obtain one false discovery rate corrected p-value for this lab-permutation test per region - PO 0.706, LP 0.065, DG 0.706, CA1 0.168, PPC $p < 0.0001$ - finding that for all regions except PPC the sum of mean lab embedded activities is not significantly different than the mean over all labs. We thus see that embedded activity differs clearly across regions but much less so across labs.

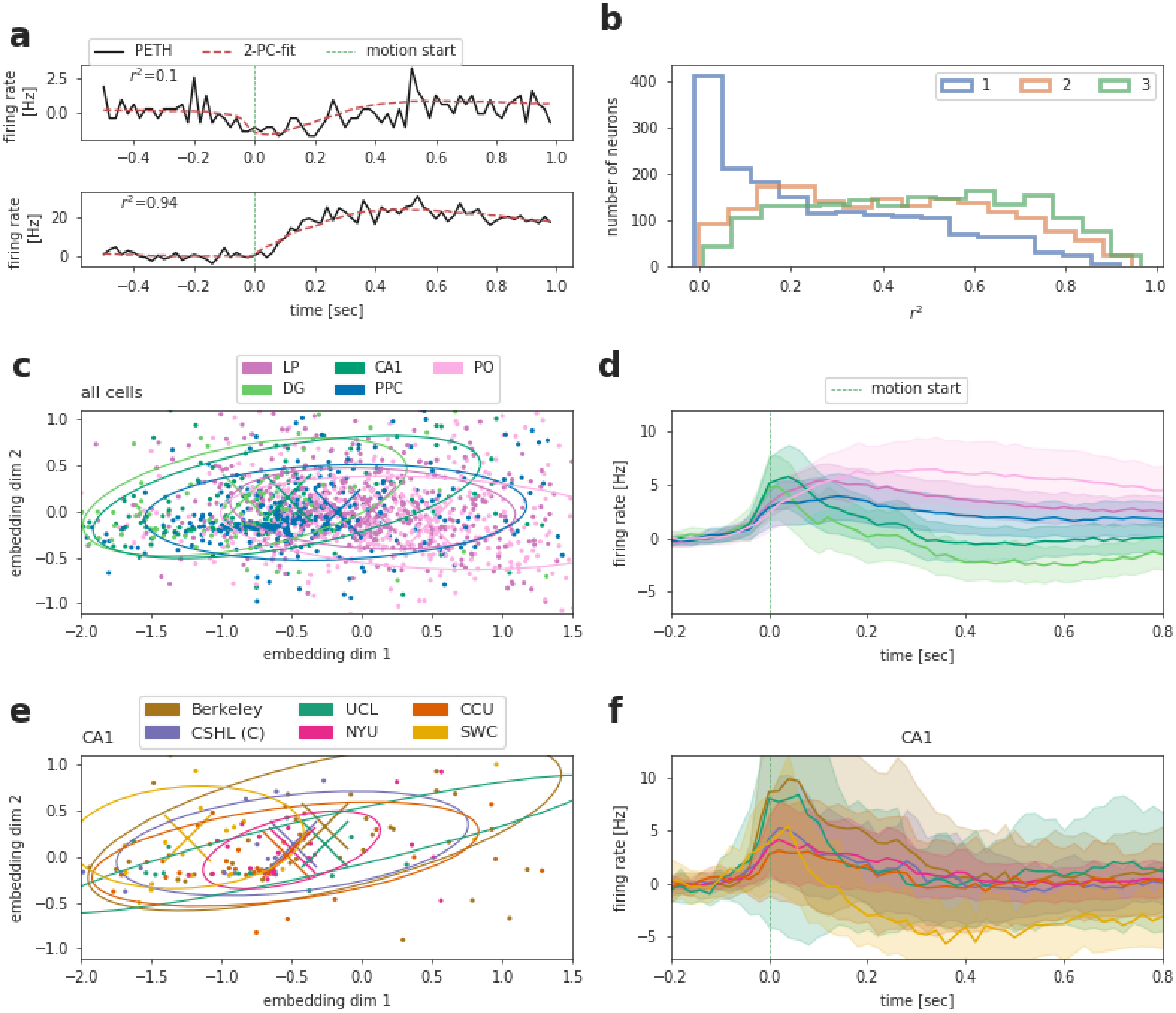


Figure 6. Principal component embedding of peri-event time histograms separates cells from different brain regions but not cells from different labs. **(a)** Two example cells' PETHs in black and 2-PC-based reconstruction; poor (top), good (bottom) fit with goodness of fit r^2 indicated on top. **(b)** Histograms of reconstruction goodness of fit across all cells based on reconstruction by 1-3 PCs. With only the first 2 PCs most PETHs are well approximated, justifying the subsequent two-dimensional embedding analysis. **(c)** Two-dimensional embedding of PETHs of all cells, colored by region (each dot corresponds to a single cell). X's and ellipses indicate the mean and standard deviation for each region. **(d)** Mean firing rates of all cells in each of the studied regions. As in the 2D embedding, mean values for PO and CA1 clearly separate. (Error bars are standard deviation across cells divided by square root of number or recordings per region). **(e)** Embedded activity of CA1 neurons plotted separately for each lab (colors). **(f)** Mean activity for all labs in CA1 (color conventions the same as in **(e)**. See supp. Figure 1 for the other regions. (Error bars are standard deviation across cells divided by square root of number of recordings per lab). Note that only 6 labs are included in this analysis, as we only include labs that have at least 3 recordings per region (see exclusion criterion Table 1).

Figure 6–Figure supplement 1. Regional 2-PC embedding and average PETH per lab

Differences in neuronal spatial position and spike characteristics are a minor source of variability across sessions

While we found little variability between laboratories in terms of electrophysiological features and task variables, we observed large variability between recording sessions and mice (Fig. 3, Fig. 5, and Fig. 5-supplemental 1). Since the spatial position of the Neuropixels probe was variable between sessions (Fig. 2), we examined variability in targeting as a potential source of differences in neuronal activity for each of the five repeated site brain regions. We also considered single-unit spike waveform characteristics as a source of variability. In the next section, we examine other potential sources of variability (e.g., mouse movements).

To investigate variability in session-averaged firing rates, we identified neurons which had firing rates different from the majority of neurons within each brain region (absolute deviation from the median firing rate being >15% of the firing rate range). These outlier neurons, which mostly turned out to be high-firing (except in PO), were compared against regular 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 (Fig. 7a), indeed spanned a wide space within that area. Interestingly, in areas other than DG, the highest firing neurons were not entirely uniformly distributed in space. For instance, in LP, high firing neurons tend to be positioned more laterally and centered on the anterior-posterior axis (Fig. 7b). In PPC and PO, the spatial position of neurons, but not differences in spike characteristics, contributes to differences in session-averaged firing rates (Fig. 7-supplemental 1b and 3c). In contrast, high-firing LP, CA1, and DG neurons have different spike characteristics compared to other neurons in their respective regions (7b and Fig. 7-supplemental 2b and 3a).

To quantify the amount of variability in average firing rates 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) as the inputs. We found similar results: In PPC, z position, or neuron depth, explained part of the variance (had a significant weight); in CA1 and DG, spike amplitude, not spatial position, explained part of the variance; in LP, x and y positions as well as spike amplitude explained some of the variance; in PO, x and y position captured more variance than the other features. In LP, where the most amount of variability can be explained by this regression model, these features account for a total of ~12% of the firing variability. In PPC, CA1, DG, and PO, they account for approximately 3%, 6%, 6%, and 5% of the variability, respectively.

Next, we examined whether neuronal spatial position and spike features contributed to variability in task-modulated activity. We found that all brain regions, except CA1, 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 seven tests in Fig. 5d). For instance, task-modulated LP neurons defined by the time-warped pre-movement test, were positioned more ventrally and centered along the anterior-posterior axis (Fig. 7c), while task-modulated LP neurons defined by the left versus right pre-movement test, tended to be more ventral (Fig. 7d). Other brain regions had less spatial differences than LP (Fig. 7- supplemental 1, 2, 3). Spike characteristics were significantly different between task-modulated and non-modulated neurons only for some tests and only in PPC, DG, and PO (Fig. 6-supplemental 1c-d and 3)b-d. On the other hand, the task-aligned Fano Factor of neurons did not have any differences in spatial position except for in PPC, where lower Fano Factors (<1) tended to be located ventrally (Fig. 7- supplemental 4a). Spike characteristics of neurons with lower vs. higher Fano Factors were only different in the LP and PO (Fig. 7- supplemental 4). Lastly, we trained a linear regression model to predict the 2D embedding of PETHs of each cell shown in Fig 6c from the x, y, z coordinates and found that spatial position contains little information ($r^2 \sim 4\%$) about the embedded PETHs of cells.

In summary, our results suggest that spatial position is a small contributor to variability for session-averaged firing rates in all brain regions except DG, and to a lesser degree for task-modulated neuronal activity in all brain regions except CA1. In all regions, spike characteristics also have a minor contribution to the observed variability. Since, overall, the contributions of spatial position and spike features were small, despite being significant, we examine other sources of variability in the next section.

