eLife

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
February 27, 2024 (this version)

Reviewed preprint version 1
December 12, 2023

Sent for peer review
September 18, 2023

Posted to preprint server
September 3, 2023

Neuroscience

Multi-day Neuron Tracking in High Density Electrophysiology Recordings using EMD

Augustine(Xiaoran) Yuan, Jennifer Colonell, Anna Lebedeva, Michael Okun, Adam S. Charles ✉, Timothy D. Harris ✉

Janelia Research Campus, Howard Hughes Medical Institute, USA • Department of Biomedical Engineering, Center for Imaging Science Institute, Kavli Neuroscience Discovery Institute, Johns Hopkins University, USA • Sainsbury Wellcome Centre, University of Sheffield, UK • Department of Psychology and Neuroscience Institute, Howard Hughes Medical Institute, USA

Abstract

Accurate tracking of the same neurons across multiple days is crucial for studying changes in neuronal activity during learning and adaptation. New advances in high density extracellular electrophysiology recording probes, such as Neuropixels, provide a promising avenue to accomplish this goal. Identifying the same neurons in multiple recordings is, however, complicated by non-rigid movement of the tissue relative to the recording sites (drift) and loss of signal from some neurons. Here we propose a neuron tracking method that can identify the same cells independent of firing statistics, which are used by most existing methods. Our method is based on between-day non-rigid alignment of spike sorted clusters. We verified the same cell identify using measured visual receptive fields. This method succeeds on datasets separated from one to 47 days, with an 84% average recovery rate.

eLife assessment

This **important** study proposes a new method for tracking neurons recorded with Neuropixel electrodes across days. The methods and the strength of the evidence are **convincing**, but the authors do not address whether their approach can be generalized to other brain areas, species, behaviors, or tools. Overall, this method will be potentially of interest to many neuroscientists who want to study long-term activity changes of individual neurons in the brain.

1 Introduction

The ability to longitudinally track neural activity is crucial to understanding central capabilities and changes of neural circuits that operate on long time-scales, such as learning and plasticity,[1–4] motor stability,[1], [5], [6] etc. We seek to develop a method capable of tracking single units regardless of changes in functional responses for the duration of an experiment spanning one to two months.

High-density multi-channel extracellular electrophysiology (ephys) recording devices enable chronic recordings over large areas over days-to-months.[7] Such chronic recordings make possible experiments targeted at improving our understanding of neural computation and underlying mechanisms. Examples include perceptual decision making, exploration and navigation.[8]–[13] Electrode arrays with hundreds to thousands of sites, for example Neuropixels, are now used extensively to record the neural activity of large populations stably and with high spatio-temporal resolution, capturing hundreds of neurons with single neuron resolution.[9], [10] Moreover, ephys retains the higher time resolution needed for single spike identification, as compared with calcium imaging that provides more spatial cues with which to track neurons over days.

The first step in analyzing ephys data is is to extract single neuron signals from the recorded voltage traces, i.e., spike sorting. Spike sorting identifies individual neurons by grouping detected action potentials using waveform profiles and amplitudes. Specific algorithms include principal components based methods[14] and,[15] and template matching methods, for example, Kilosort.[9], [11], [16], [17] Due to the high dimensional nature of the data, spike sorting is often computationally intensive on large data sets (10's to 100's of GB) and optimized to run on single sessions. Thus processing multiple sessions has received minimal attention, and the challenges therein remain largely unaddressed.

One major challenge in reliably tracking neurons is the potential for changes in the neuron population recorded (***Figure 1*** a and ***Figure 1*** b). In particular, since the probe is attached to the skull, brain tissue can move relative to the probe, e.g. during licking, and drift can accumulate over time.[18] Kilosort 2.5 corrects drift within a single recording by inferring tissue motion from continuous changes in spiking activity and interpolating the data to account for that motion.[7] Larger between-recording drift occurs for sessions on different days, and can 1) change the size and location of spike waveforms along the probe,[19] 2) lose neurons that move out of range, and 3) gain new neurons that move into recording range. Thus clusters can change firing pattern characteristics or completely appear/disappear. As a result the specific firing patterns classified as unit clusters may appear and disappear in different recordings.[9], [20]–[22] Another challenge is that popular template-matching-based spike sorting methods usually involve some randomness in template initialization.[16], [23], [24] As a result, action potentials can be assigned into clusters differently, and clusters can be merged or separated differently across runs.

Previous neuron tracking methods are frequently based on waveform and firing statistics, e.g., firing rate similarity,[25] action potential shape correlation and inter-spike interval histogram(ISI) shape.[26] When neuronal representations change, e.g., during learning[1]–[3] or representational drift,[27] neural activity statistics became less reliable. In this work, we take advantage of the rich spatialtemporal information in the multi-channel recordings, matching units based on the estimated neuron locations and unit waveforms,[28] instead of firing patterns.

As an alternative method, Steinmetz et al.[7] concatenated pairs of datasets after low resolution alignment, awkward for more than 2 datasets. We report here a more flexible, expandable and robust tracking method that can track neurons effectively and efficiently across any number of sessions.

2 Results

2.1 Procedure

Our datasets consist of multiple recordings taken from three mice (***Figure 7*** a) over 2 months. The time gap between two recordings ranges from two to 25 days. Each dataset is spike-sorted individually with a standard Kilosort 2.5 pipeline. The sorting results, including unit assignment,

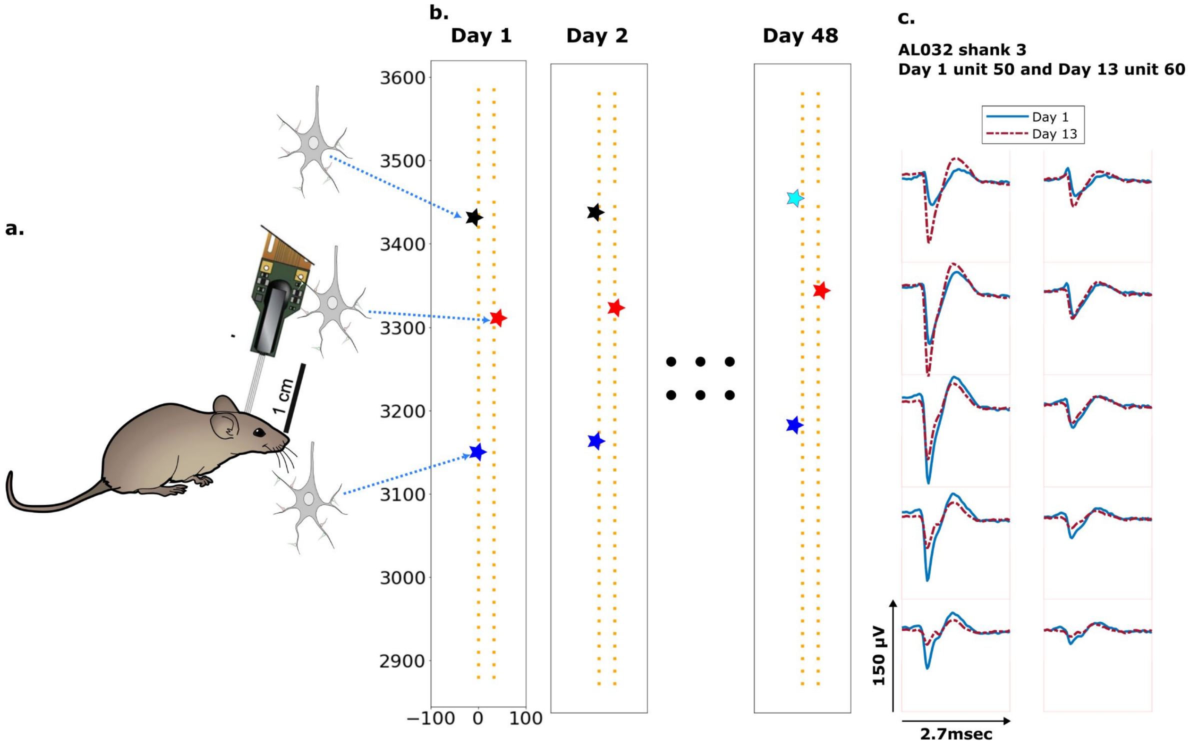

Fig. 1

Schematic depiction of drift:

a. Mice were implanted with a 4-shank Neuropixels 2.0 probe in visual cortex area V1. b. he location of a unit recorded on the probe. In this hypothetical case, the same color indicates unit The black unit is missing on day 48, while the turquoise star is an example of a new unit. Tracking and blue units across all datasets and determine that the black unit is undetected on day 48. c. Two eforms of units recorded in two datasets that likely represent the same neuron, based on similar the average waveform on one channel across 2.7 milliseconds. The blue traces are waveforms on channels (two rows above, two rows below, and one in the same row) from the first dataset (Day ected, are from the second dataset. Waveforms are aligned at the electrodes with peak amplitude, different on the two days.

spike times, etc. are used as input for our method (post-processed using ecephys spike sorting pipeline[29]) (Sec. 4.3). To ensure the sorting results are unbiased, we performed no manual curation. As the clusters returned by Kilosort can vary in quality, we only considered the subset of units labeled as ‘good’ by Kilosort, here referred to as KSgood units (Sec. 4.4). KSgood units are mainly determined by the amount of inter-spike-interval violations and are believed to represent a single unit.[16]

Our overall strategy is to run spike-sorting once per session, and then to generate a unit-by-unit assignment between pairs of datasets. When tracking units across more than two sessions, two strategies are possible: match all ensuing sessions to a single session (e.g., the first session) (Sec. 2.2 and Sec. 4.2), or match consecutive pairs of sessions and then trace matched units through all sessions (Sec. 2.4).

We refer to the subset of KSgood units with strong and distinguishable visual responses in both datasets of a comparison as reference units (See Sec. 4.4 for details). Similar to Steinmetz et al.[7] we validated our unit matching of those reference units using visual receptive field similarity. Finally, we showed that trackable units with strong visual responses are qualitatively similar to those without (***Figure S1*** to ***Figure S5***).

To provide registration between pairs of recordings, we used the Earth Mover’s Distance (EMD).[30], [31] We use a feature space consisting of a geometric distance space and a waveform similarity space, to address both rigid and non-rigid neuron motion. The EMD finds matches between objects in the two distributions by minimizing the overall distances between the established matches (Sec. 4.1.1).

