A System for Live Sorting of Neuronal Spiking Activity from Large-scale Recordings

Shreyas Muralidharan[1], Calvin Leng[2,3], Lucas Orts[1], Ethan Trepka[2], Shude Zhu[2,3], Matthew Panichello[2,3], Donatas Jonikaitis[3], Jacob Pennington[5], Marius Pachitariu[4], Tirin Moore[2,3]

Department of Electrical Engineering, Stanford University School of Engineering[1], Department of Neurobiology, Stanford University School of Medicine[2], and Howard Hughes Medical Institute[3], Stanford CA, USA, Howard Hughes Medical Institute[4], Ashburn, VA, USA, Department of Mathematics, Washington State University[5], Vancouver, WA, USA

Abstract

Online monitoring and quantification of neural signals has tremendous value both for neurofeedback experiments and for brain-computer interfaces. Unfortunately, established methods of online monitoring primarily involve the use of thresholded neural activity rather than sorted single-neuron spikes. The recent introduction of large-scale, high-density electrophysiology has enabled the recording of activity from hundreds of neurons simultaneously in both model organisms and human participants. This development highlights the need for a robust and easily implementable system for sorting spikes during data collection for 'live' analyses of neuronal signals. Here, we describe a system for live sorting of neuronal activity (LSS) based on the widely used Kilosort platform. The LSS workflow utilizes an initial period of recorded neural data to identify waveform templates using Kilosort 4. LSS then interfaces with the SpikeGLX API to retrieve small batches (e.g. 50 ms) of data and for processing online. We measured the similarity of single-neuron activity sorted live by LSS to that sorted offline in neurophysiological recordings from macaque visual cortex using Neuropixels probes. We show that LSS closely replicates the post-stimulus time histograms and visual response tuning curves of single-neurons obtained using offline sorting. Furthermore, we show that decoding neural signals online with LSS consistently outperforms online decoding of thresholded activity, and that LSS can achieve the same performance as that obtained with offline sorting.

Introduction

The use of online measurements of neural activity has played an important role in both basic and translational neuroscience. In basic neuroscience, online measurements of different levels of neural activity (e.g. membrane voltage, spiking, EEG) have been used to introduce closed-loop perturbations to address specific experimental hypotheses that directly test the contributions of neural activity to behavior[1–4]. By delivering interventions on the timescale of neural processing, such approaches allow experimenters to modulate ongoing computations and directly observe their behavioral consequences. Previous closed-loop experiments involving neuronal spiking activity have been used to demonstrate the relationship between voluntarily controlled prefrontal activity and selective attention [5] and the influence of decision variables on choice behavior [6]. More recently, this approach was used to link operant control of hippocampal place cell activity to the memory of remote locations[7]. Notably, measurements of spiking activity in these and other closed-loop studies were limited to small numbers of single neurons or to multi-unit activity.

In addition to basic neuroscience, online measurement of neural activity has proven instrumental in translational neuroscience, such as in the restoration of functions to individuals with motor or communication impairments. Brain-computer interfaces (BCIs) have emerged as powerful approaches that harness neural signals in real time to re-establish lost capabilities. Recent work utilizing spiking-based BCIs has demonstrated the decoding of handwriting[8], reaching and grasping[9], and speech [10,11] in participants otherwise impaired in those functions. These successes leverage the persistence of cortical representations that remain years after trauma or disease onset. Achieving such performance requires reliable, real-time extraction of

neural activity from large neural ensembles, requiring decoding algorithms capable of operating on millisecond timescales. However, at present, spiking-based BCIs rely solely on thresholded neural activity rather than isolated single-neuron spikes.

The recent advent of large-scale, high-density electrophysiology has enabled the recording from hundreds of neurons simultaneously. This approach uses multi-electrode arrays (MEAs) with large channel counts (100s) and spatially dense electrode layouts (e.g. 20 micron spacing) that support high-quality spike sorting of extracellular single-neuron activity. A noteworthy example of MEAs are the recently developed Neuropixels probes. Neuropixels are silicon-based probes with continuous, dense, and programmably selectable electrode contacts. To date, they have been used widely in a variety of model species including nonhuman primates and human participants [12–23].

The proliferation in the use of MEAs highlights the need to provide a system for achieving high-quality spike sorting online that is easy to implement and well-integrated with widely used large-scale neurophysiological tools. A key limitation of the standard neurophysiological workflow is that it precludes experimental interventions based on the activity of large local or distributed populations of single neurons. A wealth of studies have linked spatiotemporal patterns of neuronal population activity to fundamental components of perception, cognition and behavior [14,21,23–27]. Further evidence suggests specific roles of neuronal subtypes (e.g. 'fast-spiking' neurons) in those functions[28–32], thus indicating a need for a system to measure signals from neuronal subpopulations online. Such a system could naturally be deployed in BMI settings as well. To address this gap, we developed a system for online, 'live' spike sorting (LSS) that enables access to single-neuron activity at short latencies. Our system

builds on the widely used Kilosort platform[33,34] to facilitate adoption by the broad community of neurophysiologists.

Implementation of Live Spike Sorter

To date, a number of algorithms have been developed to transform extracellular multi-electrode recording data into single-neuron clusters, recovering each neuron's spatiotemporal electrical footprint across channels and its corresponding spike times. At present, the vast majority of laboratories utilizing MEAs, particularly Neuropixels probes, use Kilosort for spike sorting[21–23,33,35–39]. Typical MEA workflows follow a two-stage pipeline (Figure 1, top). During the recording phase, extracellular electrophysiological data is acquired and written to disk using standard data acquisition systems and software (e.g. SpikeGLX). Following the experiment, offline spike sorting is performed using the chosen sorting algorithm, producing estimates of neuron locations, spike templates, and spike times. These outputs are then analyzed in relation to experimental variables to study neural circuits. LSS establishes a new workflow in which a short training period of recording is first used to establish unit cluster positions and waveform templates (Figure 1, bottom). Next, subsequently acquired raw electrophysiological data are processed and spike sorted online, and the assignment of activity to clusters is done using those established in the training period. Activity can then be analyzed on the timescale of changing experimental variables.