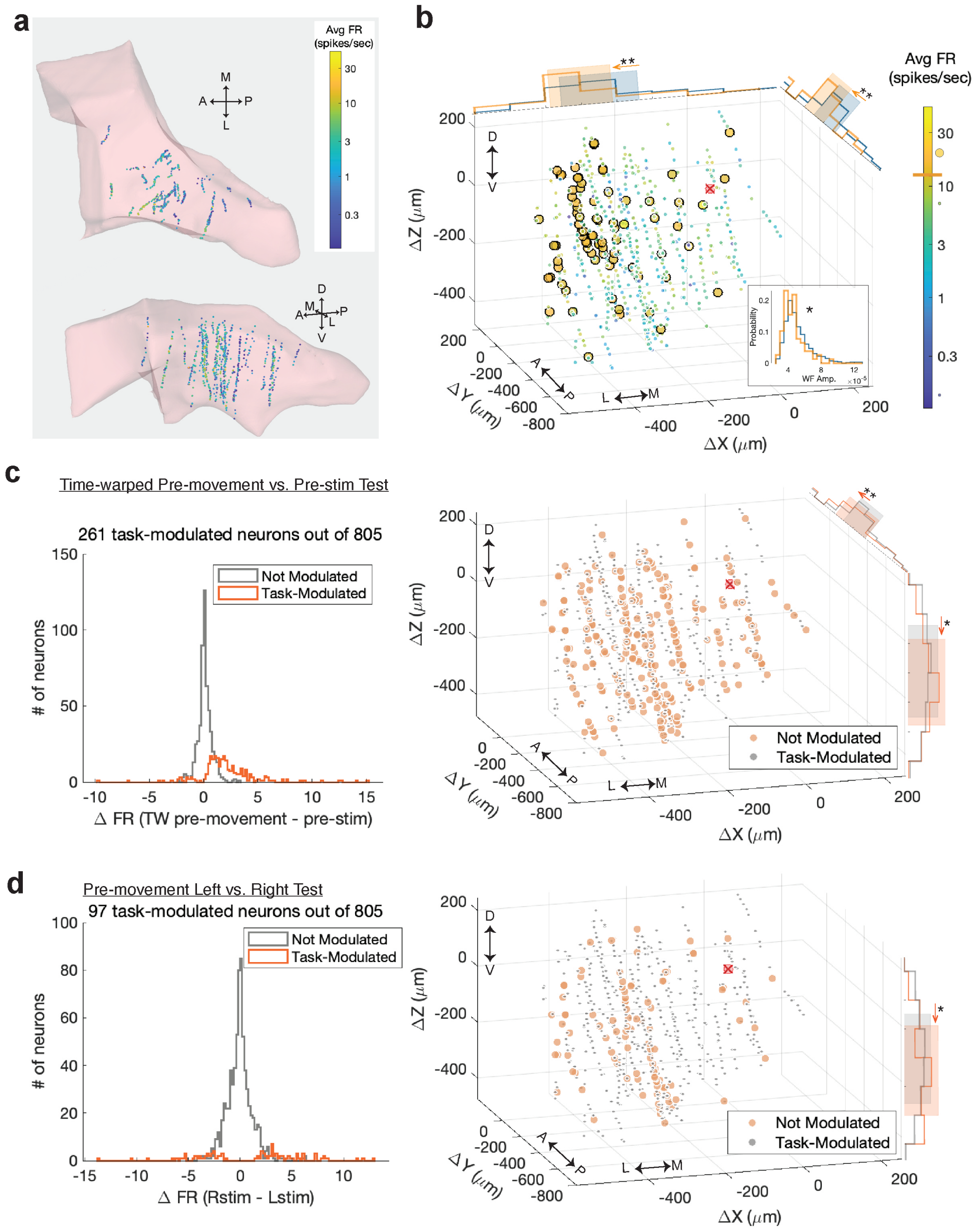


Figure 7. High-firing and task-modulated LP neurons have slightly different spatial positions than other LP neurons, potentially contributing to variability between sessions. (a) Spatial positions of recorded neurons in LP, color-coded with their firing rates averaged over the recording session. **(b)** Spatial positions of LP neurons plotted as distance from the planned target center of mass, indicated with the red x. (To enable visualization of overlapping data points, jitter was added to the unit locations.) Larger circles indicate outlier neurons (defined by a normalized firing rate deviation > 15%, resulting in a threshold of 12 sp/sec for LP, shown on the colorbar; here, 78 out of 805 neurons were outliers). Only histograms of the spatial positions and waveform features that were significantly different between the outlier (yellow) and regular (blue) units are shown (two-sample Kolmogorov-Smirnov test with Bonferroni correction for multiple comparisons; * and ** indicate corrected p-values of <0.05 and <0.01, in order). Shaded areas indicate the area between 20th and 80th percentiles of the neurons' locations. **(c)** (Left) Histogram of firing rate changes during the pre-movement period from the pre-stimulus period (using the time-warped test, Fig. 5b-c) for task-modulated (orange) and non-modulated (gray) neurons. (Right) Spatial positions of task-modulated and non-modulated LP neurons, with histograms of significant features (here, y and z positions) shown. **(d)** Same as **c** but using the pre-movement left vs. right test to identify task-modulated units.

Figure 7–Figure supplement 1. High-firing and task-modulated PPC neurons.

Figure 7–Figure supplement 2. High-firing and task-modulated CA1 neurons.

Figure 7–Figure supplement 3. High-firing and task-modulated DG and PO neurons.

Figure 7–Figure supplement 4. Time-course and spatial position of neuronal Fano Factors.

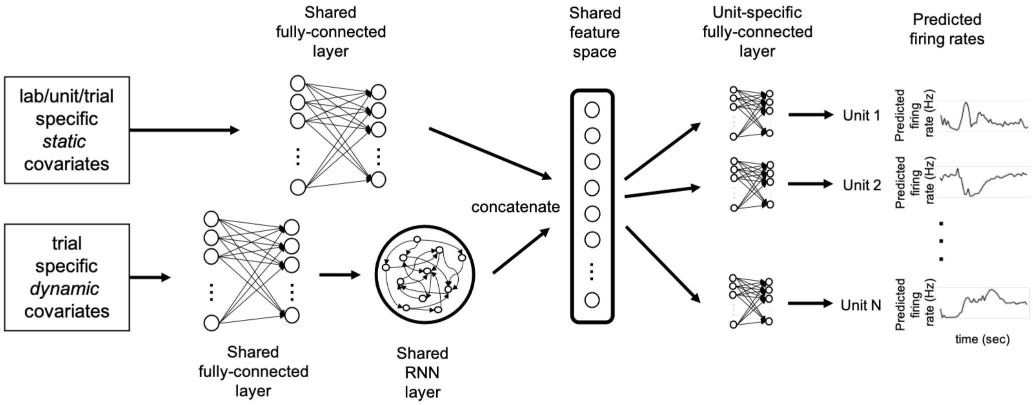

lab/unit/trial specific *static* covariates
trial specific *dynamic* covariates
Shared fully-connected layer
Shared fully-connected layer
Shared RNN layer
concatenate
Shared feature space
Unit-specific fully-connected layer
Unit 1
Unit 2
Unit N
Predicted firing rates
Predicted firing rate (Hz)
time (sec)

Figure 8. Schematic of the multi-task neural network model architecture: We adapt a multi-task neural network approach for unit-specific firing rate prediction. The model takes in a set of covariates, and outputs time-varying firing rates for each neuron for each trial. The covariates include the lab ID, 3-D unit location, and trial event times (e.g., stimulus onset); see Table 2 for a full list. The initial embedding layer of the network is shared across all units, and serves to learn a useful (nonlinear) shared set of features that all the individual units can regress onto for their predictions.

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

As discussed above, variability in neural activity between labs or between sessions can be due to many factors. These include differences in behavior between animals, differences in probe placement between sessions, and uncontrolled differences in experimental setups between labs. How can we quantify and distinguish these different sources of variability? Simple linear regression models or generalized linear models (GLMs) are likely too inflexible to capture the nonlinear contributions that many of these variables, including lab IDs and spatial positions of neurons, might make to neural activity. On the other hand, fitting a different nonlinear regression model (involving many covariates) individually to each recorded unit 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 8) that takes as input a set of covariates (including the lab ID, the unit'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 shared set of nonlinear features (shared over all recorded units) and fits a Poisson regression model on this shared feature space for each unit. (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 variables. In particular, previous reduced-rank regression approaches (***Kobak et al., 2016***; ***Izenman, 1975***) can be seen as a special case of the multi-task neural network, with a single hidden layer and linear weights in each layer.

Figure 9a shows model predictions on held-out trials for a single CA1 unit. We plot the observed and predicted peri-event time histograms and raster plots, split into left vs. right trials. As a visual overview of which behavioral covariates are highly correlated with this cell's activity on each trial, various behavioral covariates that are input into the MTNN are shown in Figure 9b. Overall, the MTNN approach accurately predicts the observed firing rates. When the MTNN and GLMs are trained on a reduced set of covariates, consisting of stimulus onset timing, stimulus contrast and side, feedback type and timing, first movement onset timing, wheel velocity, and mouse's prior, the

Variable Name	Type	Group	Note
Lab ID	Categorical / Static		
Session ID	Categorical / Static		
Unit 3D spatial position	Real / Static	Electrophysiological	In standardized CCF coordinates
Unit amplitude	Real / Static	Electrophysiological	Template amplitude
Unit waveform width	Real / Static	Electrophysiological	Template width
Paw speed	Real / Dynamic	Behavioral	Inferred from DLC
Nose speed	Real / Dynamic	Behavioral	Inferred from DLC
Pupil diameter	Real / Dynamic	Behavioral	Inferred from DLC
Motion energy	Real / Dynamic	Behavioral	
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	Behavioral	
Mouse Prior	Real / Static		Mouse's prior belief
Last Mouse Prior	Real / Static		Mouse's prior belief in previous trial
Lick	Binary / Dynamic	Behavioral	
Decision Strategy	Real / Static		Decision-making strategy (***Ashwood et al., 2022***)
Brain region	Categorical / Static	Electrophysiological	5 repeated site regions

Table 2. List of covariates input to the multi-task neural network. See Appendix for additional details.

MTNN and GLMs perform similarly on predicting the firing rates of held-out test trials. Furthermore, the MTNN trained on the full set of covariates in Table 2 outperforms the MTNN and GLMs trained on the reduced covariate set (See Figure 9 supplemental 2).

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 behavioral) 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 10 supplemental 1).

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 10. According to the single-covariate analysis (Figure 10a), face motion energy (derived from behavioral video), wheel velocity, and some task variables (e.g., stimulus information and first movement onset timing) can individually explain about 5-10% of variance of the units on average. The leave-one-out analysis (Figure 10b left) shows that most covariates have low unique contribution to the predictive power. This is because many variables 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 9b). According to the leave-group-out

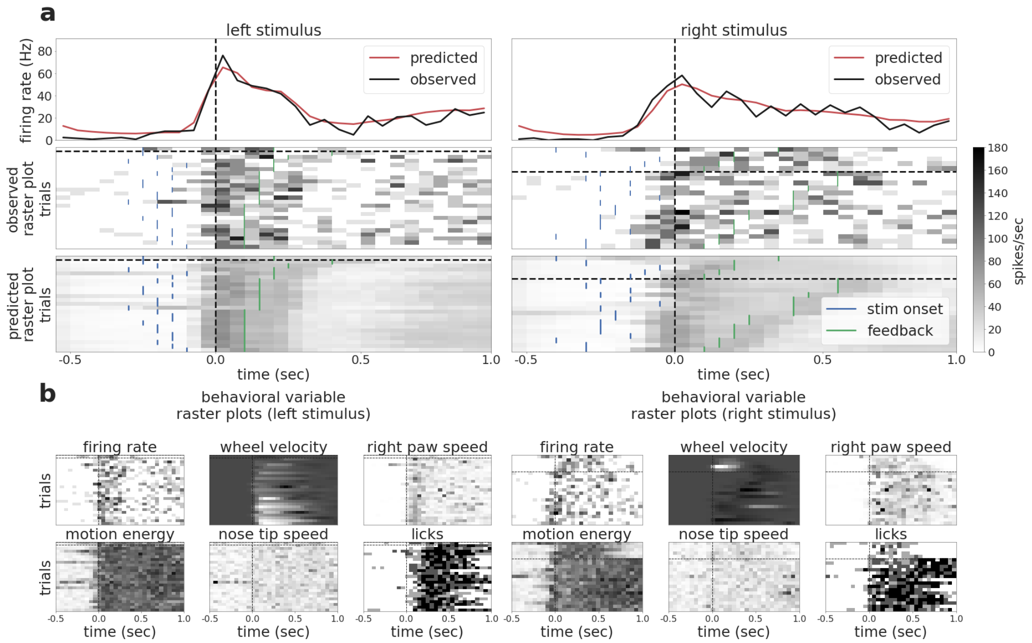


Figure 9. The MTNN model accurately estimates firing rates on held-out test trials from a CA1 neuron: **(a)** MTNN model estimates of firing rates (50 ms bin size) of a CA1 neuron from an example subject during held-out test trials. The trials are split into those that had stimulus on the left/right and are aligned to the first movement onset time (vertical dashed lines). We plot the observed and predicted peri-event time histograms **(1st row)** and the observed and predicted raster plots **(2nd and 3rd rows)**. The blue ticks in the raster plots indicate stimulus onset, and the green ticks indicate feedback times. The black horizontal dashed line separates the incorrect/correct trials (i.e., the trials above the dashed line are incorrect trials), and the trials are ordered by reaction time. The trained model does well in predicting the (normalized) firing rates. The MTNN prediction quality measured in R^2 is 0.32 on held-out test trials and 0.90 on PETHs of held-out test trials. **(b)** We plot the raster plots of behavioral variables (wheel velocity, paw speed, motion energy, nose speed, and licks), ordering the trials in the same manner as in **(a)**. We see that the MTNN firing rate predictions are modulated synchronously with the behavioral variables.

Figure 9–Figure supplement 1. Scatter plot of MTNN prediction quality (R^2) vs. mean firing rate (spikes/sec)

Figure 9–Figure supplement 2. MTNN slightly outperforms GLMs on predicting the firing rates of held-out test trials.

Figure 9–Figure supplement 3. PETHs and MTNN predictions for held-out test trials

analysis, the behavioral 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.