We use EMD in two stages: rigid drift correction and unit assignment. Importantly, the EMD distance incorporates two parameters crucial for matching units: location-based physical distance and a waveform distance metric that characterizes similarity of waveforms (Sec. 4.1.2). The EMD distance matrix is constructed with a weighted combination of the two (details in Sec. 4), i.e. a distance between two units d_{ik} is given by $d_{ik} = d_{location}{}^{ik} * d_{waveformik}$ (***Figure 2*** a). The first EMD stage estimates the homogeneous vertical movement of the entire population of KSgood units (***Figure 2*** b). This movement estimate is used to correct the between-session rigid drift in unit locations. The rigid drift estimation procedure is illustrated in figure 2b. Post drift correction, a unit’s true match will be close in both physical distance and waveform distance. Drift-corrected units were then matched at the second EMD stage. The EMD distance between assigned units can be thought of as the local non-rigid drift combined with the waveform distortion resulting from drift. We test the accuracy of the matching by comparing with reference unit assignments based on visual receptive fields (**Sec. 4.4**).

For each unit, the location is determined by fitting the peak to peak amplitudes on the 10 sites nearest the site with peak signal, based on the triangulation method in[32] (Sec. 4.1.2). The waveform distance is an L2 norm between two spatial-temporal waveforms that spans 22 channels and 2.7 msec (Sec. 4.1.2). Physical unit distances provide a way to maintain the internal structure and relations between units in the EMD. Waveform similarity metrics will distinguish units in the local neighborhood and likely reduce the effect of new and missing units (***Figure S6***).

We analyzed the match assignment results in two ways. First, we compared all subsequent datatsets to dataset 1 using recovery rate and accuracy. We define recovery rate R_{rec} as the fraction of unit assignments by our method that are the same as reference unit assignments established using visual responses (**Sec. 4.4**).

$$P(EMD \mid ref) = \frac{P(EMD \cap ref)}{P(ref)} = \frac{N_{EMD \cap ref}}{N_{ref}} \tag{1}$$

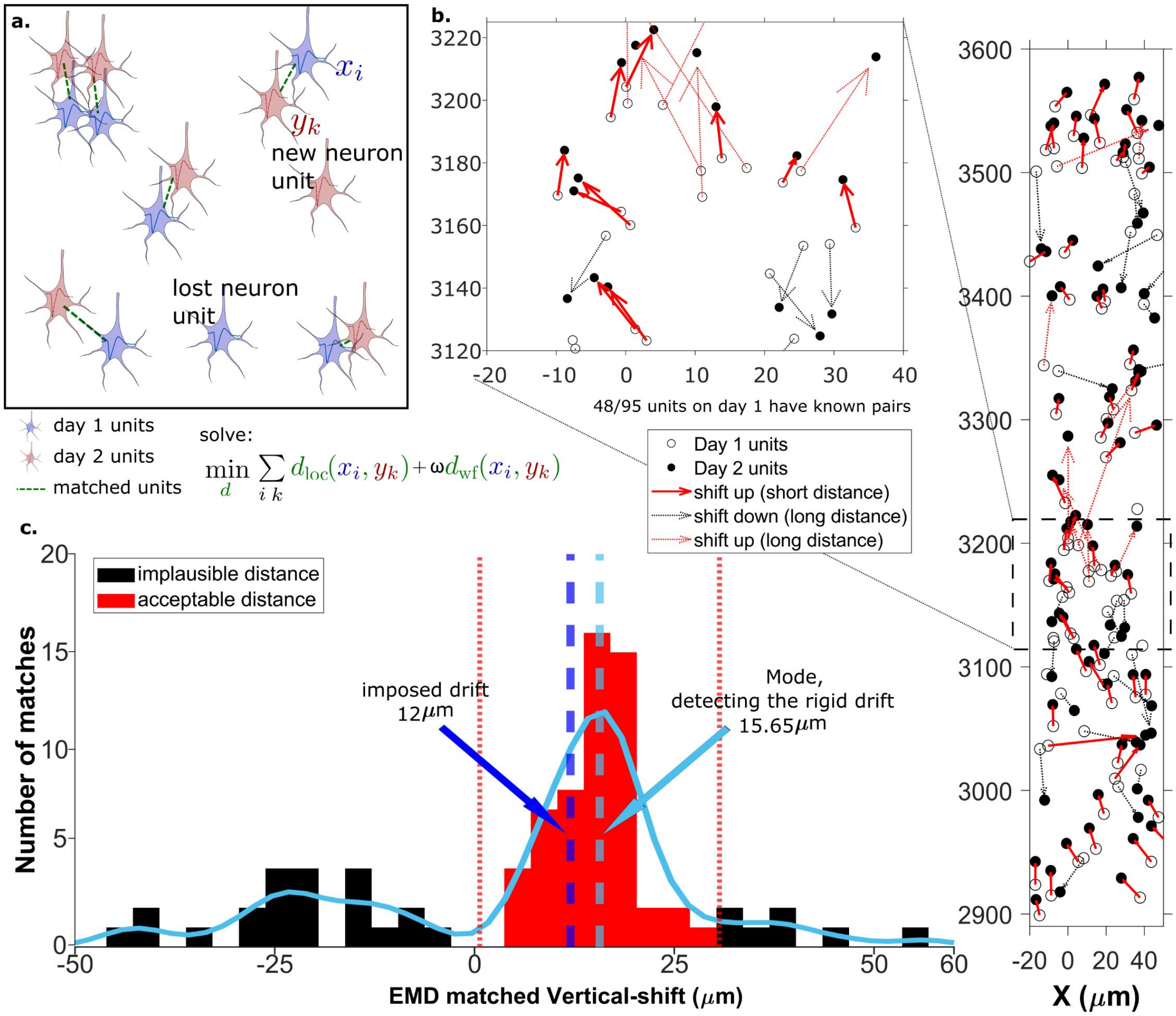

Fig. 2

The EMD can detect the displacement of single units:

a. Schematic of EMD unit matching. Each blue unit in day 1 is. Dashed lines indicate the matches to be found by minimizing the weighted sum of physical and nd filled circles show positions of units in days 1 and 2, respectively. Arrows indicate matching using ts the match direction; upward matches found with the EMD are in red and downward in black. Solid ce within 15μ*m*, while a dashed line indicates a z distance > 15μ*m*. Expanded view shows probe area gram of z-distances of matches (black and red bars) and kernel fit (light blue solid curve). The light ode (d_m = 15.65μ*m*). The dark blue dashed line shows the imposed drift (d_i = 12μ*m*). The red region *m* of the mode. The EMD needs to detect the homogeneous movement against the background, i.e. re unlikely to be the real matches due to biological constraints.

Since the EMD forces all units from the dataset with fewer neurons to have an assigned match, we use vertical z-distance to threshold out the biologically-impossible unit assignments. We then calculated the accuracy R_{acc}, i.e. the fraction of EMD unit assignments within the z-distance threshold which agree with the reference assignments.

$$P((EMD \mid ref) \cap threshold) = \frac{P((EMD \cap ref) \mid threshold)}{P(ref \mid threshold)} \tag{2}$$

We also retrieved non-reference units, i.e. matched units without receptive field information but whose z-distance is smaller than the threshold.

Second, we tracked units between consecutive datasets and summarized and analyzed the waveforms, unit locations, firing rates and visual responses (see ***Figure S1*** to ***Figure S5*** for details) of all tracked chains, i.e. units which can be tracked across at least three consecutive datasets.

2.2 Measuring rigid drift using the EMD

Drift happens mostly along the direction of probe insertion (vertical or z direction). We want to estimate the amount of vertical drift under the assumption that part of the drift is rigid, this is likely a good assumption given the small (≈ 720μm) z-range of these recordings. The EMD allows us to extract the homogeneous (rigid) movement of matched units. For ideal datasets with a few units consistently detected across days, this problem is relatively simple (***Figure 2*** a). In the real data analyzed here, we find that only ≈ 60% of units are detected across pairs of days, so the rigid motion of the real pairs must be detected against a background of units with no true match. These units with no real match will have z-shifts far from the consensus z-shift of the paired units (***Figure 2*** c).

In ***Figure 2*** the EMD match of units from the first dataset (***Figure 2*** b, open circles) to the dataset recorded the next day (***Figure 2*** b, closed circles) is indicated by the arrows between them. To demonstrate detection of significant drift, we added a 12 micron upward drift to the z-coordinate of the units from the second day. The first stage of the EMD is used to find matches using the combined distance metric as described in **section 4.1.2**. We used a kernel fit to the distribution of z-distances of all matched units to find the mode (Mode = 15.65μm); this most probable distance is the estimate of the drift (***Figure 2*** c). It is close to the actual imposed drift (d_i = 12μm).

As the EMD is an optimization algorithm with no biological constraints, it assigns matches to all units in the smaller dataset regardless of biophysical plausibility. As a result, some of the assigned matches may have unrealistically long distances. A distance threshold is therefore required to select correct pairs. For the illustration in ***Figure 2***, the threshold is set to 15μm, which is chosen to be larger than most of the z-shifts observed in our experimental data. The threshold value will be refined later by distribution fitting (***Figure S2***). In ***Figure 2*** all of the sub-threshold (short) distances belong to upward pairs (***Figure 2*** b and **c**, red solid arrows), showing that the EMD can detect the homogeneous movement direction and the amount of imposed drift.

When determining matched reference units from visual response data, we require that units be spatially nearby (within 30μm) as well as having similar visual responses. After correcting for drift, we find that we recover more reference units (***Figure S7***), indicating improved spatial match of the two ensembles. This improved recovery provides further evidence of the success of the drift correction.

2.3 A vertical distance threshold is necessary for accurate tracking

To detect the homogeneous z-shift of correct matches against the background of units without true matches, it is necessary to apply a threshold on the z-shift. When tracking units after shift correction, a vertical distance threshold is again required to determine which matches are reasonable in consideration of biological plausibility. The Receiver Operator Characteristic(ROC) curve in ***Figure 3*** shows the fraction of reference units matched correctly and the number of reference pairs retained as a function of z-distance threshold. We want to determine the threshold that maximizes the overall accuracy in the reference units (***Figure 3***, blue curve) while including as many reference units as possible (***Figure 3***, red curve).

Since reference units only account for 29% of KSgood units (units with few inter-spike-interval violations that are believed to represent a single unit), and the majority of KSgood units did not show a distinguishable visual response, we need to understand how representative the reference units are of all KSgood units.