To achieve live sorting of neuronal activity, LSS replicates the key steps of Kilosort4 in an online manner. We chose to develop LSS on the Kilosort4 framework due to its widespread use by laboratories for spike-sorting high-density electrophysiological recordings. For our purposes, we define live as the ability to sort activity from short intervals (e.g. 50 ms) of data in less than the duration of that interval. A few key implementation choices were made to achieve

online performance (Figure 2). The first is the use of GPUs, which excel at parallelized computations, such as template convolution. Second, we used the SpikeGLX API to fetch the latest batch of data as it is recorded. Each batch of data is first copied from CPU to GPU, and a series of pre-processing steps (e.g. high pass filter, whitening, projection to PCA space) are performed. A matching pursuit algorithm is used to convolve the templates found in the training period with each pre-processed batch. The matching pursuit loop iteratively subtracts off the best matching template until no spikes remain (typically 10-15 iterations). The mean time for each step is shown in Figure 2. Third, in the current version, we assume stability of the recording and assign live sorted spikes to clusters established in the training period based on Euclidean distance in PCA space. This assumption also allowed us to reuse the final drift estimation from Kilosort.

We chose to deploy and benchmark the sorter on batch sizes of 50 ms due to the time scale of fluctuation of neural states of interest. However, the live spike sorter can be deployed on smaller batch sizes for quicker latencies if needed. We benchmarked the system using both an NVIDIA A6000 GPU, and a RTX 3070 GPU. Both GPUs were able to achieve much quicker than live performance while sorting units from the entire 384 channel Neuropixels array. Our results to date suggest that the current hardware implementation could likely process live sorting of units from >384 channels. All code and step-by-step implementation instructions are available at Github: *https://github.com/MShreyasStanford/LiveSpikeSorter*

Graphical user interface

An important component of the LSS system is the user-friendly graphical interface (GUI) from which the user monitors, selects and evaluates the sorted spike data as it is logged (Figure 3). The LSS GUI consists of several panels in which key data are continuously updated, the most

basic being the spike rasters of all clustered units. In addition, another key panel continuously updates the accumulating distribution of processing times for all batches. In our current implementation of LSS, each 50-ms batch of data is processed in 9 +/- 3 milliseconds which is largely the sum of times denoted in Figure 2. These data can be used to evaluate the extent to which sorting is being performed online. It can also inform hardware decisions and whether sorting should instead be done on a subset of units. Lastly, another panel in the GUI is used to monitor single-neuron statistics for selected units. In the current version, single-neuron statistics such as binned firing rate and sorting quality metrics, e.g. inter-spike interval histograms, and autocorrelograms are included.

Validation of LSS Training Period

To determine a sufficient duration of training data for the LSS workflow, we first evaluated how well Kilosort4 retrieves the activity of ground truth, simulated units using data of varying durations (Figure 4a-c). Next, using real data, for which ground truth is unavailable, we then examined the extent to which we could retrieve units obtained in full recordings using subsets of varying durations (Figure 4e-g). To do this, we employed a large, recently published dataset of neurophysiological recordings from macaque prefrontal cortex[21]. This dataset was used because the large number of recordings provided the opportunity to assess the impact of variations in recording stability and unit yield on LSS. To leverage that variation, we also assessed the effect of probe drift during the training period on the match rate. We looked at the match rate between units recovered during various training subsets of the data and units from the whole recording for both simulated and real data. This analysis revealed a negative correlation between match rate

and probe drift rate ($r=-0.53$, $p<10^{7}$)(Figures d,h). This result emphasizes the importance of recording stability before establishing the training period for live spike sorting.

Benchmarking LSS

We benchmarked the performance of LSS in sorting spikes online using Neuropixels recordings obtained from large populations of visually responsive neurons in the macaque visual cortex. We chose visual cortex as a testing ground for LSS given that neurons there exhibit relatively reliable and interpretable, low-latency and distinctly selective responses to sensory stimulation which can be used to exemplify the benefits of single-neuron isolation. Neurons were recorded in the superior temporal sulcus (STS), primarily area MT while monkeys performed a standard fixation task (METHODS). During fixation, drifting grating stimuli spanning multiple drift directions were presented for 250 ms three times each trial. Neurophysiological data were obtained from two male monkeys across 10 sessions (7 monkey A; 3 monkey H). Each recording session began with the slow insertion of the Neuropixels NHP 1.0 probe into the STS. Using the uncurated Kilosort4 output, these sessions yielded a total of 6,983 isolated single neuronal clusters (419-957 neurons/session). In each recording session, the visual responses to the grating stimuli were measured over a duration of 5-28 minutes.

To evaluate the performance of OSS, we compared its readout of spiking activity to that obtained from the offline workflow. We compared LSS to the offline output in three ways. First, to assess whether LSS could reproduce the fine temporal spiking dynamics of individual neurons, we compared the peri-stimulus time histograms of matched units clustered via the two methods. Second, we assessed the similarity of visual motion selectivity output by LSS with that of offline by comparing the tuning curves of matched units. To perform the above two

comparisons, units from the LSS training period were matched to units obtained offline using the Kilosort matching procedure described above. From this, we obtained 1,029 matched neurons across the ten sessions. Third, we compared the performance of decoders trained on the total population activity retrieved online and offline. Each of these comparisons were carried out using the matched units derived from recordings in area MT.

Single Unit peri-stimulus time histograms(PSTHs)

We first compared the spike trains of single units retrieved online with LSS to those retrieved offline using the full set of matched units. During each recording in area MT, drifting grating stimuli were presented multiple times at the location of neuronal RFs thereby evoking multiple short latency responses from neurons during each trial (Figure 5a). This provided an opportunity to assess similarities in the temporal spiking dynamics of individual unit responses retrieved online and offline. We therefore compared the peri-stimulus time histograms (PSTHs) generated from the two methods. We found that the PSTHs were highly correlated across all matched units (median Pearson correlation = 0.96)(Figure 5b). This demonstrates that LSS captures the fine temporal spiking dynamics of populations of single neurons.

Comparison of tuning to direction of visual motion in online and offline sorted neurons

Next, we compared the responses of single units retrieved online and offline to varying directions of visual motion (Figure 6). Area MT is organized into cortical columns in which neurons within the same column exhibit similar preferences to the direction of visual motion [40]. Our dataset consists of recordings made tangential to the cortical surface and consequently pairs of neurons located further apart tended to exhibit larger differences in their preferred motion direction. For

example, units separated by several hundred microns of each other exhibited clear differences in their preferred direction of motion (Figure 6a). For each of the 1,029 matched units, we fit von Mises functions to their responses across motion direction. Of these, 309 units were well fit (r2>0.4) in both the online and offline retrieved activity. We found that the tuning peaks of these units were remarkably similar between online and offline methods (Figure 6b) (mean=0.29 deg.). Lastly, to measure the tuning similar for all matched units, we computed Pearson correlations between the online and offline tuning curves. Overall, we found that tuning curves were highly correlated between the two methods (median = 0.91)(Figure 6c). Thus, we find that LSS accurately captures the responses of single neurons across stimulus conditions.