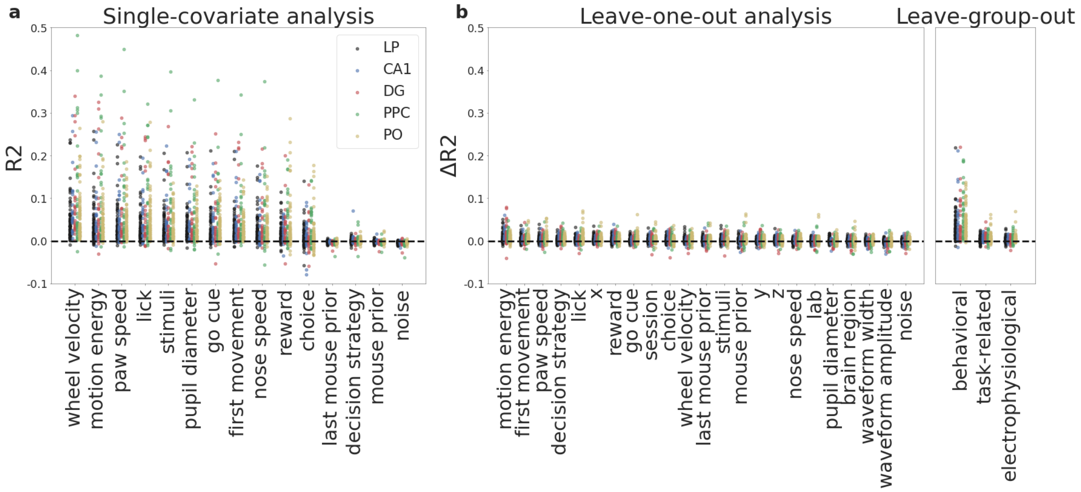


Figure 10. Single-covariate, leave-one-out, and leave-group-out analyses show the contribution of each (group of) covariate(s) to the model. Lab and session IDs have low contributions to the model. (a) Single-covariate analysis, colored by the brain region. Each dot corresponds to a single cell in each plot. **(b)** Leave-one-out analysis, colored by the brain region. The analyses are run on 246 responsive units across 20 sessions. The leave-one-out analysis shows the unique contribution of each covariate to the model, and the single-covariate analysis shows the upper limit of the contribution of each covariate to the model. The leave-group-out analysis shows how groups of electrophysiological, task-related, and behavioral covariates contribute to the model. The leave-one-out analysis shows that lab/session IDs have low effect sizes on average, indicating that within-lab 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 (i.e., lab and session IDs, unit’s 3D spatial location) are left out from the single-covariate analysis.

Figure 10–Figure supplement 1. MTNN prediction quality on the data simulated from GLMs is comparable to the GLMs’ prediction quality. The effect sizes computed by the MTNN leave-one-out analysis are similar to the effect sizes computed by the GLMs’ leave-one-out analysis

Figure 10–Figure supplement 2. Pairwise scatterplots of MTNN single-covariate effect sizes.

Discussion

We set out to test whether electrophysiological responses, notoriously variable across labs, could be reproducible across geographically separated laboratories after appropriate standardization of experiments, data processing, and analyses. After applying stringent behavioral, histological, and electrophysiological quality-control criteria, we found that electrophysiological features such as neuronal yield, firing rate, and normalized LFP power were reproducible across laboratories; their within-lab averages did not significantly deviate from the mean across labs. Similarly, the proportion of cells whose responses are tuned to behaviorally-relevant task events is reproducible across labs. Finally, a multi-task neural network approach can predict the firing rates of different units across sessions, and again, the within-lab random effects estimated by this model were comparable to between-lab random effects. Taken together, our results suggest that careful standardization can lead to reproducible electrophysiological results across laboratories.

Reproducibility in our electrophysiology studies depended on rigorous metrics of quality. We found that it was necessary to exclude a significant fraction of datasets to reach a desired level of reproducibility. Quality control was enforced for diverse aspects of the experiments, including histology, behavior, targeting, neuronal yield, and the total number of completed sessions. Among these measures, recordings with high noise and low neuronal yield were significantly represented in sessions that were excluded (40/74 sessions). A number of issues contributed here, including artifacts present in the recordings, inadequate grounding, and a decline in craniotomy health; all of these can potentially be improved with experimenter experience. Sub-standard behavior (for instance, too few trials in a session) led to the elimination of another substantial fraction of datasets. Trial counts are likely to be highly variable across labs, as there is currently no agreed upon standard for what constitutes suitable behavior for an electrophysiology experiment. This has already been shown to cause differences in the internal states visited by animals as they make decisions (***Ashwood et al., 2022***).

These observations suggest that future experiments might enjoy greater reproducibility if researchers followed, or at least reported, a number of agreed upon criteria, such as those 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***). Our work here suggests the creation of a similar set of standards for electrophysiology and behavioral experiments would be beneficial. These could include expectations for reporting (such as histological information and behavioral trial numbers) as well as suggestions for minimizing variability (e.g., agreed upon standards for the noise level that would exclude a recording).

We found probe targeting to be a large source of variability, driven by micro-manipulator positioning and anatomical discrepancies. The majority of the variance in targeting was due to the probe entry positions at the brain surface, which showed no bias in placement 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 directions; indicating that the Allen Common Coordinate Framework (CCF) (***Wang et al., 2020***) and stereotaxic coordinate systems are slightly offset. Minimizing variance in probe targeting is an important element in increasing reproducibility, as slight deviations in probe entry position and angle can lead to samples from different populations of neurons. Our approach suggests a path forward to minimize these biases: probe angles must be carefully computed from the CCF, as the CCF and stereotaxic coordinate systems do not define the same coronal plane angle. Small differences in probe location may be responsible for other studies arriving at different conclusions, highlighting the need for agreed upon methods for targeting specific areas (***Rajasethupathy et al., 2015***; ***Andrianova et al., 2022***).

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

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

Scientific advances rely on the reproducibility of scientific findings. The current study demonstrates that reproducibility is attainable for large-scale neural recordings during a standardized perceptual detection task across 9 laboratories. We offer several recommendations to increase reproducibility, including (1) standardized protocols for data collection, (2) data processing, and (3) rigorous data quality metrics. Furthermore, we have made improvements in data architecture and processing, now available to the public. Our study provides a framework for the collection and analysis of large neural datasets in a reproducible manner that will play a key role as neuroscience continues to move towards increasingly complex datasets.

Resources

Data access

Please visit https://int-brain-lab.github.io/iblenv/notebooks_external/data_release_repro_ephys.html to access the data used in this article.

Code repository

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

Protocols and pipelines

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

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, and University of California at Berkeley; the University Animal Welfare Committee of New York University; and the Portuguese Veterinary General Board.

Animals

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

Materials and apparatus

For detailed parts lists and installation instructions, see Appendix 1 (***The International Brain Laboratory, 2022a***).

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 occured 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) sampling at 60, 150, 30 Hz respectively (for details see Appendix 1 (***The International Brain Laboratory, 2022a***)). 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 was handled by Bonsai (***Lopes et al., 2015***) and 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

A detailed account of the surgical methods is in Appendix 1 (***The International Brain Laboratory et al., 2021***).

Briefly, mice were anesthetized with isoflurane and head-fixed in a stereotaxic frame. The hair was then removed from their scalp, much of the scalp and 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. The head bar was then placed in one of three stereotactically defined locations and cemented in place. The location of the future craniotomies were measured using a pipette referenced to bregma, and marked on the skull using either a surgical blade or pen. 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

For a detailed protocol on animal training, see Appendix 2 (***The International Brain Laboratory et al., 2021***).

Once the mouse is classified as having learned the biasedChoiceWorld task (criteria 'ready4ephysRig' reached, cf Appendix 2 for definition (***The International Brain Laboratory et al., 2021***)), it is transferred onto the ephys rig.

The mouse is habituated to behave on the ephys rig in a series of steps that do not involve any electrophysiology recording. First, the mouse needs 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 is reached, time delays are introduced prior to the task; these delays would eventually serve to mimic the time it would take to insert electrodes in the brain. The mouse has to maintain performance for 3 subsequent sessions (same criterion as 'ready4ephysRig'), but with a minimum of one session that has a 15 minutes delay and is a mock recording.

Electrophysiological recording using Neuropixels probes

Data acquisition

For details, see Appendix 2 and 3 (***The International Brain Laboratory, 2022b***,c).

Briefly, upon the day of electrophysiological recording, the animal was anaesthetised using isoflurane and surgically prepared. The cement and glue were removed, exposing the skull over both hemispheres. A test was made to check whether the implant could hold liquid, and if successful a grounding pin was implanted. One or two craniotomies (1 × 1 mm) were made over the marked locations. 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 with analgesics subcutaneously, and left to recover in a heating chamber until locomotor and grooming activity were fully recovered.

Once the animal was recovered from the craniotomy, it was fixed in the apparatus. Once a craniotomy was made, up to 4 subsequent recording sessions were made in that same craniotomy. Up to two probes were implanted in the brain on a given session. The probes were labelled with CM-DiI (see Appendix 4 (***The International Brain Laboratory, 2022d***) and (***Liu, 2019***)).

Spike sorting

The spike sorting pipeline used at IBL is described in details in (The International Brain Laboratory et al., 2022a). Briefly, spike sorting was performed using a modified version of the Kilosort 2.5 algorithm (Steinmetz et al., 2021). We found it necessary to improve the original code in several aspects (scalability, reproducibility, and stability, discussed below), and developed an open-source Python port; the code repository is here: (The International Brain Laboratory, 2021b).

Regarding scalability: we found that the original code failed on recording sessions with a large number of detected spikes. Therefore we improved the CPU memory usage of the code to better handle these cases.

Regarding reproducibility: spike sorting algorithms are still in heavy development; we needed to tag and validate code versions and parameter settings internally so we could release the algorithm to our data-processing computers across multiple labs on our own schedule. We also defined a set of integration tests on short (100 seconds) recordings, using hybrid ground-truth datasets (Pachitariu et al., 2016) to validate algorithm changes before new version releases.

Regarding stability: we observed a number of clear artifacts in the raw Neuropixels output ("dead" channels, simultaneous "glitch" artifacts across multiple channels, mis-alignment errors, etc.) that were not handled properly by previous algorithms. We developed new methods to handle each of these artifact types, resulting in significantly more stable sorting outputs. See (The International Brain Laboratory et al., 2022a) for full details.

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. 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 and perfuse-fixed with PBS followed by 4% formaldehyde solution (Thermofisher 28908) in 0.1M PB pH 7.4. 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. For full details, see Appendix 5 (The International Brain Laboratory, 2022e).

For imaging, brains were equilibrated with 50mM PB solution and embedded into 5% agarose gel blocks. The brains were imaged using 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 using StitchIt (Campbell, 2021). Whole brain coronal image stacks were acquired with a resolution of 4.4 x 4.4 x 25.0 μm in XYZ, with a two-photon laser wavelength of 920nm, and power of 35% of 1800mW from the source laser, yielding approximately 150mW at the block face. Serial section microscopy proceeded with 2 z slices taken for each 50μm tissue slice, at a depth of 30μm and 55μm from the tissue surface. Two channels of image data was acquired on two PMTs for green (bandpass filter ET525/50m) and red (bandpass filter ET570lp) fluorescence.

Whole brain images were downsampled to 25μm XYZ pixels 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. For full details, see Appendix 7 (The International Brain Laboratory, 2022g).

Probe track tracing and alignment

Neuropixels probe tracks were manually traced to yield a probe trajectory using Lasagna (***Campbell et al., 2020***), a Python-based image image viewer equipped with a plugin tailored for this task. Traced probe track data was uploaded to an Alyx server (***Rossant et al., 2021***); a database designed for experimental neuroscience laboratories. Neuropixels channels were then manually aligned to anatomical features along the trajectory using electrophysiological landmarks with [ephys alignment tool] (***Faulkner, 2020***) (***Liu et al., 2021***). For full details, see Appendix 6 (***The International Brain Laboratory, 2022f***).