We found the distribution of z-distances of reference pairs is different from the distribution of all KSgood units (***Figure 4***a, top and middle panel). While both distributions may be fit to an exponential decay, the best fit decay constant is significantly different (Kolmogorov-Smirnov test, reject H0, $p = 5.5 \times 10^{-31}$). Therefore, the accuracy predicted by the ROC of reference pairs in **Figure 3** will not apply to the set of all KSgood pairs. The difference in distribution is likely due to the reference units being a special subset of KSgood units in which units are guaranteed to be found in both datasets, whereas the remaining units may not have a real match in the second dataset. To estimate the ROC curve for the set of all KSgood units, we must estimate the z-distance distribution for a mixture of correct and incorrect pairs.

We assume that the distribution of z-distances $P(\Delta)$ for reference units is the conditional probability $P(\Delta \mid H)$; that is, we assume all reference units are true hits. The distribution of z-distances for all KSgood units $P(\Delta)$ includes both hits and false positives. The distance distribution of false positives is the difference between the two (**Sec. 8.4**, ***Equation 6***).

A Monte Carlo simulation determined that the best model for fitting the z-distance distribution of reference units $P(\Delta \mid H)$ is a folded Gaussian distribution (***Figure 4***a, middle panel) and an exponential distribution for false positive units. The KSgood distribution is a weighted combination of the folded Gaussian and an exponential:

$$P(AllUnits) = f * P(FoldedGaussian) + (1 - f) * P(Exponential) \tag{3}$$

We fit the KSgood distribution to ***Equation 3*** to extract the individual distribution parameters and the fraction of true hits (f). The full distribution can then be integrated up to any given z-threshold value to calculate the false positive rate. (***Figure 4***a, top panel, see **Sec. 8.4** for details).

Based on the the estimated false positive rate (***Figure 4***a, bottom panel), we used a threshold of $10\mu m$ (***Figure 3***, black dotted line) to obtain at least 70% accuracy in the KSgood units. We used the same threshold to calculate the number of matched reference units and the corresponding reference unit accuracy (***Figure 4***b, green bars).

Note that this threshold eliminates most of the known false positive matches of reference pairs (***Figure 4***b, red fraction) at the cost of recovering fewer correct pairs (***Figure 4***b, green bars). The recovery rate varies from day to day; datasets separated by longer times tend to have higher tracking uncertainty (***Figure S10***).

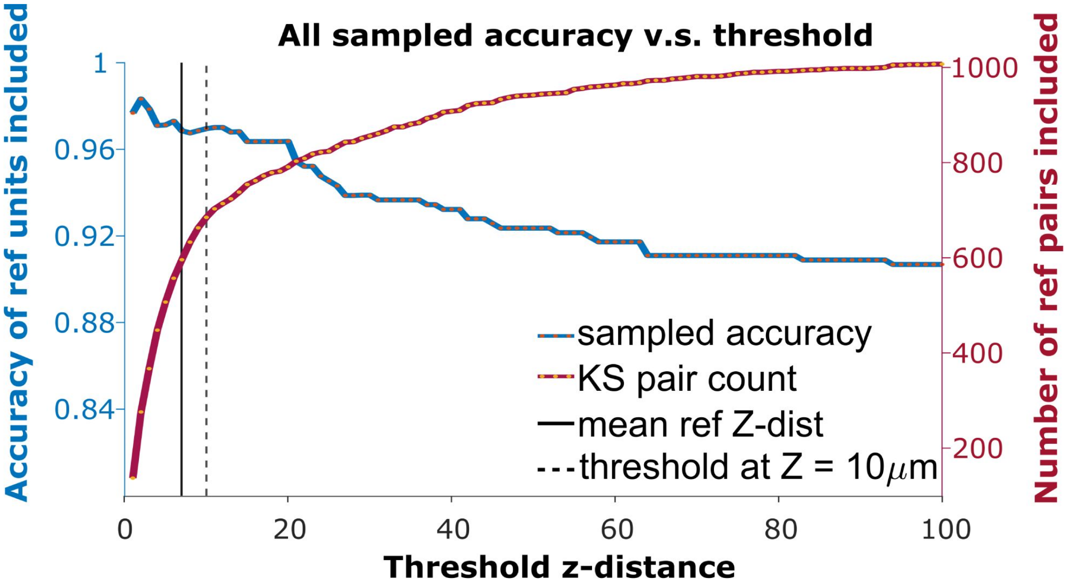

Fig. 3

The ROC curve of matching accuracy vs. distance.

The blue curve shows the accuracy for reference units. The red line nce units included. The solid vertical line indicates the average z distance across all reference pairs dashed vertical black line indicates a z-distance threshold at z = 10μ*m*.

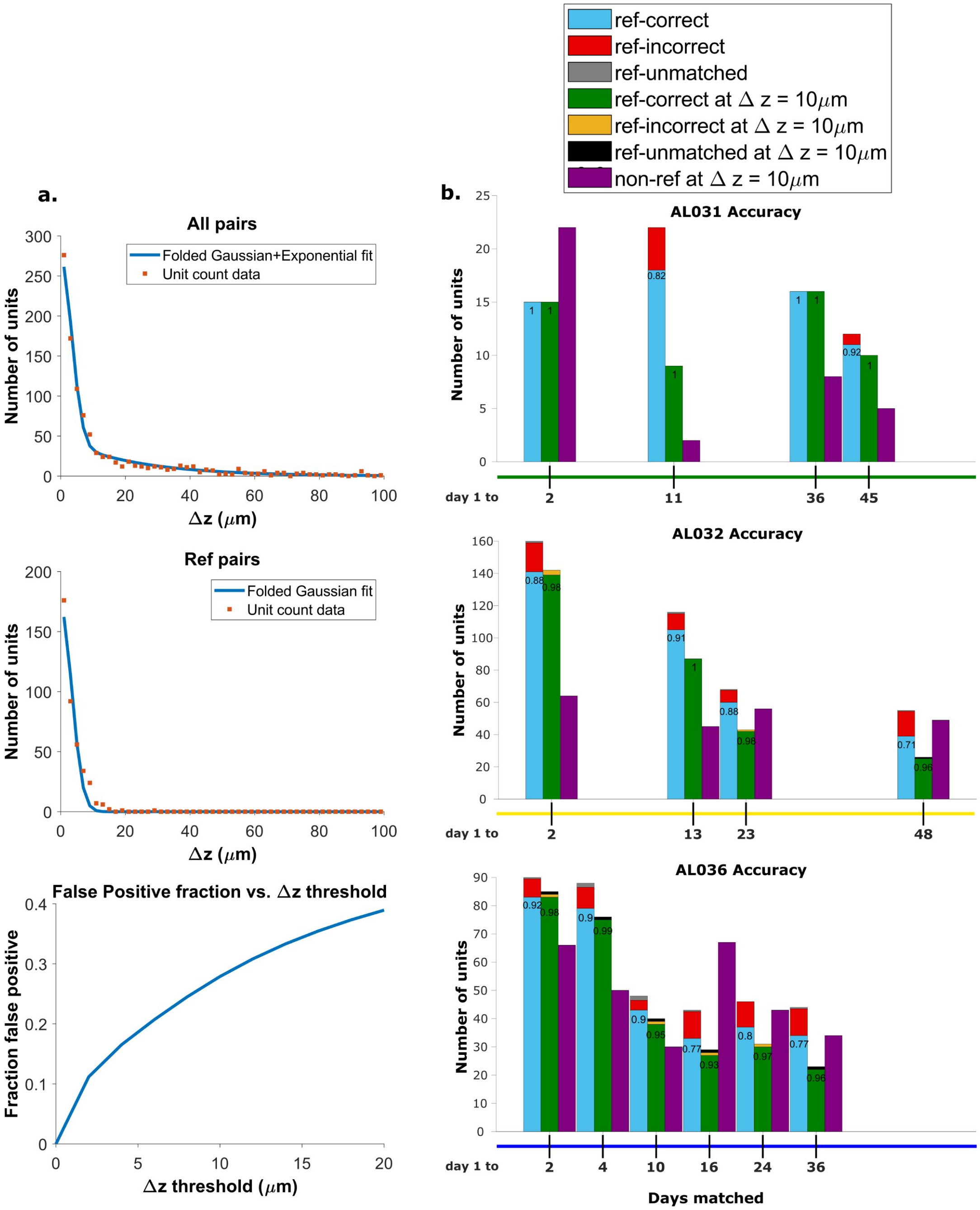

Fig. 4

Recovery rate, accuracy and putative pairs:

a. The histogram distribution fit for all KS-good units (top) and reference units alone (middle). False positives for reference units are defined as units matched by EMD but not matched when using receptive fields. The false positive fraction for the set of all KSgood units is obtained by integration. z = 10μ*m* threshold has a false positive rate = 27% for KSgood units. b. Light blue bars represent the number of reference units successfully recovered using only unit location and waveform. The numbers on the bars are the recovery rate of each datatset, and the red portion indicates incorrect matches. Incorrect matches are cases where units with a known match from receptive field data are paired with a different unit by EMD; these errors are false positives. The green bars show matching accuracy for the set of pairs with z-distance less than the 10μ*m* threshold. The orange portion indicates incorrect matches after thresholding. The false positives are mostly eliminated by adding the threshold. Purple bars are the number of putative units (unit with no reference information) inferred with z-threshold = 10μ*m*.

In addition to the units with visual response data, we can track units which have no significant visual response (***Figure 4***b, purple bars). All comparisons are between subsequent datasets and the day 1 dataset.

2.4 Units can be tracked in discontinuous recordings for 48 days

To assess long-term tracking capabilities, we tracked neurons across all datasets for each mouse. ***Figure 5*** shows a survival plot of the number of unit chains successfully tracked over all durations. All units in the plot can be tracked across at least three consecutive datasets, a chain as the term is used here. We categorized all trackable unit chains into three types: reference chains, mixed chains and putative chains. Reference chains have receptive field information in all datasets. Putative chains have no reference information in any of the datasets. Mixed units have at least one dataset with no receptive field information. There are 133 reference chains, 135 mixed chains and 84 putative chains across all the subjects. Among them, 46 reference, 51 mixed, and 9 putative units can be followed across all datasets. We refer to them as fully trackable units. One example trackable unit in each group is shown in ***Figure 6***, ***Figure S16***, and ***Figure S17***.