Decoding direction of motion from neural data

A major advantage of sorting single-unit data online is that the sorted activity can then be used to decode signals conveyed by populations of individual neurons. Thus, to further validate OSS, we measured the extent to which population decoding from activity sorted online achieved comparable performance to offline decoding. Using data from each of the recording sessions from area MT, we trained multi-class logistic regression classifiers separately on spiking activity retrieved online with LSS and offline (METHODS). Each classifier was trained to decode one of twelve directions of grating drift across 100 ms time bins around the time of stimulus onset. In addition, we trained classifiers on multi-unit threshold-crossings employing methods widely used for online measurements of neuronal activity [9,11,41](METHODS). This enabled direct comparisons of decoding performance between the three methods.

We first compared the performance of decoders trained on all units (features) available from each session's data using offline, live sorted and thresholded activity. Across the 10

sessions, as expected, classifiers trained on offline sorted activity decoded the direction of motion extremely well following stimulus onset. In many sessions, decoding performance during the stimulus period approached 100% (88.54 %+/- 2.93%; chance performance=12%)(Figure 7a). Performance of classifiers trained on thresholded activity was less robust (68.47%+/- 3.26%). Classifier performance from thresholded activity was worse than that obtained from offline sorted activity in every recording session, and was reduced on average by 20%. This loss in performance was significant across sessions ($p<10^{-5}$). In contrast, the performance obtained from classifiers trained on sorted activity using LSS closely approximated that of offline classifiers (87.05%+/- 3.26%). Overall, the loss in performance from online sorted activity averaged only 2.3%, and was significantly less than the loss observed with thresholded activity ($p<10^{-4}$). Thus, we observed a substantial net benefit from online sorting with LSS over the multi-unit thresholding commonly used to decode neuronal signals online.

Notably, the three different sorting algorithms yielded unequal numbers of units in each session (Figure 7a), thus hindering fair comparisons of decoding performance. In all sessions for example, offline sorting yielded more units than that obtained from live sorting, presumably due to a greater duration in the dataset. Thus, we sought to compare decoding performance independent of the difference in units. To do so, we computed standard feature dropping curves[42,43] from each session's data for offline, live sorted and thresholded activity. These curves were then fitted with asymptotic functions to extract the number of features at which performance saturated (asymptotic performance)(Figure 7b) as well as the rate of saturation (METHODS). Across the 10 sessions, the performance of decoders trained on offline sorted data saturated at larger numbers of units compared to live sorted (mean difference in units 42.5 units; $p<0.05$) and thresholded activity (mean difference 35.9 units, $p>0.05$). More importantly,

comparisons of the asymptotic performance demonstrated that only the thresholded activity underperformed that obtained from the offline sorted data (mean difference in performance 17.3%, $p<0.001$). By contrast, performance obtained from live sorted data was statistically indistinguishable from offline sorting (mean difference in performance +/- 1.4% , $p=0.13$). Thus, when controlling for the number of units, we found that LSS could achieve the same performance as that obtained with offline sorting.

We found that neither the probe drift rate ($r=0.049$, $p>0.05$) nor the duration of the test recordings ($r=-0.222$, $p>0.05$) significantly correlated with the small loss in performance from using online sorted activity.

DISCUSSION

Here, we have described LSS, a system for sorting neuronal spiking activity as it is collected live from large-scale microelectrode probes. The system leverages the technical advantages and broad use of Kilosort4 to provide a robust and easily implementable system for obtaining online analyses of single-neuron spiking data. Live sorting with LSS is achieved by utilizing an initial period of recorded neural data to identify waveform templates and then assigning incoming data to trained templates. We benchmarked LSS by comparing its output to that obtained with offline sorting. We show that LSS closely replicates the post-stimulus time histograms and visual response tuning curves of single-neurons obtained using offline sorting. Furthermore, we show that decoding neural signals online with LSS consistently outperforms online decoding of thresholded activity, and that LSS achieves virtually the same performance as that obtained with offline sorting.

As this is the first iteration of LSS, we have considered a number of ways in which the system might be advanced further to improve upon its performance and alignment with the quality of offline sorted data. First, estimation of drift can be done online using principles from the Kilosort4 algorithm, allowing for the improvement of tracking of clusters identified in the training period. Furthermore, simultaneous updating of training clusters could be achieved by running offline Kilosort on data written to disk during live sorting. Thus, the training templates could be updated to enable live sorting of units that may become detectable after the conclusion of the training period. Second, manual curation of units found in the training period could easily be achieved to impose stricter criteria on live sorting. Finally, integration with other data acquisition platforms such as Open Ephys could enable more widespread utility of LSS.

Acknowledgements

The authors would like to thank Danielle Lopes and Stephen Cital for veterinary assistance, and Frank Lanfranchi for helpful comments. This work was supported by NIH grants EY014924 and NS11662302, and a Ben Barres Professorship to T.M..

Methods

Experimental subjects

Neural recordings were collected in a total of ten sessions from two male Rhesus macaques (A: age 14 years, weight 11 kg, H: age 11 years, weight 17kg). All surgical and experimental procedures were approved by the Stanford University Institutional Animal Care and Use Committee and were in accordance with the policies and procedures of the National Institutes of Health.

Electrophysiological recordings

We recorded spiking activity from neurons using primate Neuropixels probes (IMEC). In each recording session, the dura was perforated with a 21-gauge pointed cannula mounted on an independent micromanipulator. A single Neuropixels NHP probe was then advanced through the cannula using either a hydraulic Narishge microdrive or a motorized drive system with custom 3D-printed grid inserts. Recording sites were configured as 384 simultaneously active channels in a contiguous block (3.84mm span), providing dense sampling along the probe shank. Neuropixels probes digitized signals at the headstage and streamed separate bands for action potentials (AP; 300 Hz high-pass; 30 kHz sampling) and local field potential (LFP; 1 kHz low-pass; 2.5 kHz sampling). Neural activity was monitored online and saved to disk using SpikeGLX (https://billkarsh.github.io/SpikeGLX/).

Behavioral task and recording setup.