Permutation tests

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 between laboratories: the sum of the absolute differences between laboratory means and overall mean. The null-hypothesis is that there is no difference between the different laboratory means, i.e. the assignment of mice to laboratories is completely random. We constructed the corresponding null-distribution by permuting these assignments between laboratories and mice randomly 10000 times (leaving the relative numbers of mice in laboratories intact) and computing the test statistic on these randomised samples. Given this constructed 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.

Dimensionality reduction of peri-event time histograms via principal component analysis

In Figure 6 we use principal component analysis (PCA) to embed peri-event time histograms (PETHs) into a two-dimensional feature space for visualization and further analysis. Our overall approach is to compute PETHs, split into fast-reaction-time and slow-reaction-time trials, then concatenate these PETH vectors for each cell to obtain an informative summary of each cell's activity. Next we stack these double PETHs from all labs into a single matrix and use PCA to obtain a low-rank approximation of this PETH matrix.

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

Let the stack of these double PETH vectors be Y, being a $N \times 2T$ matrix, where N is the total number of neurons recorded across 5 brain regions and labs. Running principal components analysis (PCA) on Y (singular value decomposition) is used to obtain the low-rank approximation $UV \approx Y$. This provides a simple low-d embedding of each cell: U is $N \times k$, with each row of U representing a k-dimensional embedding of a cell that can be visualized easily across labs and brain regions. V is $k \times 2T$ and corresponds to the k temporal basis functions that PCA learns to best approximate Y. Figure 6(a) shows two cells of Y and the corresponding PCA approximation from UV.

The scatter plots in Figure 6 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 6c.

Video analysis

Some of the behavioral time series used in the neural network analysis are derived from video recordings of the animals. Full details of the video analysis pipeline are here: (***The International Brain Laboratory et al., 2022b***), and the code is available here: (***The International Brain Laboratory, 2021a***).

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

Marker-less tracking of body parts is achieved using Deeplabcut (***Mathis et al., 2018***), a deep-learning-based tool that is used within a fully automated pipeline in IBL to track various body parts such as the paws. The pipeline first detects 3 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:

  `'nose_tip', 'tongue_end_r', 'tongue_end_l'`

- 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.

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, 3d 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 et al., 2022b***) for further details and example quality control images.

Multi-task neural network model to quantify sources of variability

Data preprocessing

For the Multi-task neural network (MTNN) analysis, we used data from 20 sessions recorded in CCU, CSHL (C), SWC, Berkeley, and NYU. We included various covariates in our feature set (e.g. go-cue signals, stimulus/reward type, Deep Lab Cut behavioral outputs). For the “decision strategy" covariate, we used the posterior estimated state probabilities of the 4-state GLM-HMMs trained on the sessions used for the MTNN analysis (***Ashwood et al., 2022***). 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 within 0.4 seconds before the first movement onset time and feedback time is within 0.9 seconds after the first movement onset time. Finally, we selected responsive units whose mean firing rate is greater than 5 spikes/second for further analyses. For sessions with more than 15 responsive units, we randomly sampled 15 units.

Model Architecture

Given a set of covariates in Table 2, the MTNN predicts the target sequence of firing rates from 0.5 seconds before first movement onset to 1 second after, with bin width set to 50 ms (30 time bins). More specifically, a sequence of feature vectors $x_{\text{dynamic}} \in \mathbb{R}^{D_{\text{dynamic}} \times T}$ that include dynamic covariates, such as Deep Lab Cut (DLC) outputs, and wheel velocity, and a feature vector $x_{\text{static}} \in$

$\mathbb{R}^{D_{\text{static}}}$ that includes static covariates, such as the lab ID, unit's 3-D location, are input to the MTNN to compute the prediction $y^{pred} \in \mathbb{R}^{T}$, where D_{static} is the number of static features, D_{dynamic} is the number of dynamic features, and T is the number of time bins. The MTNN has initial layers that are shared by all units, and each unit has its designated final fully-connected layer.

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

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

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

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

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

$$y^{pred}_{t} = f(w^{T}_{(s,u)} \text{concat}(e_{\text{static}}, h^{(forward)}_{t}, h^{(backward)}_{t}) + b_{(s,u)}) \tag{6}$$

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 unit-specific fully-connected layer, indexed by (s, u). Each part of the MTNN architecture can have an arbitrary number of layers. For our analysis, we used two fully-connected shared layers for static covariates (Eqn. (2)) and three-layer bidirectional RNNs for dynamic covariates, with the embedding size set to 64.

Model training

The model was implemented in PyTorch and trained on a single GPU. The training was performed using Stochastic Gradient Descent on the Poisson negative loglikelihood (Poisson NLL) loss with learning rate set to 0.1, momentum set to 0.9, and weight decay set to 10^{-15}. We used a learning rate scheduler such that the learning rate for the i-th epoch is 0.1×0.95^{i}, and the dropout rate was set to 0.2. We also experimented with mean squared error (MSE) loss instead of Poisson NLL loss, and the results were similar. The batch size was set to 512.

The dataset consists of 20 sessions, 246 units and 6878 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 10 supplemental 1, we first trained GLMs on the same set of 246 responsive neural units from 20 sessions used for the analysis in Figure 10, 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 unit. For each simulated unit, we used these kernels of the trained GLM to simulate its firing rates for 350 randomly initialized trials. The random trials were 1.5 seconds long with 50 ms bin width. For all trials, the first movement onset timing was set to 0.5 second after the start of the trial, and the stimulus contrast, side, onset timing and feedback type, timing were randomly sampled. We used wheel velocity traces and mouse's priors from real data for simulation. We finally ran the leave-one-out analyses with GLMs/MTNN on the simulated data and compared the effect sizes estimated by GLMs and MTNN.

Acknowledgments

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

References

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

Ashwood ZC, Roy NA, Stone IR, Urai AE, Churchland AK, Pouget A, Pillow JW. Mice alternate between discrete strategies during perceptual decision-making. Nature Neuroscience. 2022; 25(2):201–212.

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

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

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

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995; 57(1):289–300. https://www.jstor.org/stable/2346101, publisher: [Royal Statistical Society, Wiley].

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Liu L. Painting Neuropixels probes and other silicon probes for electrophysiological recordings. protocolsio. 2019; doi: dx.doi.org/10.17504/protocols.io.wxqffmw.

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

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

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

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

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

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

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

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

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

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

Pachitariu M, Steinmetz NA, Kadir SN, Carandini M, Harris KD. Fast and accurate spike sorting of high-channel count probes with KiloSort. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, editors. *Advances in Neural Information Processing Systems*, vol. 29; 2016. .

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

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

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

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

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

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

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

Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. In: *9th Python in Science Conference*; 2010. .

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

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

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

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

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

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

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

The International Brain Laboratory. Appendix 1: IBL electrophysiological (Ephys Neuropixels) rig setup instructions: Hardware and Software. figshare. 2022; doi: 10.6084/m9.figshare.17307077.

The International Brain Laboratory. Appendix 2: IBL protocol for electrophysiology recording using Neuropixels probe. figshare. 2022; doi: 10.6084/m9.figshare.19697896.

The International Brain Laboratory. Appendix 3: IBL protocol for craniotomy surgery. figshare. 2022; doi: 10.6084/m9.figshare.19697827.

The International Brain Laboratory. Appendix 4: Protocol for labeling the tip of Neuropixels probes. figshare. 2022; doi: 10.6084/m9.figshare.19698130.

The International Brain Laboratory. Appendix 5: IBL protocol for perfusion and shipment of brain sample. figshare. 2022; doi: 10.6084/m9.figshare.19698061.

The International Brain Laboratory. Appendix 6: IBL protocol for registering the electrode location using LASAGNA. figshare. 2022; doi: 10.6084/m9.figshare.19698166.

The International Brain Laboratory. Appendix 7: IBL protocol for mouse brain reconstruction and registration. figshare. 2022; doi: 10.6084/m9.figshare.19698895.

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

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

The International Brain Laboratory, Birman D, Bonacchi N, Buchanan K, Chapuis G, Huntenburg J, Meijer G, Paninski L, Schartner M, Svoboda K, Whiteway M, Wells M, Winter O. Video hardware and software for the International Brain Laboratory. figshare. 2022; doi: 10.6084/m9.figshare.19694452.

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

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

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

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

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

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

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

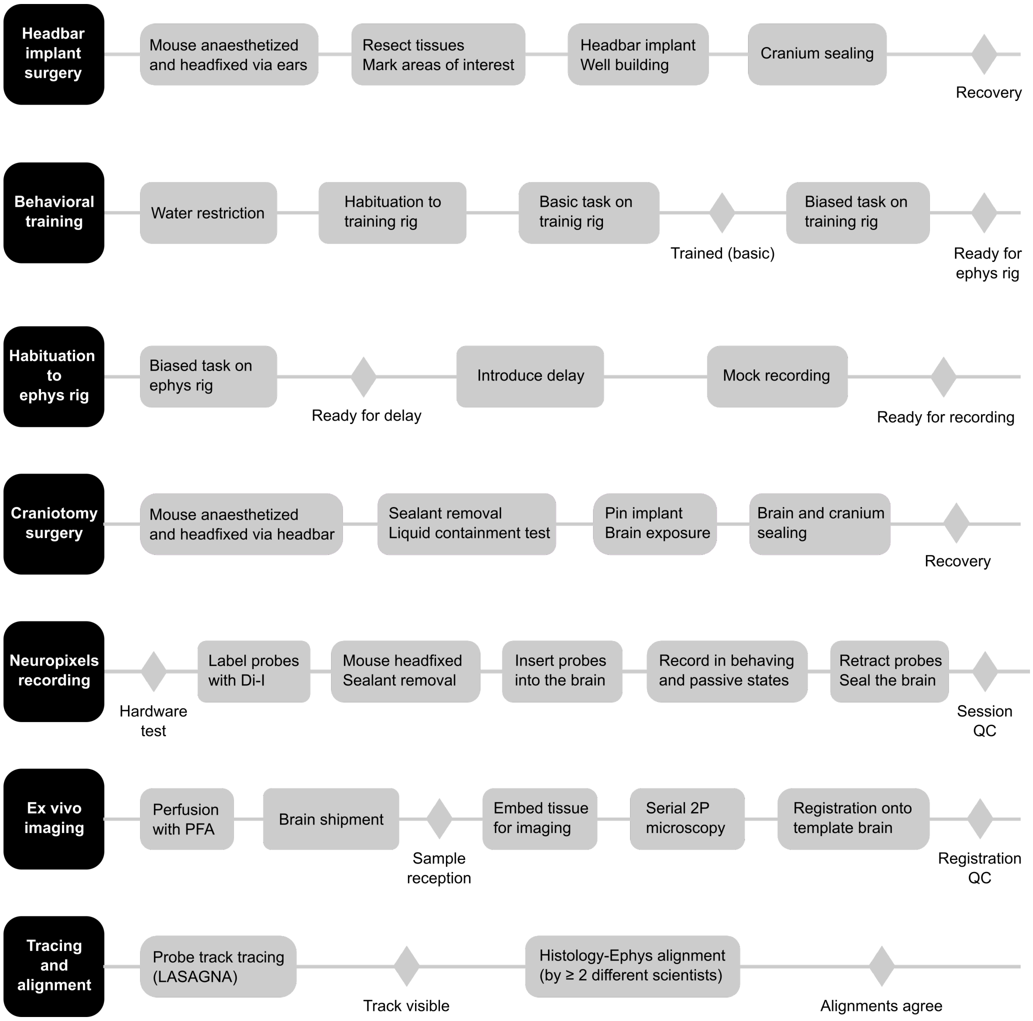
Waaga T, Agmon H, Normand VA, Nagelhus A, Gardner RJ, Moser MB, Moser EI, Burak Y. Grid-cell modules remain coordinated when neural activity is dissociated from external sensory cues. Neuron. 2022; .