We hypothesize that the three groups of units are not qualitatively different from each other, that is, all units are equally trackable. In order to check for differences among the three groups, we analyzed the locations, firing rates, waveforms, and receptive fields of the fully trackable units in the three groups: reference, putative, and mixed.

The spatial-temporal waveform similarity is measured by the L2 distance between waveforms (**Sec. 4.1.2**). A Kruskal-Wallis test is performed on the magnitude of L2 change between all pairs of matched waveforms among the three groups. There is no statistical difference in the waveform similarity in reference, putative, and mixed units (H = 0.59, p = 0.75) (***Figure S1***). There is no significant difference in the physical distances of units per dataset (H = 1.31, p = 0.52) (***Figure S2***, bottom panel), nor in the location change of units (H = 0.23, p = 0.89) (***Figure S2***, top panel).

Firing rate is characterized as the average firing rate fold change of each unit chain, with firing rate of each unit in each dataset normalized by the average firing rate of that dataset. There is no difference in the firing rate fold change in the three groups of units (H = 1, p = 0.6) (***Figure S3***).

The receptive field similarity between units in different datasets is described by visual fingerprint (vfp) correlation and Peristimulus Time Histogram (PSTH) correlation between units, and the similarity score, the sum of the two correlations (**Sec. 4.4**). The change in vfp between matched units is similar among the three groups (H = 2.23, p = 0.33). Similarly, the change in PSTH is not different among the three groups (H = 1.61, p = 0.45) (***Figure S4***).

3 Discussion

We present here an EMD-based neuron tracking algorithm that provides a new, automated way to track neurons over long-term experiments to enable the study of learning and adaptation with state-of-the-art high density electrophysiology probes. We demonstrate our method by tracking neurons up to 48 days without using receptive field information. Our method achieves 90% recovery rate on average for neurons separated up to one week apart and 78% on average for neurons five to seven weeks apart (***Figure 4***b, blue bars). We also achieved 99% accuracy up to one week apart and 95% five to seven weeks apart, when applying a threshold of 10 μm (***Figure 4***b, green bars). It also retrieved a total of 552 tracked neurons with partial or no receptive field information, 12 per pair of datasets on average. All the fully trackable unit chains were evaluated by waveforms and estimated locations. Our method is simple and robust; it only requires spike sorting be performed once, independently, per dataset. In order to be more compatible and

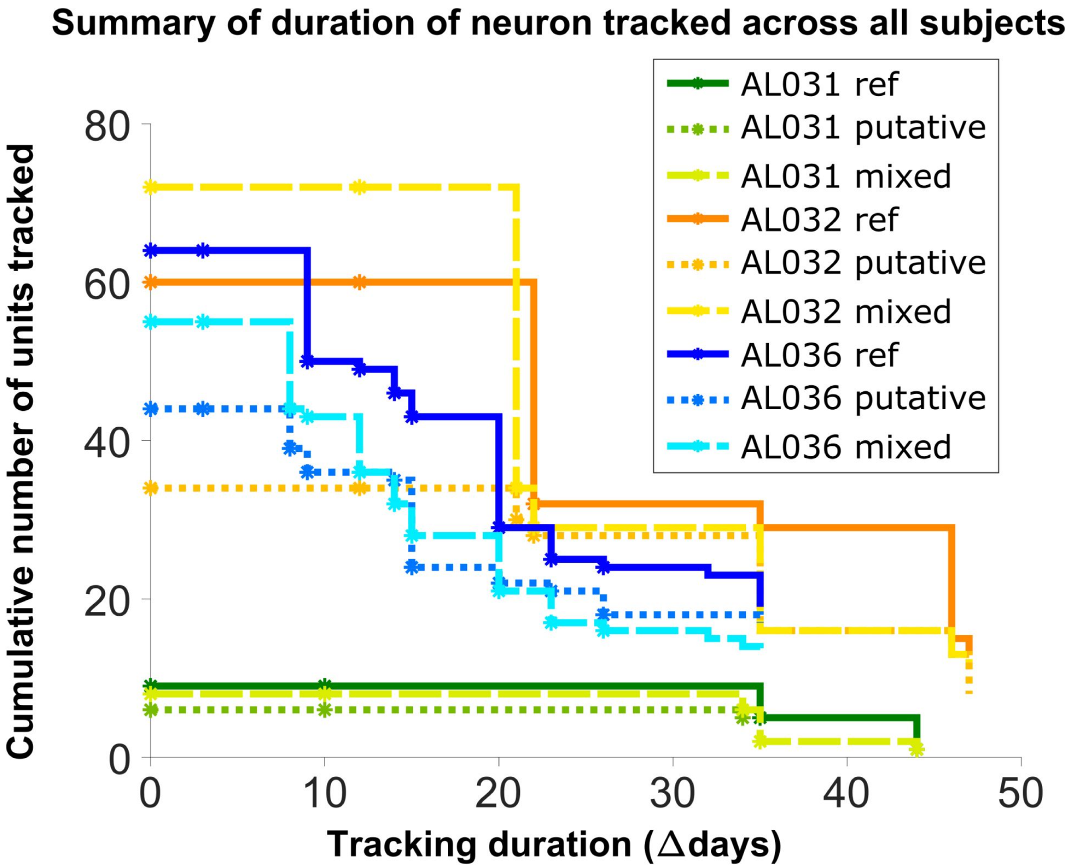

Fig. 5

Number of reference units (deep blue, dark orange and green for different subjects), putative (medium green, medium orange and blue) units, and mixed units (light green, yellow, and light blue) tracked for different durations. The loss rate is similar for different chain types in the same subject. Note that chains can start on any day in the full set of recordings, so the different sets of neurons have chains with different spans between measurements.

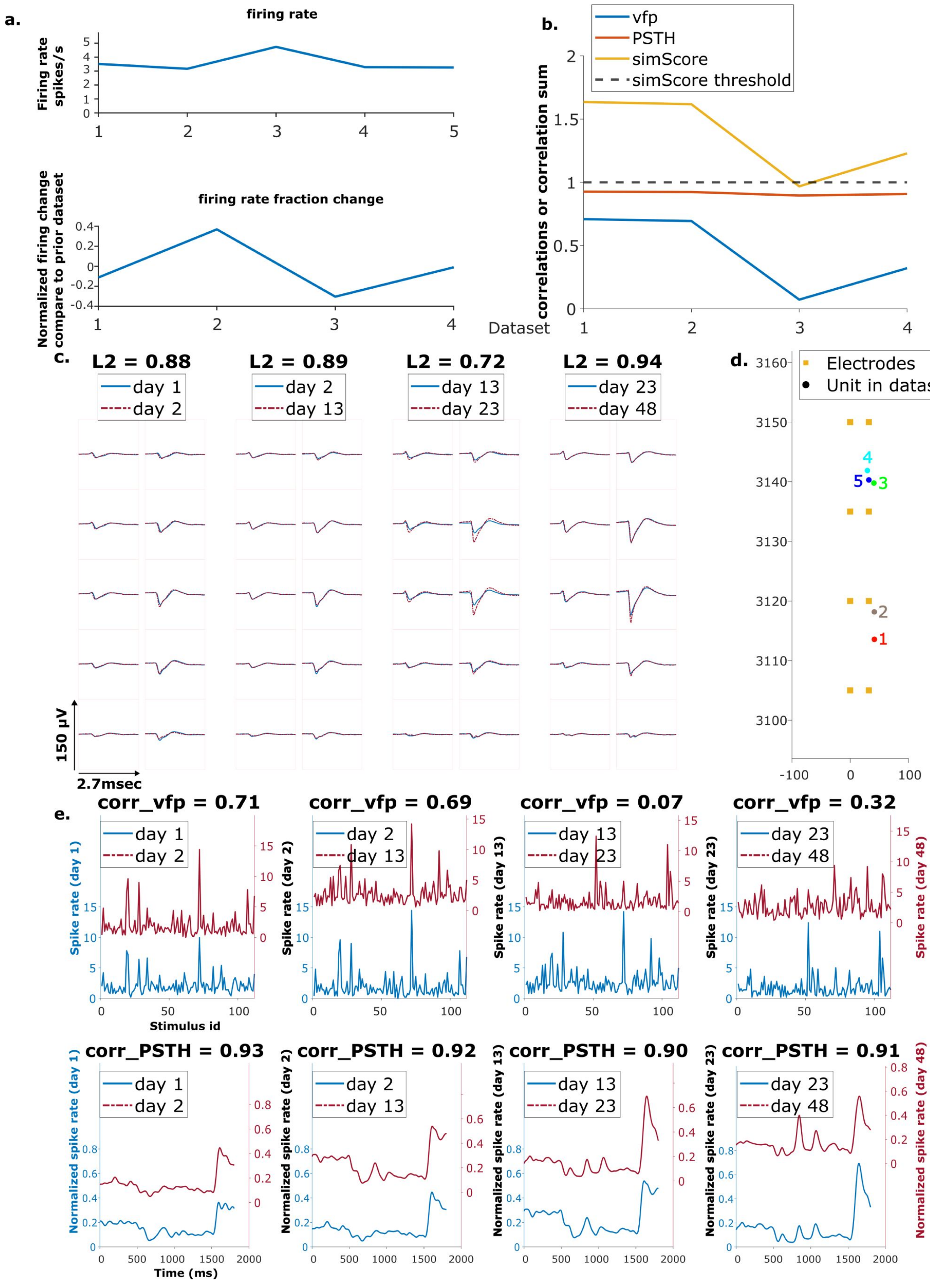

Fig. 6

Example mixed chain:

a. Above: Firing rates of this neuron on each day (Day 1, 2, 13, 23, 48). Below: Firing rate fractional change compared to the previous day. b. Visual response similarity (yellow line), PSTH correlation (orange line), and visual fingerprint correlation (blue line). The similarity score is the sum of vfp and PSTH. The dashed black line shows the threshold to be considered a reference unit. c. Spatial-temporal waveform of a trackable unit. Each pair of traces represents the waveform on a single channel. d. Estimated location of this unit on different days. Each colored dot represents a unit on one day. The orange squares represent the electrodes. e. The pairwise vfp and PSTH traces of this unit.

generalizable with existing sorting methods, we chose Kilosort, one of the most widely used spike sorting methods.[33], [34] We show the capability of our method to track neurons with no specific tuning preference (***Figure S16***).