Neural recordings were conducted in the superior temporal sulcus (STS) while monkeys performed a fixation task in which visual stimuli were presented in the peripheral visual field to map receptive fields (RFs) and to measure direction of motion selectivity. The location of the STS was identified using anatomical MRIs in both monkeys and confirmed based on the functional properties of recorded neurons including RF size and robust selectivity to planar visual motion which indicated that most neurons were in area MT. Neural data were recorded with Neuropixels 1.0 NHP probes. Experiment code was written in MATLAB (MathWorks, version R2020b) using the Psychophysics Toolbox. Stimuli were presented on a ViewPixx LCD monitor with 1920x1080 pixels resolution and 100 Hz refresh rate (VPixx Technologies). The monkeys viewed the display from 42 cm. Eye position was monitored at 1000 Hz using monocular corneal reflection and pupil tracking with an Eyelink 1000 Plus (SR Research Ltd., Ottawa, ON, Canada). Eye-tracker calibration was performed with a five point protocol at the beginning of each recording session. Video fixation task. In the video fixation task, each trial started when the monkey fixated a central spot (blue, 0.5 deg. for monkey T; red, 1 deg. for monkey A) for 300 ms. The fixation spot was presented offset from the center of the screen (location reported individually for each session in the dataset) on a mid-gray (RGB=127) background. The monkey received a juice reward for maintaining fixation for the duration of the trial.

Visual Stimuli

For receptive field (RF) mapping, the monkey fixated a central spot to initiate a trial. Subsequently, a series of RF mapping stimuli (20 for monkey T, 13 for monkey A) were then presented at various eccentricities, each for 100 ms with no inter-stimulus interval. The stimuli in the RF mapping task were sinusoidal gratings (spatial frequency 1 cycle/degree) inside a Gaussian envelope (sigma 1 deg.) with eight different carrier orientations. The carrier wave changed phase continuously at 15 Hz. The monkey received a juice reward for maintaining fixation for the duration of the trial. Gratings drifted at a rate of 20 deg/sec.

Training Length Validation

We determined unit matches as in Pachitariu et al, 2024. Briefly, we computed false positive rate (FP) and false miss rate (FM) for pairs of spike trains and defined their match score as 1-FP-FM. Units with a score of at least 0.8 are considered matched. Match rate is defined as the number of matched units divided by the number of units compared. Since different recording lengths were used, we only compared spike trains on the four 5-minute segments shared by all recordings.

For simulated data, we determined matches against ground truth units. For real data, no ground truth was available, so we tested for matches against the Kilosort4 sorting result for the full 1-hour recording. Since sorting results are noisy, we estimated an upper bound for match rates on the real data as the median match rate for 50-minute segments. We sorted five 50-minute segments of each recording, offset by 1-minute intervals. Additionally, since results for the real data may contain artifacts, units that drop in or out, and noisy MUA units that are not suitable for matching, we only matched units labeled 'good' by Kilosort4 with at least a 1 Hz spike rate for the tested portion of the spike trains.

We estimated drift rate from the per-batch drift amounts used for drift correction by Kilosort4. We smoothed these drift amounts with a 2nd-order Savitzky-Golay filter with a window size of 361 batches (12 minutes) to reduce noise, then summed across timepoints the velocity values (first-order absolute differences between consecutive batches). We estimated drift separately for each recording segment.

Implementation of LSS

LSS can be implemented with any CUDA enabled GPU. We tested the system with CUDA version 11.3. LSS can be found at this link: https://github.com/calvinleng97/OnlineSpikes_test/, and we recommend using the microsoft visual studio compiler and linker (cl.exe). We have not tested compilation of this project under any standard other than C++17. Components of the system, such as the LSS or Python GUI can be modified to fit specific use cases. Detailed instructions can be found at the github.

Neuron Dropping Curves

The following function was used to fit the neuron dropping curves for all datasets. Here, k denotes the number of units, c the chance-level accuracy, A the asymptotic decoding accuracy, and τ the rise time constant in units of neurons.

$$y(k) = c + (A - c)\left(1 - e^{-k/\tau}\right)$$

Thresholded activity

For comparisons of decoding performance with units extracted offline or with LSS, we decoded visual motion direction from area MT neurons using unsorted activity. To achieve this, we trained classifiers on multi-unit threshold-crossings 4.5 standard deviations below the mean of the voltage trace, similar to previous studies.[9,11,41]

References

1. Ezzyat, Y. *et al.* Closed-loop stimulation of temporal cortex rescues functional networks and improves memory. *Nat. Commun.* **9**, 365 (2018).
2. Shibata, K., Watanabe, T., Sasaki, Y. & Kawato, M. Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science* **334**, 1413–1415 (2011).
3. deBettencourt, M. T., Cohen, J. D., Lee, R. F., Norman, K. A. & Turk-Browne, N. B. Closed-loop training of attention with real-time brain imaging. *Nat. Neurosci.* **18**, 470–475 (2015).
4. Sitaram, R. *et al.* Closed-loop brain training: the science of neurofeedback. *Nat. Rev. Neurosci.* **18**, 86–100 (2017).
5. Schafer, R. J. & Moore, T. Selective attention from voluntary control of neurons in prefrontal cortex. *Science* **332**, 1568–1571 (2011).
6. Peixoto, D. *et al.* Decoding and perturbing decision states in real time. *Nature* **591**, 604–609 (2021).
7. Coulter, M. E. *et al.* Closed-loop modulation of remote hippocampal representations with neurofeedback. *Neuron* **113**, 949–961.e3 (2025).
8. Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M. & Shenoy, K. V. High-performance brain-to-text communication via handwriting. *Nature* **593**, 249–254 (2021).
9. Hochberg, L. R. *et al.* Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature* **485**, 372–375 (2012).
10. Moses, D. A. *et al.* Neuroprosthesis for decoding speech in a paralyzed person with Anarthria. *N. Engl. J. Med.* **385**, 217–227 (2021).

11. Willett, F. R. *et al.* A high-performance speech neuroprosthesis. *Nature* **620**, 1031–1036 (2023).

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

13. Steinmetz, N. A. *et al.* Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science* **372**, eabf4588 (2021).

14. Steinmetz, N. A., Zatka-Haas, P., Carandini, M. & Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576**, 266–273 (2019).

15. Peters, A. J., Fabre, J. M. J., Steinmetz, N. A., Harris, K. D. & Carandini, M. Striatal activity topographically reflects cortical activity. *Nature* **591**, 420–425 (2021).

16. Jia, X. *et al.* High-density extracellular probes reveal dendritic backpropagation and facilitate neuron classification. *J. Neurophysiol.* **121**, 1831–1847 (2019).

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

18. Allen, W. E. *et al.* Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science* **364**, 253 (2019).

19. Vesuna, S. *et al.* Deep posteromedial cortical rhythm in dissociation. *Nature* **586**, 87–94 (2020).

20. Gardner, R. J. *et al.* Toroidal topology of population activity in grid cells. *Nature* **602**, 123–128 (2022).

21. Panichello, M. F. *et al.* Intermittent rate coding and cue-specific ensembles support working memory. *Nature* **636**, 422–429 (2024).