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

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

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

Supplementary figures


Headbar implant surgery
Mouse anaesthetized and headfixed via ears
Resect tissues Mark areas of interest
Headbar implant Well building
Cranium sealing
Recovery
Behavioral training
Water restriction
Habituation to training rig
Basic task on trainig rig
Trained (basic)
Biased task on trainig rig
Ready for ephys rig
Habituation to ephys rig
Biased task on ephys rig
Ready for delay
Introduce delay
Mock recording
Ready for recording
Craniotomy surgery
Mouse anaesthetized and headfixed via headbar
Sealant removal Liquid containment test
Pin implant Brain exposure
Brain and cranium sealing
Recovery
Neuropixels recording
Hardware test
Label probes with Di-I
Mouse headfixed Sealant removal
Insert probes into the brain
Record in behaving and passive states
Retract probes Seal the brain
Session QC
Ex vivo imaging
Perfusion with PFA
Brain shipment
Sample reception
Embed tissue for imaging
Serial 2P microscopy
Registration onto template brain
Registration QC
Tracing and alignment
Probe track tracing (LASAGNA)
Track visible
Histology-Ephys alignment (by ≥ 2 different scientists)
Alignments agree

959

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

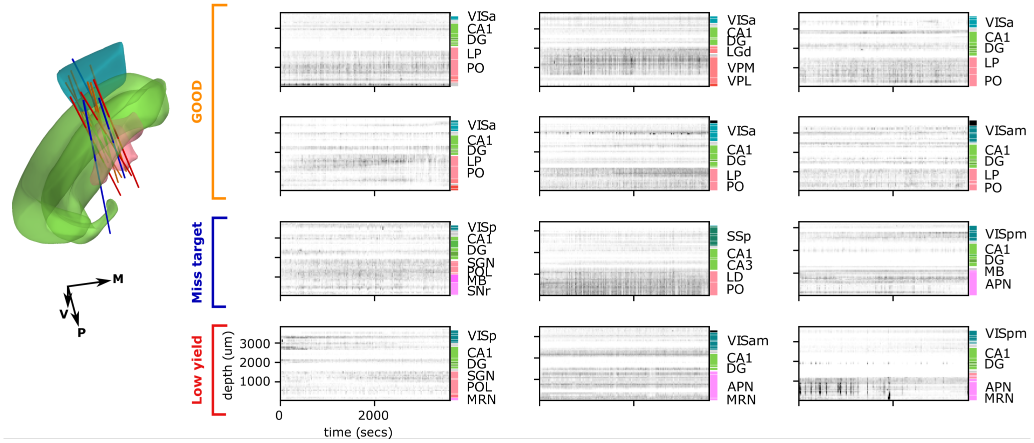


Figure 1–Figure supplement 2. Spiking activity qualitatively appears as heterogeneous across recordings. Example raster plots of neural activity recorded from the repeated site in N=12 mice. The raster plots in the first top two rows originate from sessions marked as being of good quality. The middle and bottom rows are raster plots from recordings that were excluded, based either on the probe misplacement, or the low number of detected units.

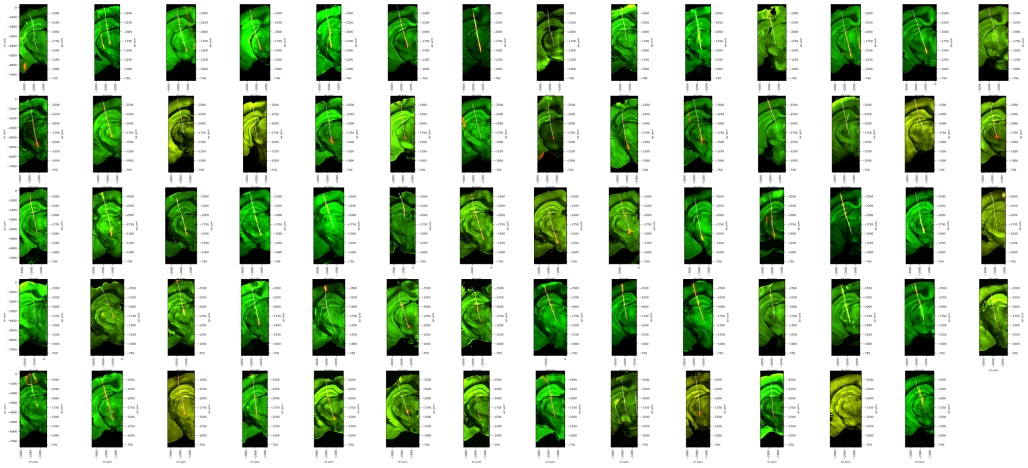
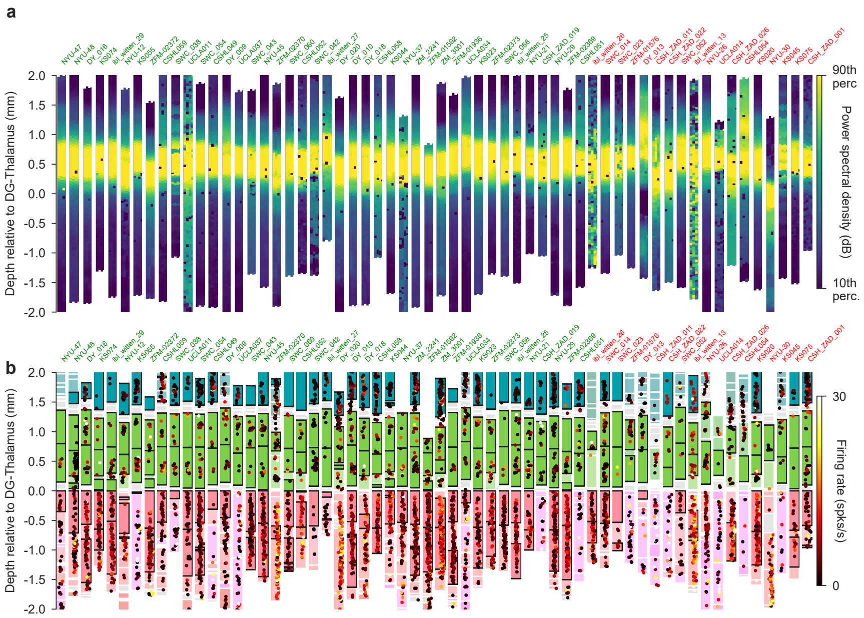


Figure 2–Figure supplement 1. Plots of all subjects with a repeated site insertion that were included in analysis of probe placement. Coronal tilted slices are made along the linearly interpolated best-fit to the histology insertion, shown through the raw histology (green: auto-fluorescence data for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



962 **Figure 3–Figure supplement 1.** Recordings that failed quality control were often visible outliers. **a**, Power spectral density between 20 and 80 Hz of all insertions, including those that failed to meet quality criteria. Recordings are labelled with the subject name above them; names in green passed quality control whereas names in red did not. **b**, Plots as in **a** but with firing rates of single neurons according to the depth at which they were recorded.

963

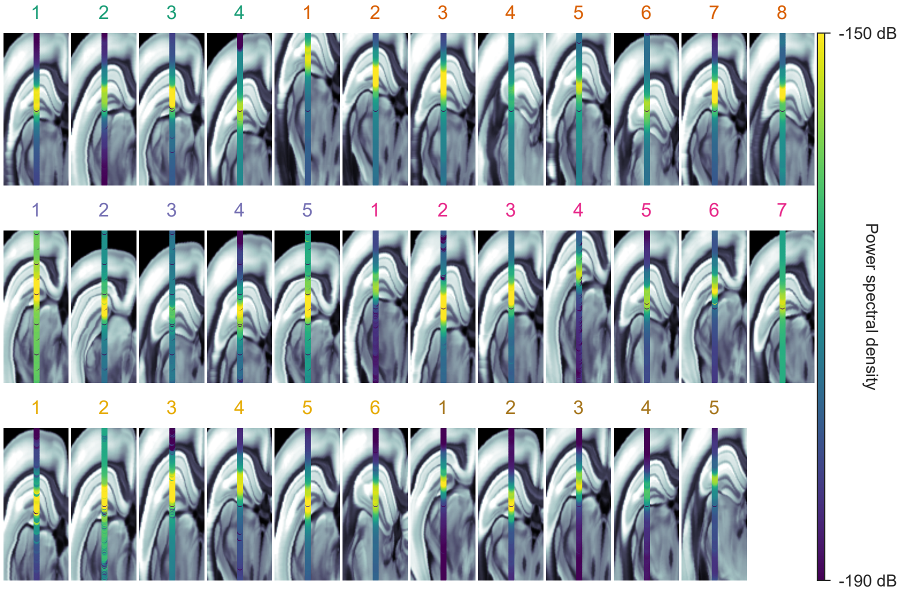


Figure 3–Figure supplement 2. Power spectral density between 20 and 80 Hz recorded along each probe shown in figure 3 overlaid on a coronal slice through brain. Each coronal slice has been rotated such that the probe lies along the vertical axis.

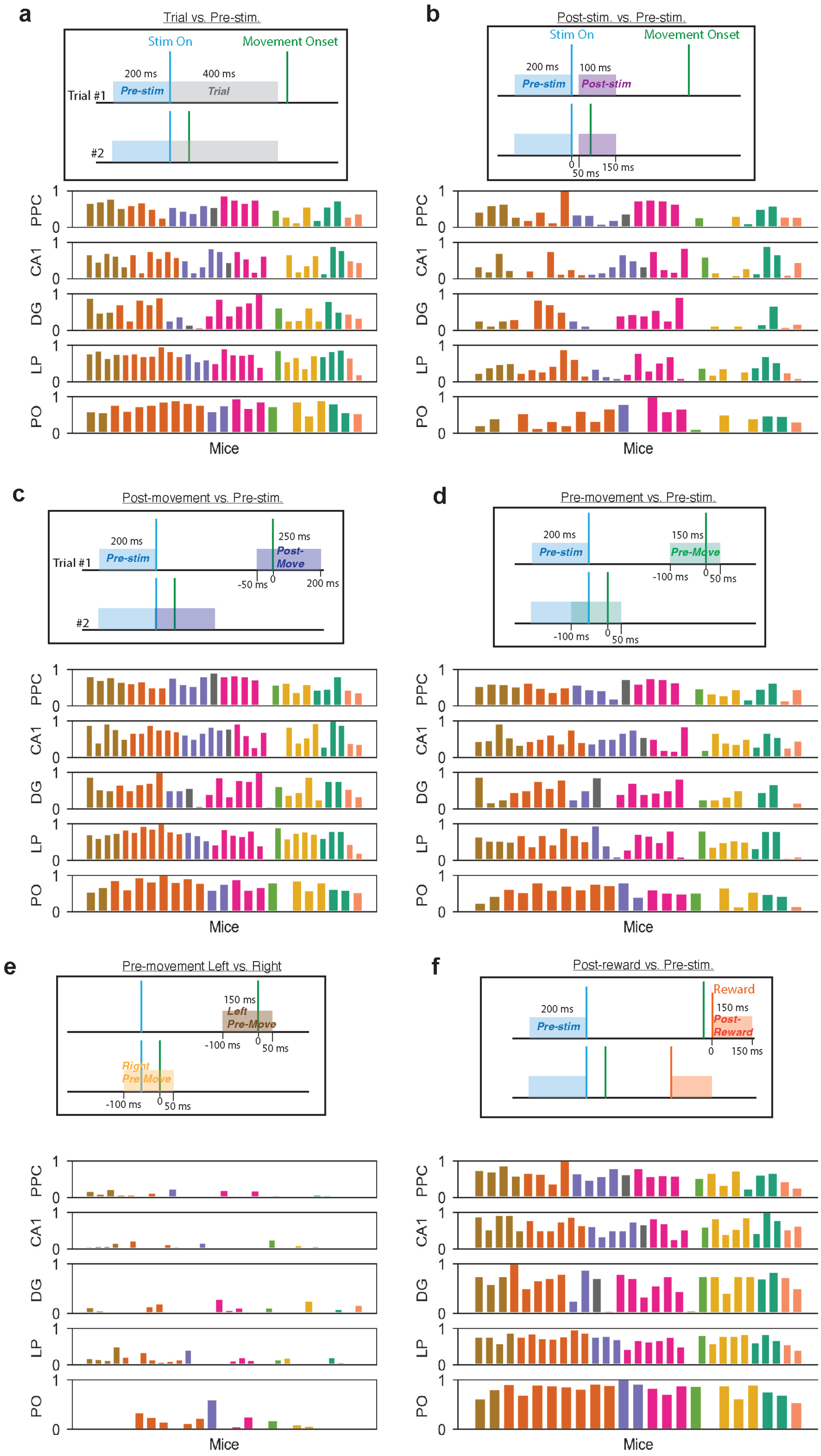


Figure 5–Figure supplement 1. (a)-(g) Schematics of six different tests performed (in addition to the test in Figure 5b) for finding task-modulated neurons. The two example trials show potential caveats of using each method; for instance, in **a**, the trial period may or may not include movement, depending on the reaction time in each trial. Below each schematic, the proportion of task-modulated neurons for the test is shown, across mice and brain regions, colored by lab ID.