The method includes means to identify dataset pairs with very large drift. In our data, we can detect large drift because such datasets have very few reference units, and significantly different EMD cost (**Sec. 8.6**). For example, datasets 1 and 2 in animal AL036 have very few reference units compared to other datasets (see ***Figure S11***, AL036). This observation is consistent with the overall relationship between the EMD cost and recovery rate (***Figure S12***). Datasets with higher cost tend to have lower unit recovery rate and higher variation in recovery rates. Therefore, these two datasets were excluded in the tracking analysis.

Our validation relies on identifying reference units. The reference unit definition has limitations. The similarity score is largely driven by PSTHs (***Figure 6***, ***Figure S11***), the timing of stimulus triggered response, rather than vfp, the response selectivity. As a result, a single neuron can be highly correlated, i.e. similarity score greater than 1, with more than 20 other neurons. For example, in subject AL032 shank 2, one neuron on day 1 has 22 highly correlated neurons on day 2, 4 of which are also within the distance of 30μm. Non-reference units may also have very similar visual responses: we note that 33 (5 putative neurons and 28 mixed neurons) out of 106 trackable neurons have a similarity score greater than 1 even for days with no reference unit assignment. Coincidentally similar visual responses could potentially contribute to inaccurate assignment of reference units and irregularity in trackable unit analysis. These errors would reduce the measured accuracy of the EMD matching method; since the accuracy is very high (***Figure 4***), the impact of mismatches is low.

We note that the ratio of reference units over KSgood units decreases as recordings are further separated in time (***Figure S13***). This reduction in fraction of reference units might be partially due to representational drift as well as the fact that the set of active neurons are slightly different in each recording. The visual fingerprint similarity of matched neurons decreased to 60% after 40 days (see reference 7 supplement).

We developed the new tracking algorithm based on an available visual cortex dataset, and used a prominent sorting algorithm (Kilosort 2.5) to spikesort the data. We had reference data to assess the success of the matching and tune parameters. Applying our algorithm in other brain areas and with other sorters may require parameter adjustment. Evaluation of the results in the absence of reference data requires a change to the fitting procedure.

The algorithm has only two parameters: the weighting factor ω that sets the relative weight of waveform distance vs. physical distance, and the z-distance threshold that selects matches that are likely correct. We found that recovery rate, and therefore accuracy, is insensitive to the value of ω for values larger than 1500, so this parameter does not require precise tuning. However, the false positive rate is strongly dependent on the choice of z-distance threshold.

When reference information (unit matches known from receptive fields or other data) is available, the procedure outlined in **section 8.4** can be followed. In that case, the distribution of z-distances of known pairs is fit to find the width of the distribution for correct matches. That parameter is then used in the fit of the z-distance distribution of all pairs to ***Equation 3***. Integrating the distributions of correct and incorrect pairs yields the false positive rate vs. z-distance, allowing selection of a z-distance threshold for a target false positive rate.

In most cases, reference information is not available. However, the z-distance distributions for correct and incorrect pairs can still be estimated by fitting the distribution of all pairs. In section 8.4, ***Figure S9*** we show the results of fitting the z-distribution of all pairs without fixing the width of the distribution of correct matches. The result slightly underestimates this width, and the estimated false positive rate increases. This result is important because it suggests the accuracy estimate from this analysis will be conservative. We detail the procedure for fitting the z-distance distribution Methods section (Alg. 2).

As suggested in Dhawale et al.,[5] discontinuous recordings will have more false positives. Improving spike sorting and restricting the analysis to reliably sorted units will help decrease the false positive rate. Current spike sorting methods involve fitting many parameters. Due to the stochastic nature of template initialization, only around 60% to 70% units are found repeatedly in independently executed analysis passes. This leads to unpaired units which decreases EMD matching accuracy. Future users may consider limiting their analysis to the most reliably detected units for tracking; requiring consensus across analysis passes or sorters is a possible strategy. Finally, more frequent data acquisition during experiments will provide more intermediate stages for tracking and involves smaller drift between consecutive recordings.

4 Methods

Our neuron tracking algorithm uses the Earth Mover's Distance (EMD) optimization algorithm. The minimized distance is a weighted combination of physical distance and 'waveform distance': the algorithm seeks to form pairs that are closest in space and have the most similar waveforms. We test the performance of the algorithm by comparing EMD matches to reference pairs determined from visual receptive fields (Sec. 4.4). We calculate two performance metrics. The 'recovery rate' is the percentage of reference units that are correctly matched by the EMD procedure. The 'accuracy' is the percentage of correctly matched reference units that pass the z-distance threshold (***Figure 4***a). 'Putative units' are units matched by the procedure which do not have reference receptive field information. 'Chains' are units that can be tracked across at least three consecutive datasets. The full procedure is summarized in **Algorithm 1**.

Algorithm 1 Neuron Matching Procedure Algorithm 1 Neuron Matching Procedure

Input: channel map, unit cluster label, cluster mean waveforms(with K_{loc} = 2 and K_{wf} = 5 rows and K_{col} = 2 columns of channels), and spike times

Step 1 Estimate unit locations

Estimate background amplitude for each unit

```
for all KSgood units $u_n \in U$ do
    if peak-top-peak voltage $V_{ptp} > 60\mu V$ then
        Get $u_n$'s waveform on channels $C_m$
        Get the peak-to-peak amplitudes $V_{ptp_c}$ of $u_n$ background-subtracted waveforms on channels
          $C_{u_n} = \{mc_{u_n} - k_{loc}, ..., mc_{u_n} + k_{loc}\}$ where $mc_{u_n}$ is the peak channel
        Estimate the neuron's 3D location as in:[32]
          $f(x, y, z) = \sum_{c \in Cu_n} (V_{ptp_c} - \frac{1}{\sqrt{(x-x_c)^2+(z-z_c)^2+y^2}})^2$ where x, z, and y are the horizontal location,
          vertical location, and distance of the unit from the probe, respectively.
        Find an estimate of the global minimizer of $f, x_{u_n}, y_{u_n}, z_{u_n}$ using least-squares optimization
    end
end
Step 2 Compute waveform similarity metrics
for waveforms $wf_{xi} \in U_{N1}$ and $wf_{yk} \in U_{N2}$ where $U_{N1}, U_{N2}$ are the set of all units in the two datasets
  do
    Centered at peak channel $mc_{xi}$ and $mc_{yk}$, respectively
    Get the sets of channels for each unit: $C_{u_n} = \{mc_{u_n} - k_{wf}, ..., mc_{u_n} + k_{wf}\}$
    There are $K_{wf} * 2 * K_{col} + 2 = 22$ channels for each unit
    Compute the waveform similarity metric as $(1/22) * \sum_{c \in Cu_{xi}, Cu_{yk}} L2(wf_{xi} -
      wf_{yk}) / max(L2(wf_{xi}), L2(wf_{yk}))$ for each of the 22 channels
end
```

Step 3 Between-session drift correction

Run the EMD with distances in physical and waveform space

Estimate z-distance mode of all matched pairs with Gaussian kernel fit Apply correction on physical distances of all units ε $U_2 : z_{corr} = z - z_{mode}$

Step 4 Unit matching

Run the EMD with corrected physical distance and waveform metrics

Set z-distance threshold to select unit pairs likely to be the same neuron

Output: cost $\Sigma\, d_{EMD}$, unit assignments

4.1 Algorithm

4.1.1 Earth Mover's Distance

The EMD is an optimization-based metric developed in the context of optimal transport and measuring distances between probability distributions. It frames the question as moving dirt, in our case, units from the first dataset, into holes, which here are the neural units in the second dataset. The distance between the "dirt" and the "holes" determines how the optimization program will prioritize a given match. Specifically, the EMD seeks to minimize the total work needed to move the dirt to the holes, i.e., neurons in day 1 to day 2, by solving for a minimum overall effort, the sum of distances.[30], [31]

$$\begin{aligned} \min_{d_F} \quad & \sum_{i\,k} D(x_i, y_k), \, where \;\, D = d_{loc} + \omega d_{wf} \\ \text{subject to} \quad & f_{ik} \in [0,1] \;\forall i, k \\ & \sum_k (f_k) \le length(Y) \\ & \sum_i (f_i) \le length(X) \\ & \sum(F) = \min(\sum X, \sum Y) \end{aligned} \tag{4}$$

in which $d_{loc} \in \square D^3$ is the three-dimensional physical distance between a unit from the first dataset x_i, and a unit from the second dataset y_k. $d_{wf} \in \square D^1$ is a scalar representing the similarity between waveforms of units x_i and y_k. ω is a weight parameter that was tuned to maximize the recovery rate of correctly matched reference units. F is the vector of matched objects between the two datasets (See ***Figure S14*** for details about selecting weight).

The EMD has three benefits:

- It allows combining different types of information into the'distance matrix' to characterize the features of units.
- The EMD can detect homogeneous movement of units (***Figure 2*** c), thus providing a way for rigid drift correction, as described in **section 4.1.3**.
- By minimizing overall distances, the EMD has tolerance for imperfect drift correction, error in the determination of unit positions, and possible non-rigid motion of the units.

However, since the EMD is an optimization method with no assumptions about the biological properties of the data, it makes all possible matches. We therefore added a threshold on the permissible z-distance to select physically plausible matches. Supplement ***Figure S14*** shows the recovery rate change as a function of weight parameters to combine neuron location and waveform metrics into a distance matrix.

4.1.2 Calculating the EMD distance metric

The unit locations are estimated by fitting 10 peak-to-peak (PTP) amplitudes from adjacent electrodes and the corresponding channel positions with a 1/R distance model.[32] Unlike Boussard, et al.,[32] we operate on the mean waveforms for each unit rather than individual spikes. We found using the mean waveform yields comparable results and saves significant computation time. Unit locations are three-dimensional coordinates estimated relative to the probe, where the location of the first electrode on the left column at the tip is considered the origin. The mean waveform is computed by averaging all the spike snippets assigned to the cluster by KS 2.5.

For 10 channels $c \in C_{un}$, find the location coordinates x_{un}, y_{un}, z_u that minimizes the difference between measured amplitudes V_{PTP} and amplitudes estimated with locations $\frac{\alpha}{\sqrt{(x-x_c)^2+(z-z_c)^2+y^2}})$:

$$\min \sum_{c \in C_{u_n}} \left(V_{PTP_c} - \frac{1}{\sqrt{(x-x_c)^2 + (z-z_c)^2 + y^2}} \right)^2 \tag{5}$$

The locations are used to calculate the physical distance portion of the EMD distance.