22. Lanfranchi, F. F., Wekselblatt, J., Wagenaar, D. A. & Tsao, D. Y. A compressed hierarchy

for visual form processing in the tree shrew. *Nature* **646**, 872–882 (2025).

23. Stine, G. M., Trautmann, E. M., Jeurissen, D. & Shadlen, M. N. A neural mechanism for terminating decisions. *Neuron* **111**, 2601–2613.e5 (2023).
24. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. *Annu. Rev. Neurosci.* **36**, 337–359 (2013).
25. Michaels, J. A. *et al.* Sensory expectations shape neural population dynamics in motor circuits. *Nature* (2025) doi:10.1038/s41586-025-09690-9.
26. Marshel, J. H. *et al.* Cortical layer-specific critical dynamics triggering perception. *Science* **365**, eaaw5202 (2019).
27. Engel, T. A. *et al.* Selective modulation of cortical state during spatial attention. *Science* **354**, 1140–1144 (2016).
28. Shin, H. & Moore, C. I. Persistent gamma spiking in SI nonsensory fast spiking cells predicts perceptual success. *Neuron* **103**, 1150–1163.e5 (2019).
29. Ceccarelli, F. *et al.* Static and dynamic coding in distinct cell types during associative learning in the prefrontal cortex. *Nat. Commun.* **14**, 8325 (2023).
30. Diester, I. & Nieder, A. Complementary contributions of prefrontal neuron classes in abstract numerical categorization. *J. Neurosci.* **28**, 7737–7747 (2008).
31. Hussar, C. R. & Pasternak, T. Memory-guided sensory comparisons in the prefrontal cortex: contribution of putative pyramidal cells and interneurons. *J. Neurosci.* **32**, 2747–2761 (2012).
32. Merrikhi, Y. *et al.* Spatial working memory alters the efficacy of input to visual cortex. *Nat. Commun.* **8**, 15041 (2017).
33. Pachitariu, M., Sridhar, S., Pennington, J. & Stringer, C. Spike sorting with Kilosort4. *Nat.*

Methods **21**, 914–921 (2024).

34. Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M. & Kenneth D., H. Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. *bioRxiv* (2016) doi:10.1101/061481.
35. Paulk, A. C. *et al.* Large-scale neural recordings with single neuron resolution using Neuropixels probes in human cortex. *Nat. Neurosci.* **25**, 252–263 (2022).
36. International Brain Laboratory *et al.* A brain-wide map of neural activity during complex behaviour. *Nature* **645**, 177–191 (2025).
37. Leonard, M. K. *et al.* Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature* **626**, 593–602 (2024).
38. Gilliland, R. L., Selvanayagam, J., Zanini, A., Johnston, K. D. & Everling, S. Neural activity for complex sounds in the marmoset anterior cingulate cortex. *Commun. Biol.* **7**, 1310 (2024).
39. Chen, S. *et al.* Brain-wide neural activity underlying memory-guided movement. *Cell* **187**, 676–691.e16 (2024).
40. Albright, T. D. Direction and orientation selectivity of neurons in visual area MT of the macaque. *J. Neurophysiol.* **52**, 1106–1130 (1984).
41. Trautmann, E. M. *et al.* Accurate estimation of neural population dynamics without spike sorting. *Neuron* **103**, 292–308.e4 (2019).
42. Lebedev, M. A. How to read neuron-dropping curves? *Front. Syst. Neurosci.* **8**, 102 (2014).
43. Wessberg, J. *et al.* Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature* **408**, 361–365 (2000).

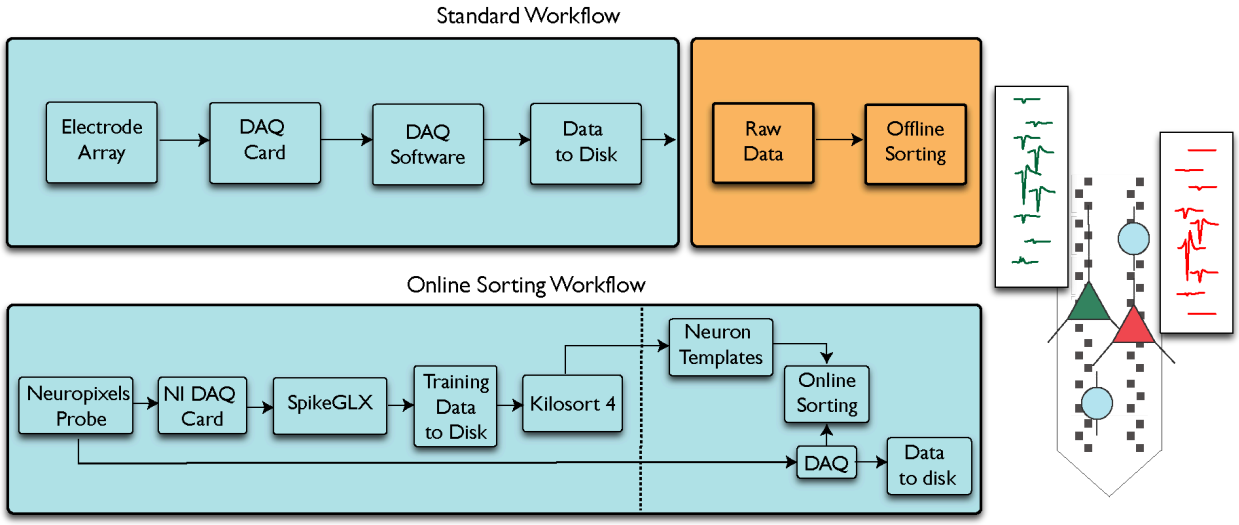

Figure 1. Standard and live Spike Sorting Workflows. Top, in the standard workflow, data are acquired during a recording session (shown in blue), and written to disk. Subsequently, stored experimental data are accessed for offline (shown in orange) sorting (e.g. Kilosort 4) to retrieve single-neuron information (waveform locations, templates, and spike times) which are then used for analyses. Bottom, in the online spike sorting (OSS) workflow, a two-step process (separated by dotted line) begins with an initial training recording. Sorting is then performed on the training data. Templates and neuron information from these data are then used to sort subsequent data online, while simultaneously writing the same data to disk. Spike data sorted online can then be analyzed during the recording, e.g. for neurofeedback. Right cartoon depicts multi-electrode recordings from adjacent neurons and the post processed waveform templates across locations..