965

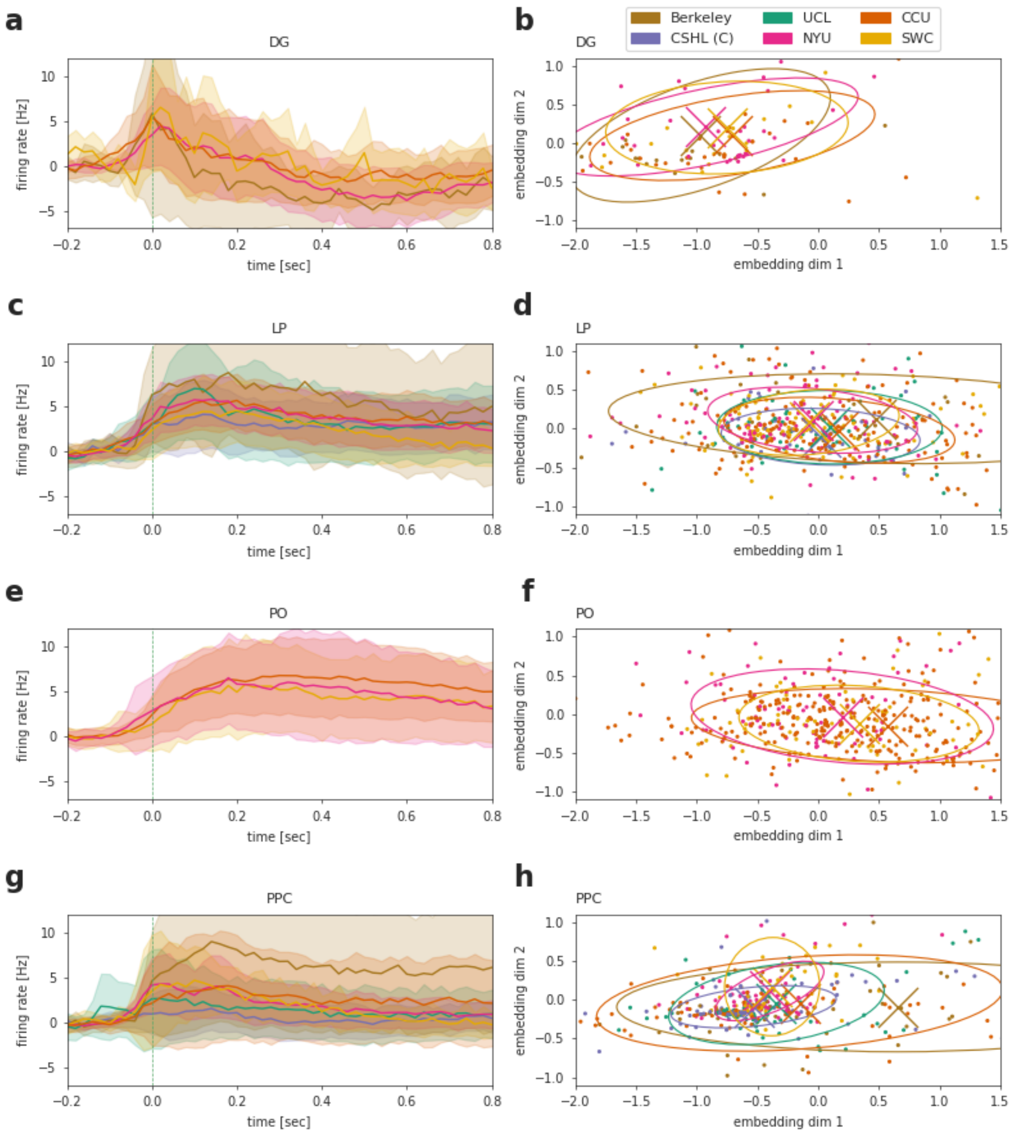


Figure 6–Figure supplement 1. Same as 6(e,f), for the remaining regions. Note that only Berkeley lab in region PPC differs significantly from the mean of all labs.

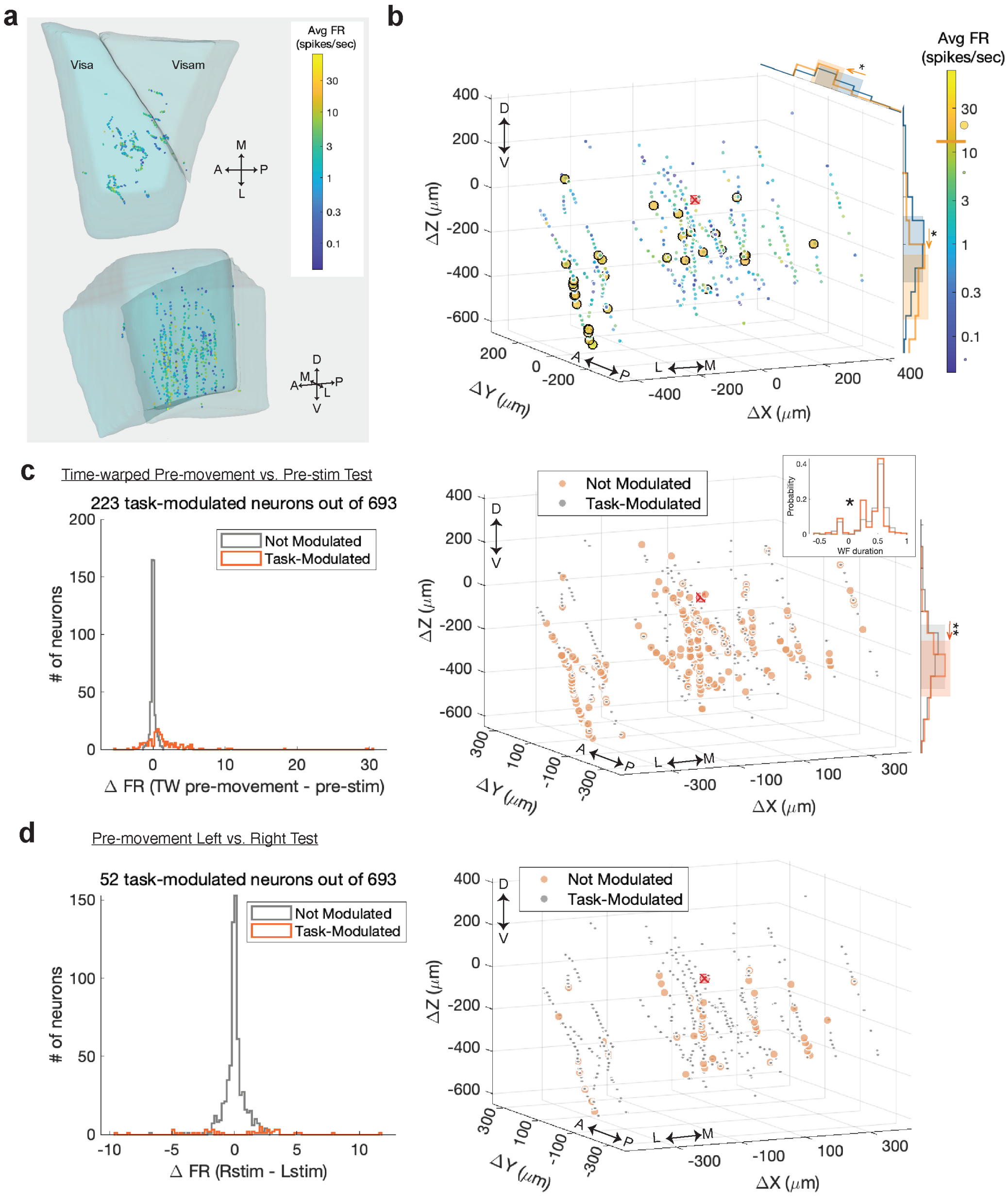


Figure 7–Figure supplement 1. High-firing and task-modulated PPC neurons are located in deeper layers than other PPC neurons. **(a-d)** Similar to Figure 7 but for PPC.

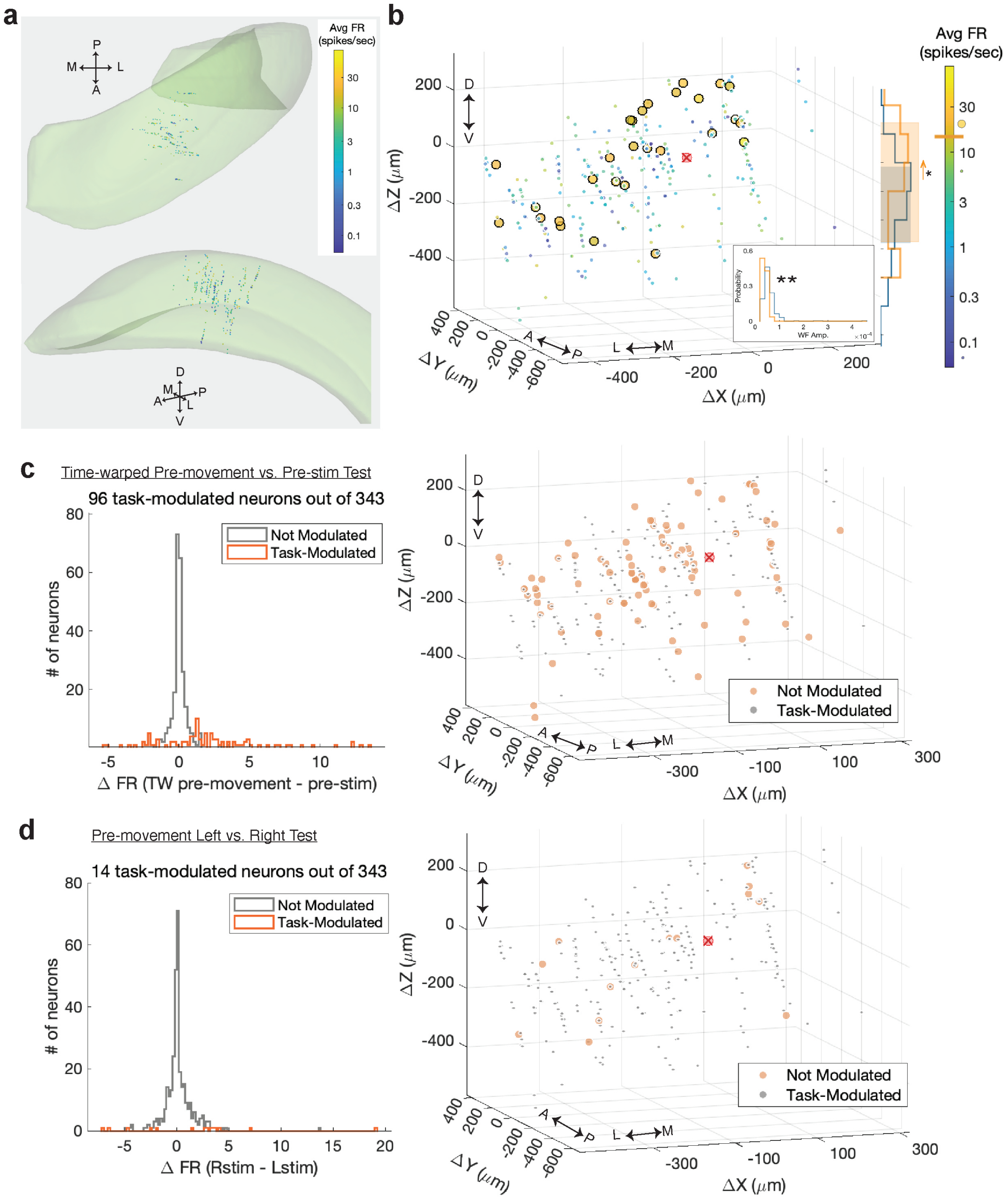


Figure 7–Figure supplement 2. High-firing, but not task-modulated, CA1 neurons are positioned more dorsally and have lower spike amplitudes than other CA1 neurons. **(a-d)** Similar to Figure 7 but for CA1.