For the waveform similarity metric, we want to describe the waveform characteristics of each unit with its spatial-temporal waveform at the channels capturing the largest signal. The waveform similarity metric between any two waveforms u_{n1} and u_{n2} in the two datasets is a scalar calculated as a normalized L2 metric (see Alg.1 Step 2) on the peak channels, namely the channel row with the highest amplitude and 5 rows above and below (a total of 22 channels). The resulting scalar reflects the 'distance' between the two units in the waveform space and is used to provide information about the waveform similarity of the units. It is used for between-session drift correction and neuron matching. ***Figure 1***c shows an example waveform of a reference unit.

4.1.3 Between-session Drift Correction

Based on previous understanding of the drift in chronic implants, we assumed that the majority of drift occurs along the direction of the probe insertion, i.e. vertical z-direction. This rigid drift amount is estimated by the mode of the z-distance distribution of the EMD assigned units using a normal kernel density estimation implemented in MATLAB. We only included KSgood units.[16] The estimated drift is then applied back to correct both the reference units and the EMD distance matrix by adjusting the z coordinates of the units. A post-correction reference set is compared with the post-correction matching results for validation.

4.2 Determining Z Distance Threshold

Determining the z-distance threshold to achieve a target false positive rate requires estimating the widths of the z-distance distributions of correct and incorrect pairs. If reference data is available, the z-distance distribution of the known correct pairs should be fit to a folded Gaussian as described in 8.4. The width of the folded Gaussian, which is the error in determination of the z-positions of units, is then fixed in the fit of the z-distribution of all pairs found by the algorithm outlined in Algorithm **4**.**1**.**1**. If no reference data is available, the width of the distribution of correct pairs is determined by fitting the z-distance distribution of all pairs to ***Equation 3*** with the folded Gaussian width as one of the parameters. This procedure is detailed in **Algorithm 2**. We show two examples of model fitting without reference information in section ***Figure S9***.

Algorithm 2 Determining an appropriate z distance threshold

Algorithm 2 Determining an appropriate z distance threshold

Input: Z distances of all matched units, target false positive rate, width σ of the z-distance distribution of correct pairs, if available

Step 1 Fit z distance distribution of all pairs to decompose into distributions of correct and incorrect pairs

Fit the z-distance distribution of all pairs to the sum of a folded Gaussian (for correct pairs) and an exponential (for incorrect pairs). If the width σ of the distribution of correct pairs is known from reference data, fix at that value. Otherwise, include in the fit parameters. (See **section 8.4** for details). The functional form is: $P(z) = d(f N e^{-\frac{z^2}{2\sigma^2}} + \frac{1-f}{c} e^{-\frac{z}{c}})$ Where: f = fraction of correct pairs; σ = width of the distribution of correct pairs; c = decay constant of distribution of incorrect pairs; d = amplitude normalization; and $N = \frac{2}{\sigma\sqrt{2\pi}}$, the normalization factor of the folded Gaussian.

Step 2 Determine z threshold to achieve a target false positive rate

For Neuropixels 1.0 and 2.0 probes, the width of the z-distance distribution of correct matches (σ) should be <10 μm; a larger width, or a very small value of the fraction of correct pairs suggests few or no correct matches. In this case, the EMD cost is likely to be large as well (See ***Figure S11***

Animal AL036 first two rows).

For a range of z values, integrate the z-distance distribution of incorrect pairs from 0 to z, and divide by the integral of the distribution of all pairs over that range. This generates the false positive rate vs. z-distance threshold, as shown in ***Figure S9***. (Code available at: *https://github.com/AugustineY07/Neuron_Tracking/tree/main/Pipeline/Plot/Fit*)

Output: σ (uncertainty of position estimation), threshold at the target false positive rate

4.3 Dataset

The data used in this work are recordings collected from two chronically implanted NP 2.0 fourshank probes and one chronically implanted one-shank NP 2.0 probe in the visual cortex of three head fixed mice (***Figure 7***b, see Steinmetz et al.[7] for experiment details). The recordings were taken while 112 visual stimuli were shown from three surrounding screens (data from Steinmetz et al.[7] Supplement Section 1.2). The same bank of stimuli was presented five times, with order shuffled. The 4-shank probes had the 384 recording channels mapped to 96 sites on each shank.

We analyzed 65 recordings, each from one shank, collected in 17 sessions (5 sessions for animal AL031, 5 sessions for animal AL032, and 7 sessions for animal AL036). The time gap between recordings ranges from one day to 47 days (***Figure 7***a), with recording durations ranging from 1917 to 2522 seconds. The sample rate is 30kHz for all recordings. There are a total of 2958 KSgood units analyzed across all animals and shanks, with an average of 56 units per dataset (***Figure 7***d and ***Figure S15***).

4.4 Reference set

To track clusters across days, Steinmetz et al.[7] concatenated two recording sessions and took advantage of the within-recording drift correction feature of Kilosort 2.0 to extract spikes from the two days with a common set of templates. They first estimated the between session drift of each recording from the pattern of firing rate and amplitude on the probe and applied a position correction of an integer number of probe rows (15μm for the probes used). Then two corrected recordings were concatenated and sorted as a single recording. This procedure ensured that the same templates are used to extract spikes across both recordings, so that putative matches are extracted with the same template. A unit from the first half of the recording is counted as the same neuron if its visual response is more similar to that from the same cluster in the second half of the recording than to the visual response of the physically nearest neighbor unit. Using this procedure and matching criteria, 93% of the matches were correct for recordings < 16 days apart, and 85% were correct for recordings from 3-9 weeks (See Steinmetz et al.,[7] **Fig. 4**). In addition, although mean fingerprint similarity decreases for recordings separated by more than 16 days, this decline is only 40% for the same unit recorded from 40 days apart (see Steinmetz et al.[7] Supplement S3). This procedure, while successful in their setting, was limited to the use of integral row adjustments of the data for between-session drift correction and relied on a customized version of Kilosort 2.0. Although up to three recordings can be sorted together, they must come from recording sessions close in time. In addition, a separate spike sorting session needs to be performed for every pair of recordings to be matched, which is time consuming and introduces extra sorting uncertainty.

To find units with matched visual responses, we examine the visual response similarity across all possible pairs. The visual response similarity score follows Steinmetz et al.,[7] and consists of two measurements. 1) The peristimulus time histogram (PSTH), which is the histogram of the firing of a neuron across all presentations of all images, in a 1800 msec time window starting 400 msec before and ending 400 msec after the stimulus presentation. The PSTH is calculated by

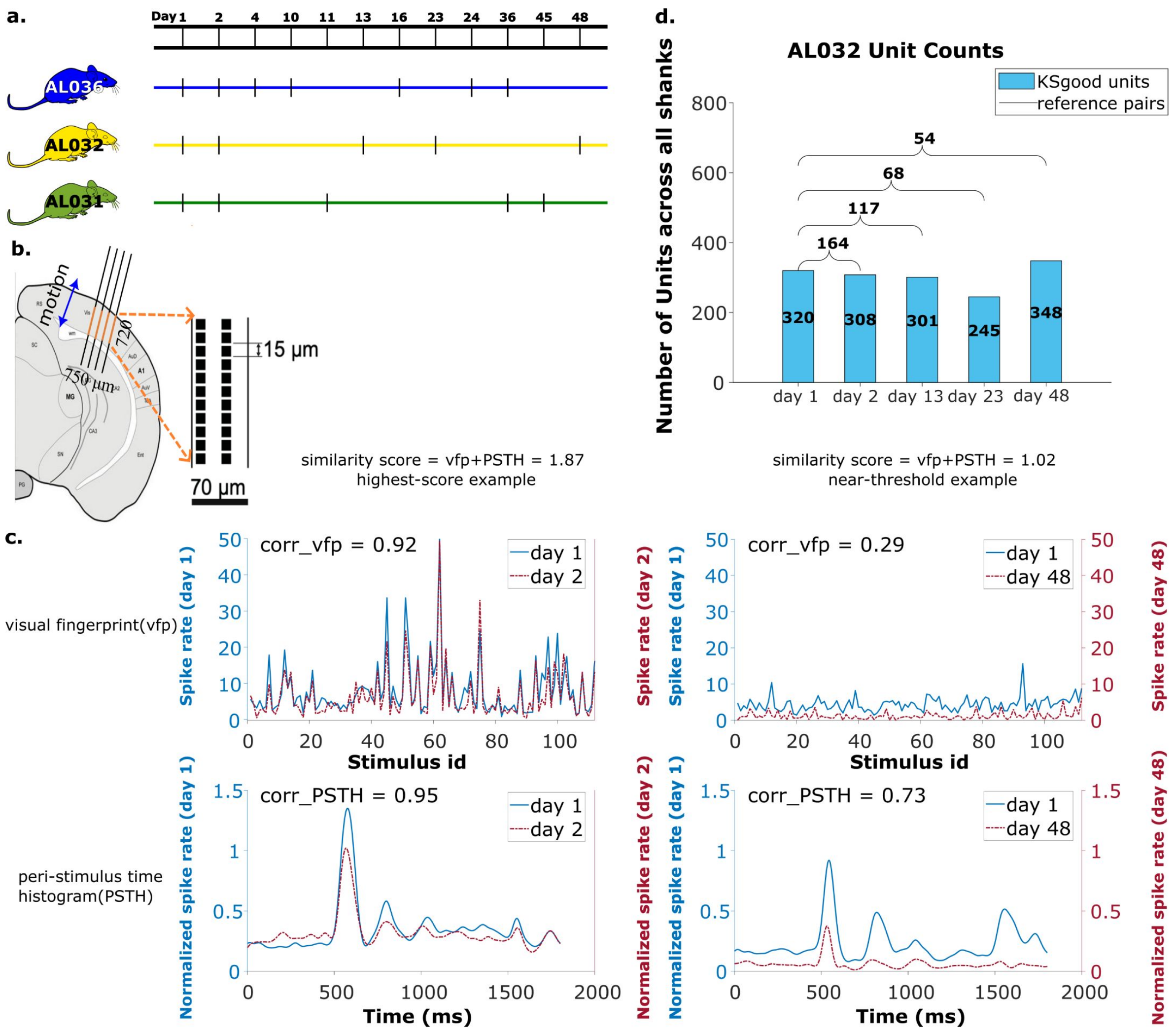

Fig. 7

Summary of dataset:

a. The recording intervals for each animal. A black dash indicates one recording on that day. b. All tex V1 with a 720 µ*m* section of the probe containing 96 recording sites. The blue arrow indicates mples of visual fingerprint(vfp) and peri-stimulus time histogram(PSTH) from a high correlation (left shold (right column) correlation unit. Both vfp and PSTH values vary from [-1,1]. d. Kilosort-good and al AL032, including units from all four shanks.

histrogramming spike times relative to stimulus on time for all stimuli, using 1 ms bins. This histogram is then smoothed with a Gaussian filter. 2) The visual fingerprint(vfp) is the average response of the neuron to each of the 112 images. The vfp is calculated by averaging the spike counts in response to each natural image from the stimulus onset to 1 second afterwards across 5 shuffled trials.