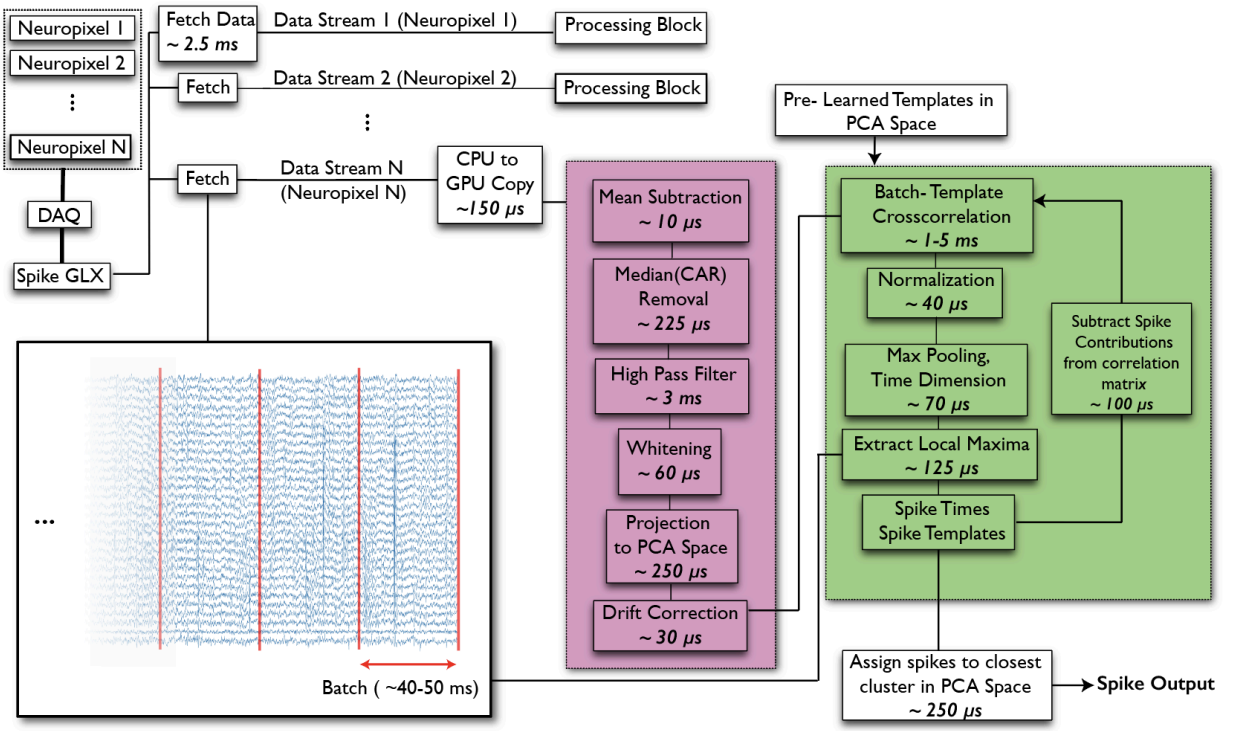

Figure 2. Processing steps for LSS. Bottom left inset shows a portion of a continuous data stream, processed in small (~20-50 ms) batches (red lines). Each data stream (1-N) is retrieved independently via a fetch call to the SpikeGLX API, followed by a processing block, consisting of pre-processing and spike retrieval. Each batch of data (e.g. N) goes through a series of pre-processing steps (purple) after a CPU to GPU copy. This is followed by spike retrieval via an orthogonal matching pursuit (OMP) algorithm (green). The OMP loop terminates when no further spikes remain. The location of each spike in PCA space is used to assign that spike to its closest cluster in the training data.

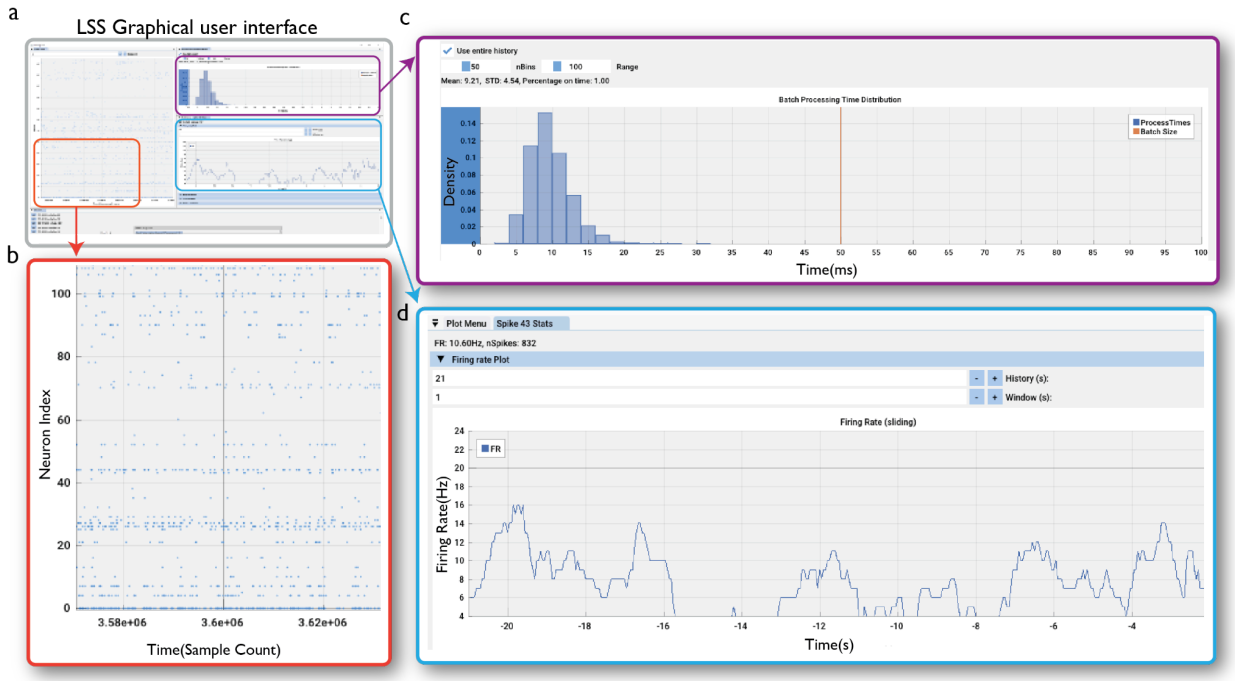

Figure 3. Graphical User Interface for LSS. A, Live raster of spiking activity across sorted neurons. The x axis displays sample count from the start of SpikeGLX acquisition. B, histogram of processing times for every 50-ms batch of data, relative to batch duration (orange line). C, Panel to view selected single-neuron statistics, e.g. binned firing rate as shown. Other metrics include spike sorting quality metrics, such as inter-spike interval distributions and autocorrelograms.