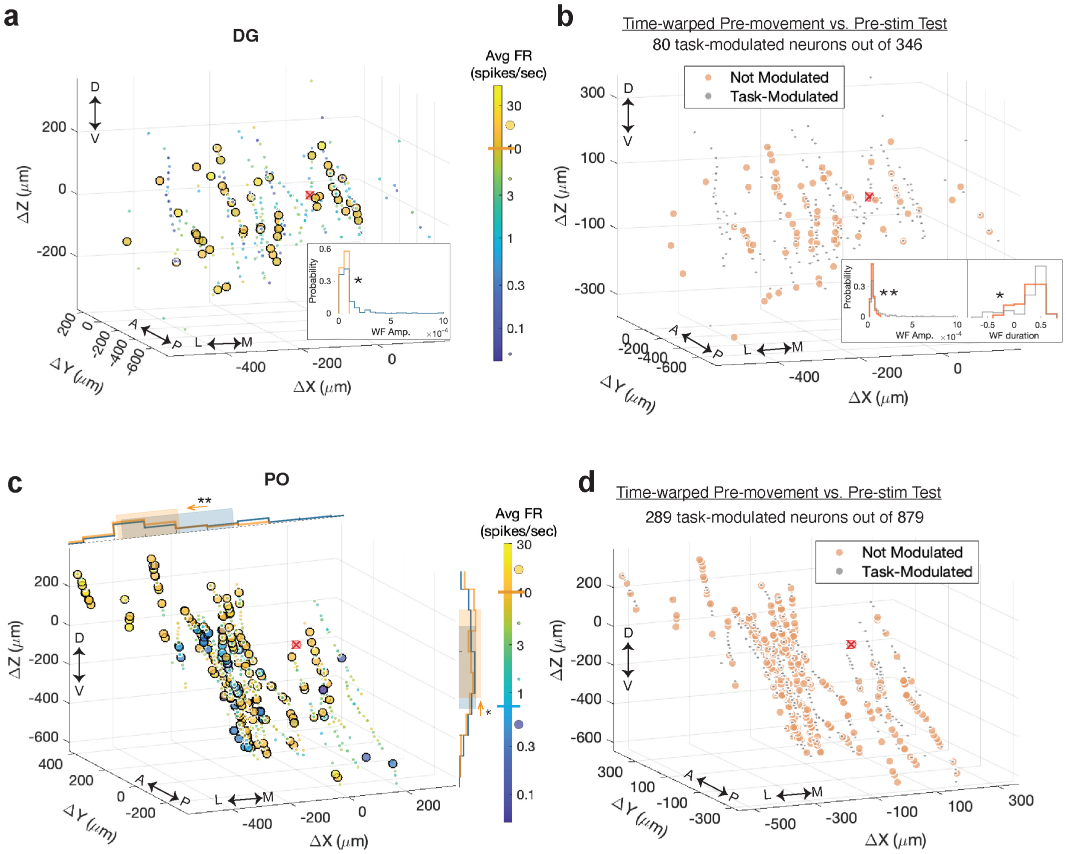


Figure 7–Figure supplement 3. Spatial positions and spike characteristics of outlier and task-modulated neurons in DG and PO are different from other neurons. **(a)** Spatial positions of DG neurons plotted as distance from the planned target center of mass, indicated with the red x. From comparisons of spatial position and waveform features, histogram of only those that were significantly different between the outliers (yellow) and regular neurons (blue) are shown: here, high-firing neurons have smaller waveform amplitudes. **(b)** Spatial positions of task-modulated and non-modulated DG neurons (using the time-warped pre-movement test) with the histogram of significant features shown (here, waveform amplitude and duration). For some other task-modulation tests (not shown), spatial positions of DG neurons were also significantly different. **(c-d)** Same as **a-b** but for PO neurons. In **c**, outliers included high and low firing neurons, making up 209 out of 879 neurons (44 of which are low firing). Shaded areas indicate the 20th and 80th percentiles of the neuron's spatial positions.

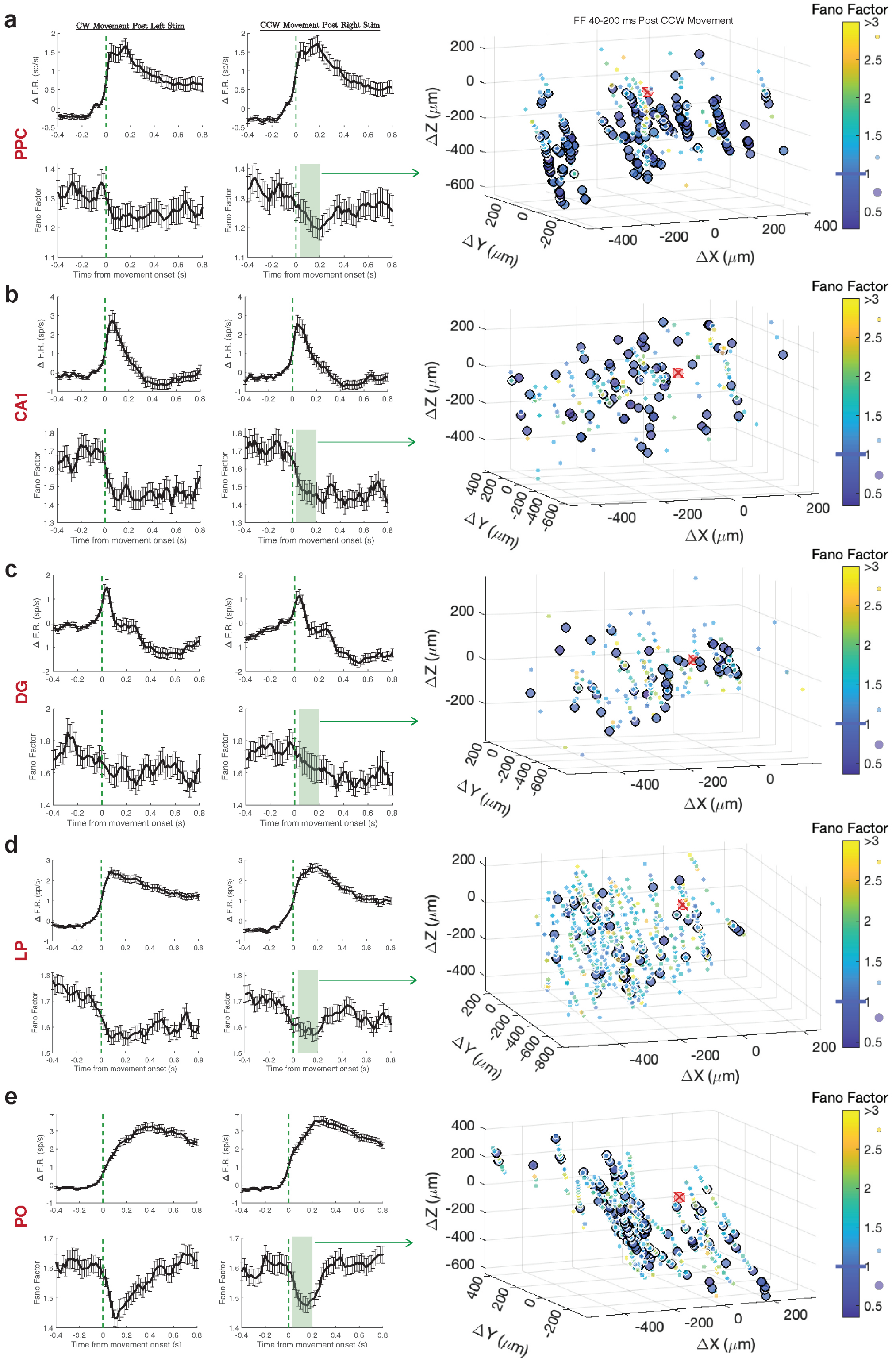


Figure 7–Figure supplement 4. Time-course and spatial position of neuronal Fano Factors. **(a)** *Left column*: Change in firing rate (top) and Fano Factor (bottom) averaged over all PPC neurons when aligned to movement onset after presentation of left or right full-contrast stimuli (correct trials only; 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 PPC (deeper neurons have lower Fano Factors) possibly due to higher drift in the activity of neurons closer to the surface over long recordings, from drying of the craniotomy. In the thalamus, spike characteristics between high and low Fano Factor neurons were significantly different (not shown).

970

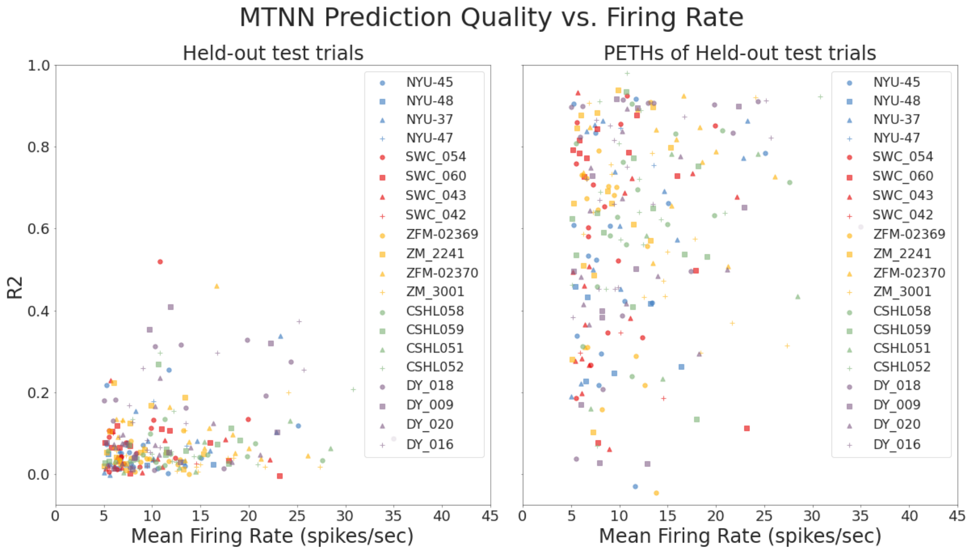


Figure 9–Figure supplement 1. For each unit in each session, we plot the MTNN prediction quality on held-out test trials against the firing rate of the unit averaged over the test trials. Each lab/session is colored/shaped differently. R^2 values on concatenations of the held-out test trials are shown on the left, and those on PETHs of the held-out test trials on the right.