Following Steinmetz et al.,[7] the similarity score between two neurons is the sum of the correlation of the PSTH and the correlation of the vfp across two sessions. The two correlations have values in the range (-1,1), and the similarity score ranges from (-2, 2).

The pool of reference units is established with three criteria: 1) The visual response similarity score of the pair, as described above, is greater than 1 and their physical distance, both before and after drift correction, is smaller than 30µm. We impose the 30 µm threshold on both pre- and post-correction data because the drift is relatively small in our case, and we can reduce false positives by constraining the reference units to be in a smaller region without losing units. In general, one could apply the threshold only on corrected data (after drift correction). 2) A Kruskal-Wallis test is applied on all trials of the vfps to ensure the triggered response to the stimulus is significantly distinguishable from a flat line. 3) Select units from each recording that meet the good criteria in Kilosort. Kilosort assigns a label of either single-unit (good) or multi-unit (MUA) to all sorted clusters based on ISI violations.[16] This step aims to ensure included units are well separated. If there are multiple potential partners for a unit, the pair with the highest similarity score is selected as the reference unit. The complete pool of reference units includes comparisons of all pairs of recordings for each shank in each animal. The portion of units with qualified visual response ranges from 5% to 61%, depending on the time gap between datatets (***Figure S13***). Overall, these reference units made up 29% of all KSgood units (***Figure S15***) across all three animals in our dataset. ***Figure 7*** c shows examples of visual responses from a high similarity reference unit and a reference unit with similarity just above threshold.

5 Code sharing

All code used can be accessed at: *https://github.com/AugustineY07/Neuron_Tracking*.

Acknowledgements

This work was supported supported in part by NIH grant U01 NS115587. We thank Claudia Böhm and Albert Lee for allowing us to use their data in ***Figure S9***.

7 Declaration of interests

The authors declare no competing interests.

8 Supplement

8.1 Trackable units statistics

To show that trackable reference, putative, and mixed units are qualitatively similar, we summarized the median, maximum and minimum change of firing rate, visual receptive field, and location in the box plots in ***Figure S1*** to ***Figure S5***. A Kruskal-Wallis test performed for each feature suggested no difference among the three groups (see **Sec. 2.4** for details).

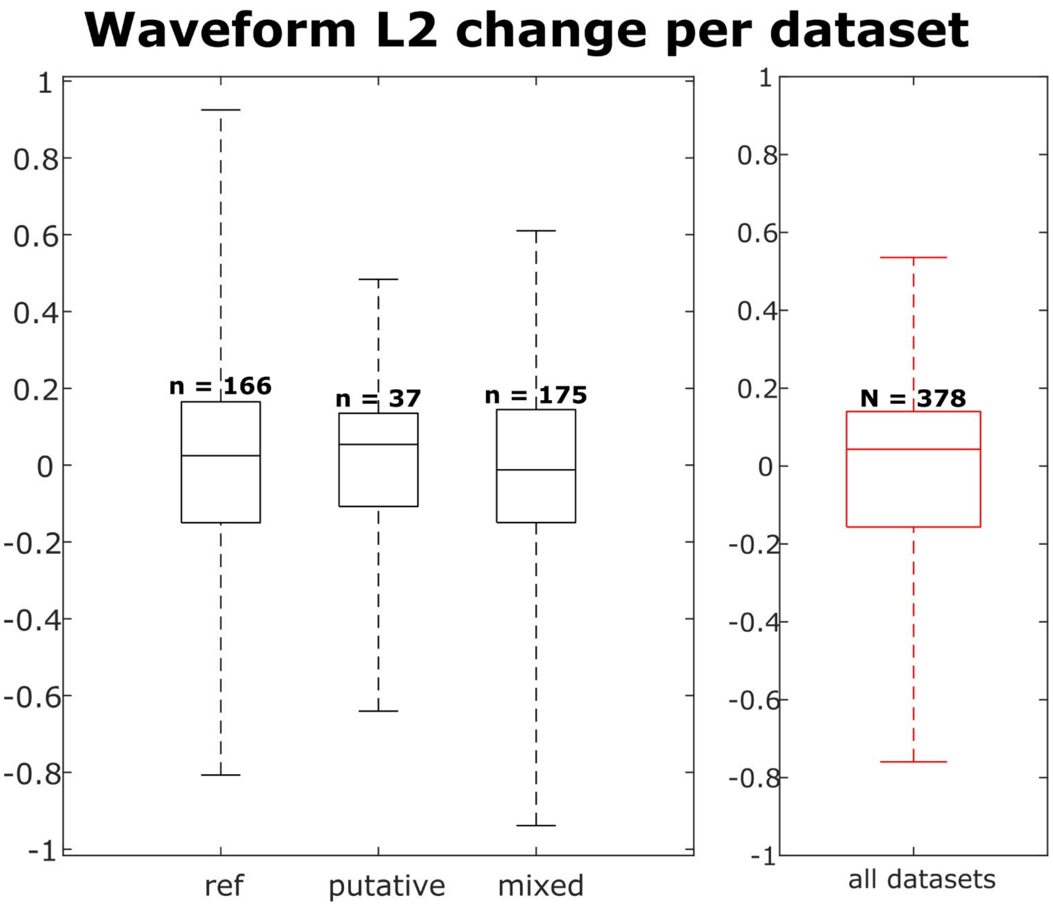

Fig. S1

Distribution of waveform L2 similarity change per dataset for each neuron group and across all neurons.

Box plots ns, and 75% percentile. Whiskers at the ends of the box plot show maximum and minimum values. t comparisons, i.e. (number of units)×(number of datatsets - 1).

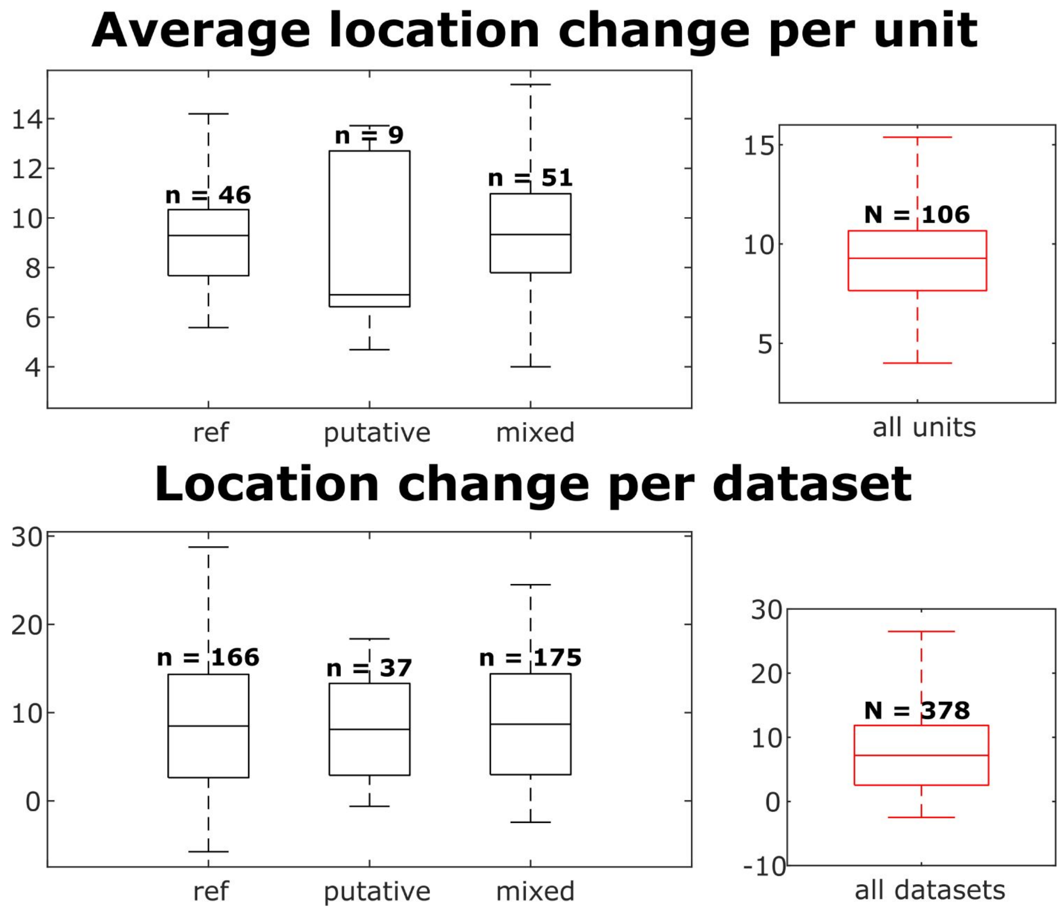

Fig. S2

Distributions of individual unit location changes over whole chains (top) and unit location changes between pairs of datasets (bottom), for each neuron group and across all neurons. Box plots indicate 25% percentile, medians, and 75% percentile. Whiskers at the ends of the box plot show maximum and minimum values. In the top plot, n and N are the number of units. In the bottom plot, n and N are the number of unit comparisons, i.e. (number of units)×(number of datatsets - 1).