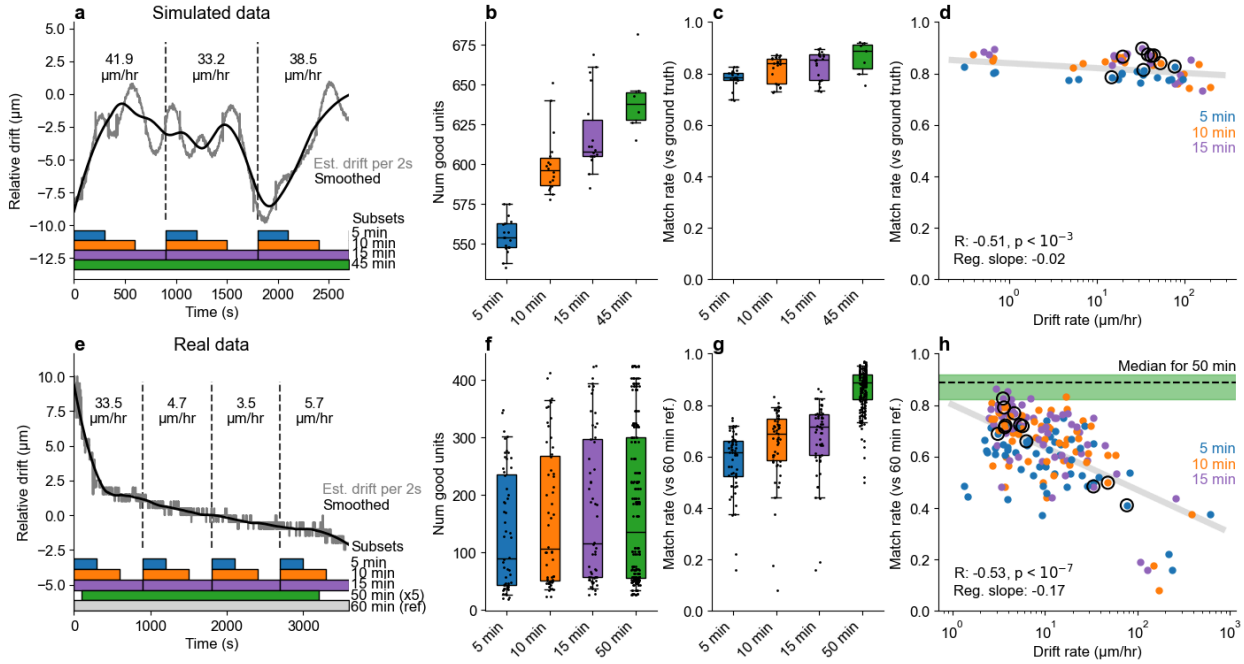

Figure 4. Effect of recording duration and drift on spike sorting. A, Estimated drift rate for example simulated recording with realistic drift. Sorting was performed on 5-, 10-, 15- and 45-minute subsets. B, Number of good units identified by Kilosort4 for each subset. Boxes are 25/75% quartiles, center is median. Whiskers indicate the 95th and 5th percentiles. C, Proportion of 600 ground truth units correctly identified by Kilosort4 ("match rate"). D, Match rate as a function of drift rate, colored by subset, with linear regression fit. Circled points are examples in a. E-H, As a-d, but for real data. Since no true ground truth was available, we used as ground truth the Kilosort4 sorting result of the full 60-minute recordings. Dashed upper bound in h is estimated from repeated sorting of slightly shifted 50-minute subsets to illustrate the noise in the spike sorting process.

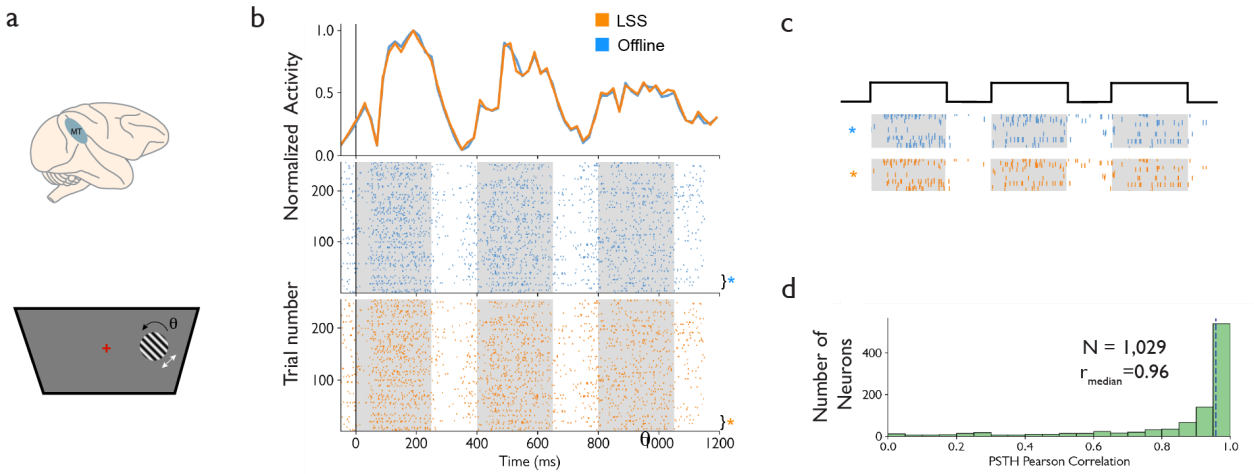

Figure 5. Comparison of live and offline sorted peri-stimulus time histograms. A, Activity of neurons in area MT (top) were measured in response to drifting gratings varying in direction (bottom). B, Response of an example area MT unit to three presentations of a drifting grating across multiple trials. Top, peri-stimulus time histograms are shown for activity retrieved with LSS (orange) and offline (blue). Bin size, 100 ms; step size, 10 ms. Bottom, raster plots of live and offline activity. Gray shading denotes the onset and duration of stimulus presentations. C, Magnified view of spike rasters from several trials labeled (}) in B. D, Distribution of correlations between peri-stimulus time histograms for all online and offline sorted units.