971

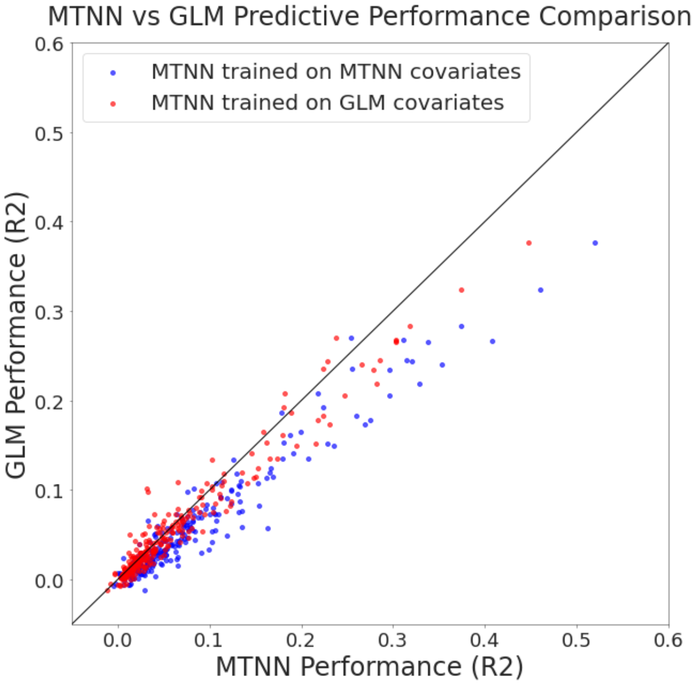
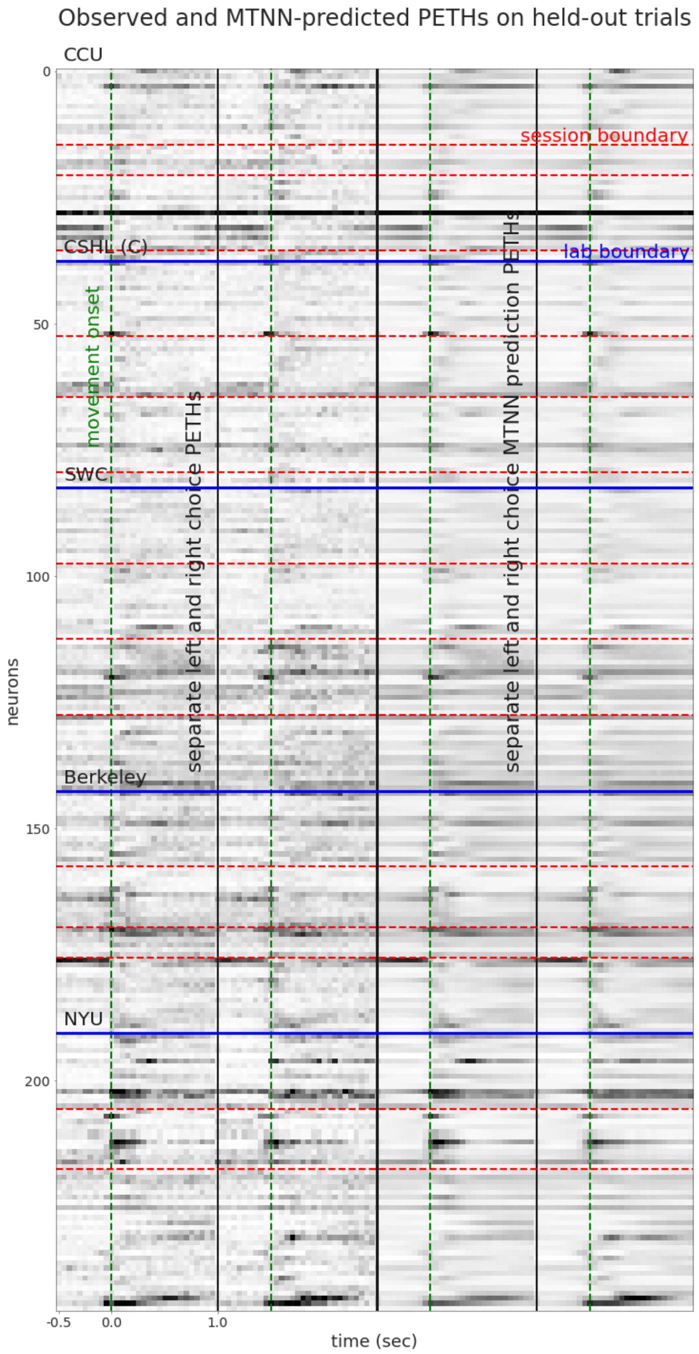
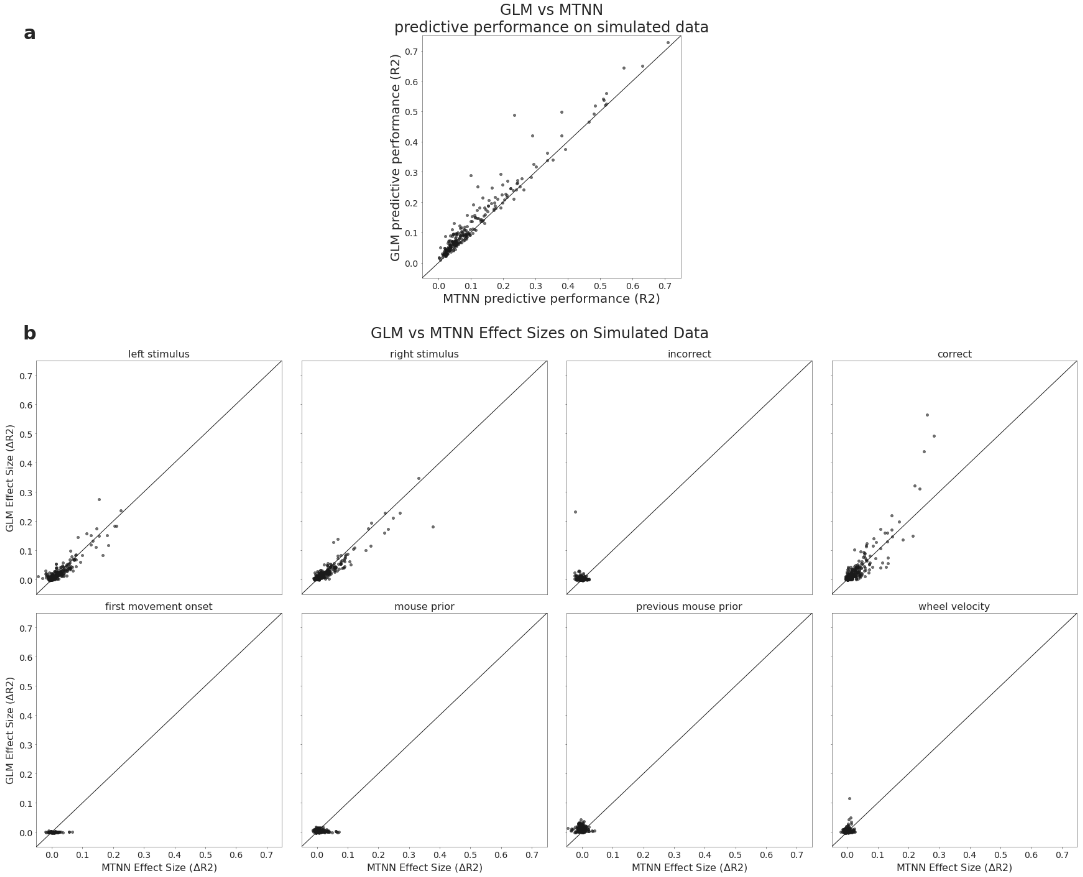


Figure 9–Figure supplement 2. MTNN and GLMs performs similarly on predicting the firing rates of held-out trials when trained on a reduced set of covariates, which includes stimulus onset time, stimulus side and contrast, feedback time and type, first movement onset time, wheel velocity, and mouse's prior. MTNN trained on the full set of covariates in Table 2 outperforms the MTNN/GLMs trained on the reduced covariate set.



972

Figure 9–Figure supplement 3. The left half shows for each neuron the trial averaged activity for left choice trials and next to it right choice trials. The vertical green lines show the first movement onset. The horizontal red lines separate recording sessions while the blue lines separate labs. The right half of each of these images shows the MTNN prediction of the left half. The trial-averaged MTNN predictions for held-out test trials captures visible modulations in the PETHs.



973

Figure 10–Figure supplement 1. 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 neural unit. 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 6 covariates. The effect sizes estimated by the MTNN and GLM leave-one-out analyses are comparable.

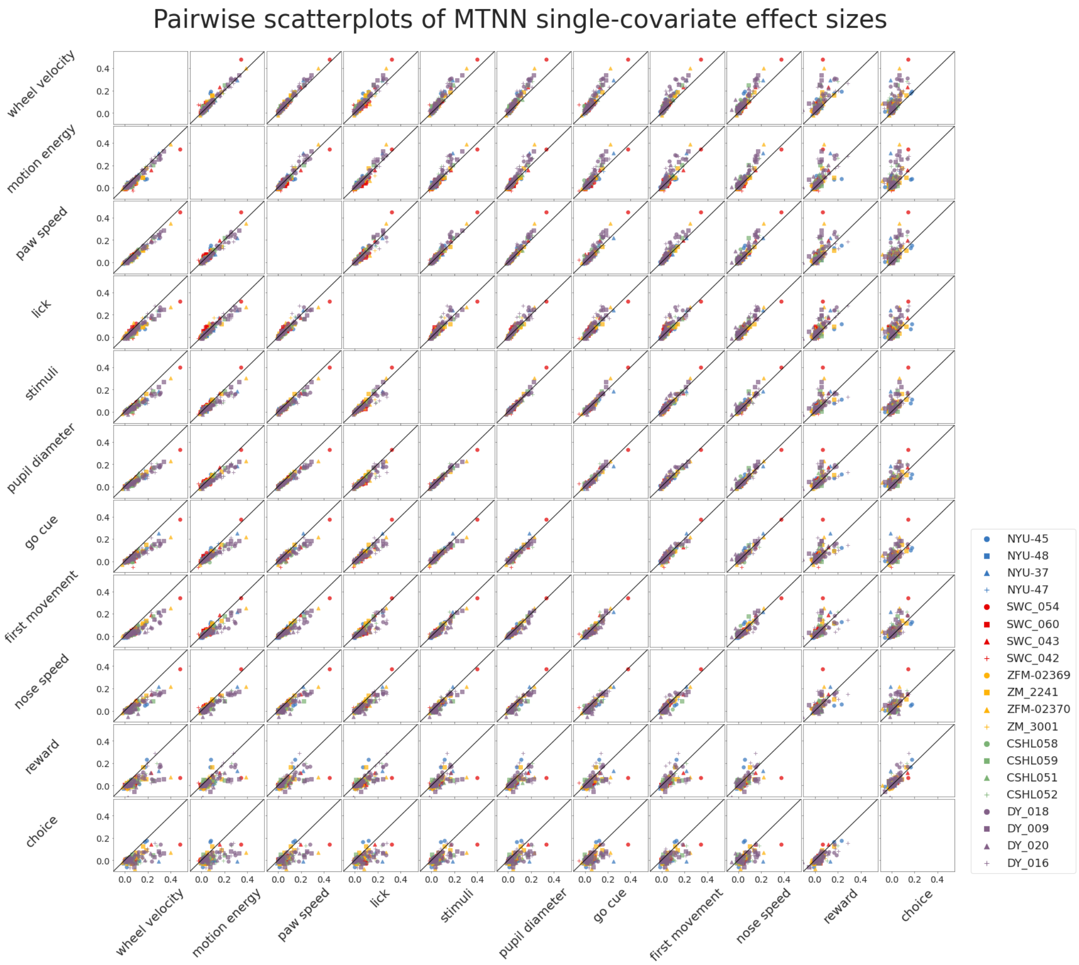


Figure 10–Figure supplement 2. We plot pairwise scatterplots of MTNN single-covariate effect sizes. Each dot represents the effect sizes of one neural unit and is colored by lab. There is no outlier lab. The effect sizes are highly correlated.