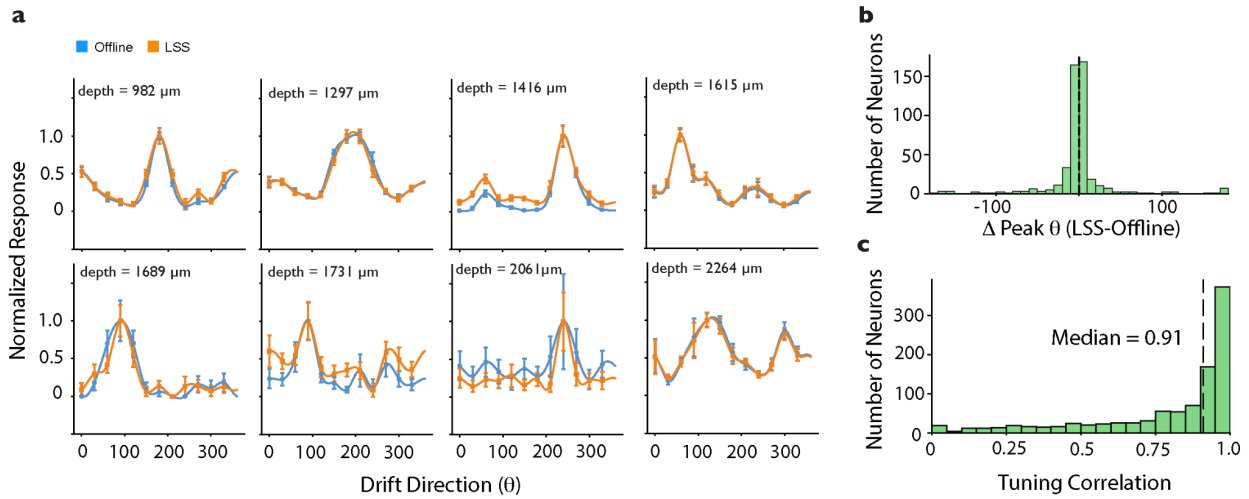

Figure 6. Comparison of online and offline sorted tuning curves. A, tuning to the direction of visual motion in drifting grating stimuli of eight example area MT neurons in a single recording session sorted offline (blue) and with LSS (orange). Responses of each neuron were normalized to the peak firing rate and fitted with von Mises functions. Error bars denote SEM. Recording depths are shown for each neuron from shallow to deep. B, Distribution of differences between online and offline tuning peaks for neuronal responses well fit by the von Mises function. C, Distribution of correlations between tuning curves for all online and offline sorted units.

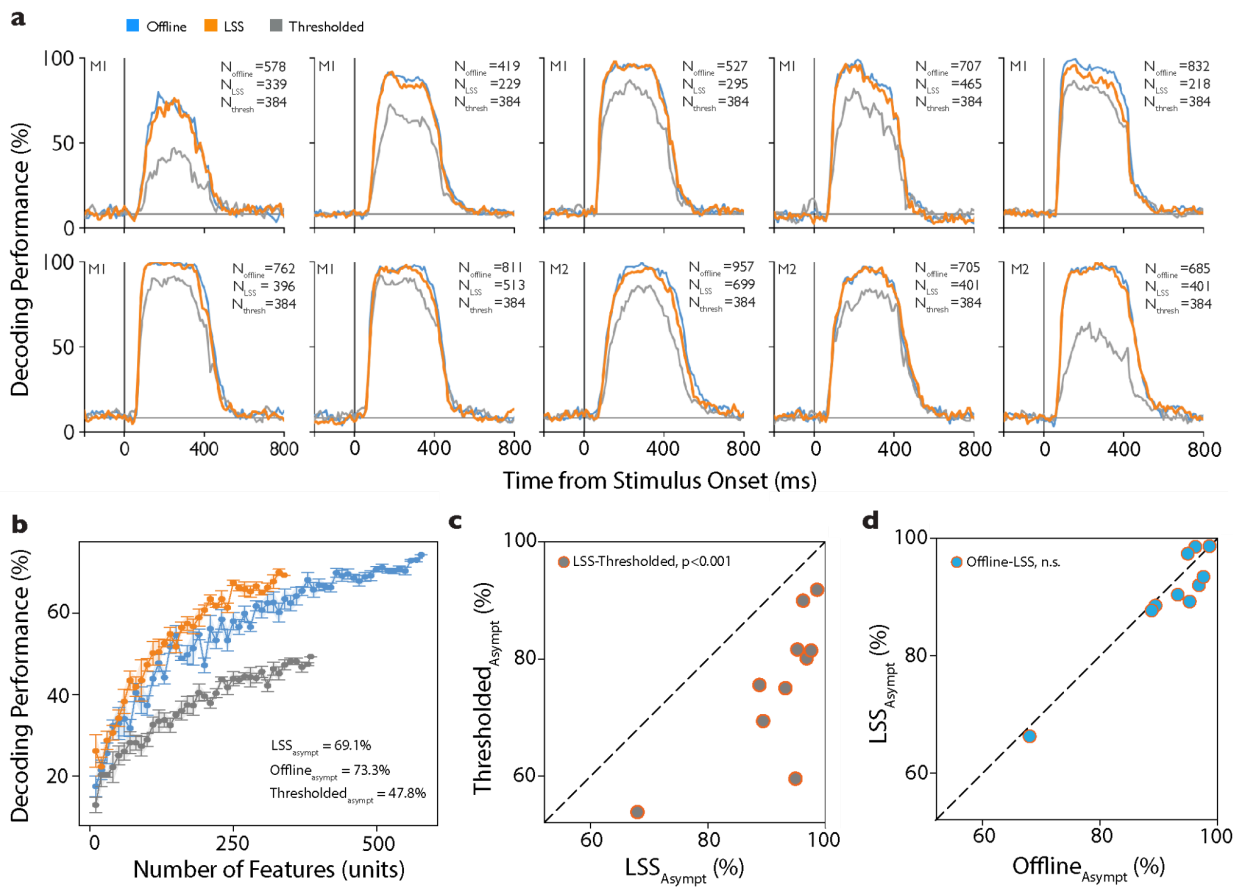

Figure 7. Comparison of population decoding from offline sorted, live sorted and thresholded activity. A, Classifier decoding of visual motion direction across time for populations of area MT neurons in 10 recording sessions in two monkeys. Decoding performance compares offline sorted spikes, live sorted spikes, and thresholded activity (gray). For the latter, activity was thresholded at 4.5 standard deviations below the mean voltage at all 384 channels. Each session trace shows decoding performance aligned to stimulus onset. Gray line denotes chance performance (12%). B, Feature-dropping curves for live sorted, offline sorted and thresholded activity in a single recording session. C, Comparison of asymptotic performance between live sorted and thresholded activity. D, Comparison of asymptotic performance between live sorted and offline sorting. Dotted lines denote unity.