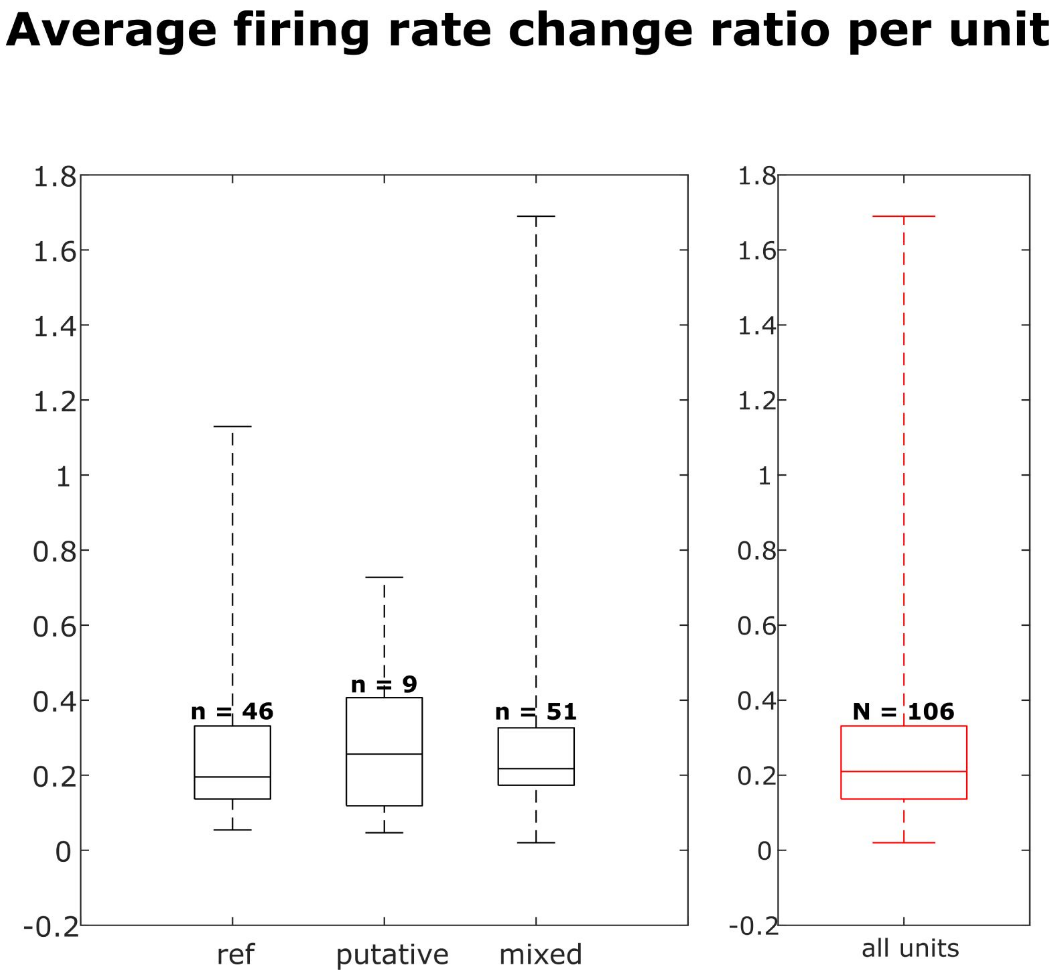
Average firing rate change ratio per unit

Fig. S3

Distribution of firing rate fold change per dataset for each neuron group and across all neurons. Box plots indicate 25% percentile, medians, and 75% percentile. Whiskers at the ends of the box plot show maximum and minimum values. n and N represent the number of units.

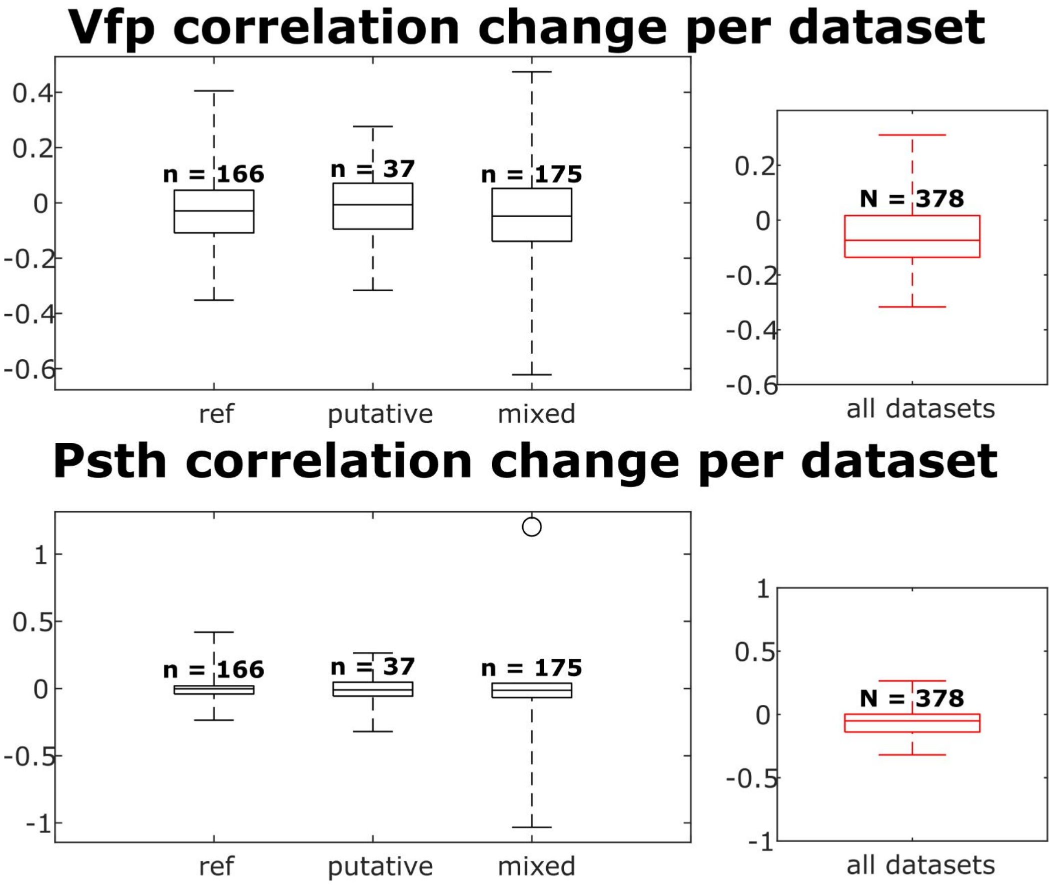

Fig. S4

The visual fingerprint and PSTH change distributions per dataset for each neuron group and across all neurons. Box plots indicate 25% percentile, medians, and 75% percentile. Whiskers at the ends of the box plot show maximum and minimum values. n and N are the number of unit comparisons, i.e.(number of units)×(number of datatsets - 1).

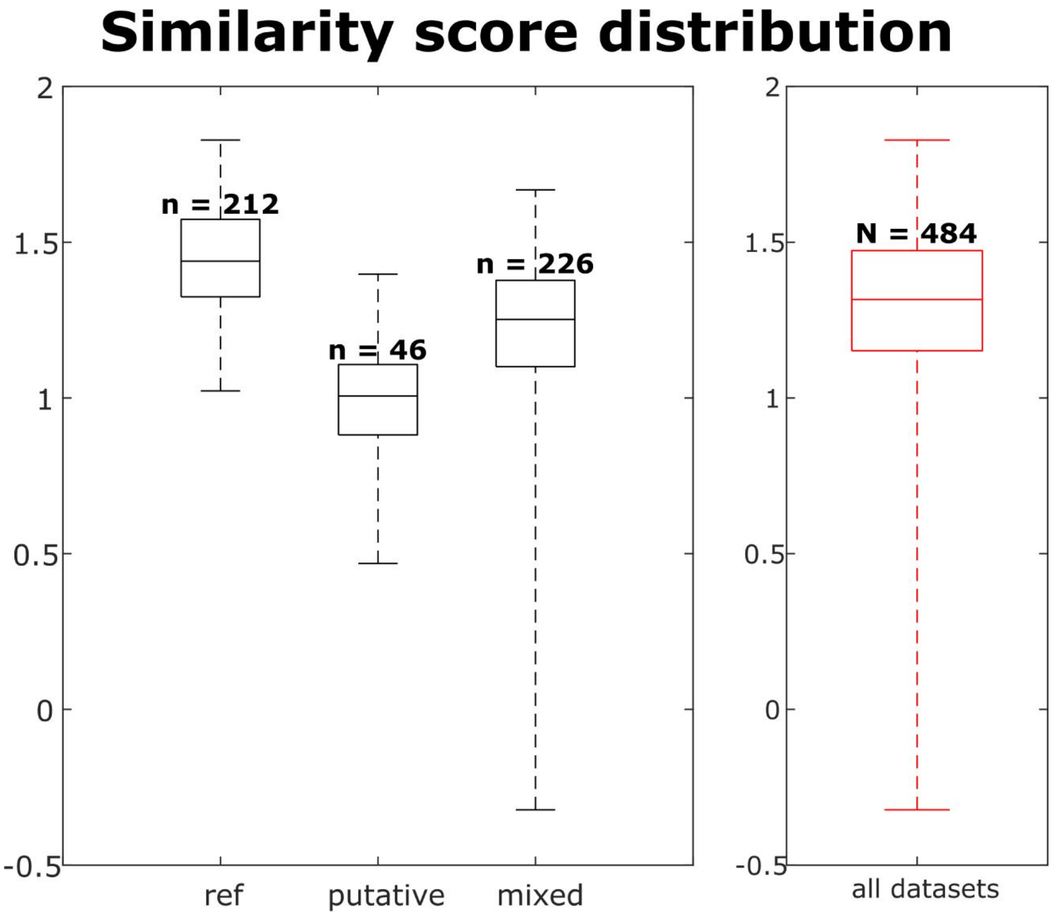

Fig. S5

The similarity score distribution per dataset for each neuron group and across all neurons.

Box plots indicate 25% percentile, medians, and 75% percentile. Whiskers at the ends of the box plot show maximum and minimum values. n and N are the number of observations of the units, i.e. Σ_{units} (observations of this unit)

8.2 Similarity score heatmap

We identify reference pairs as units that are close in space (peak channels separated by < 30μm) and high similarity score (>1). Multiple partners can meet these criteria due to oversplitting – these correspond to blocks of high scores in the heatmap. We only include a unit as a reference if its highest similarity score counterpart in the other dataset is within the 30μm distance threshold.

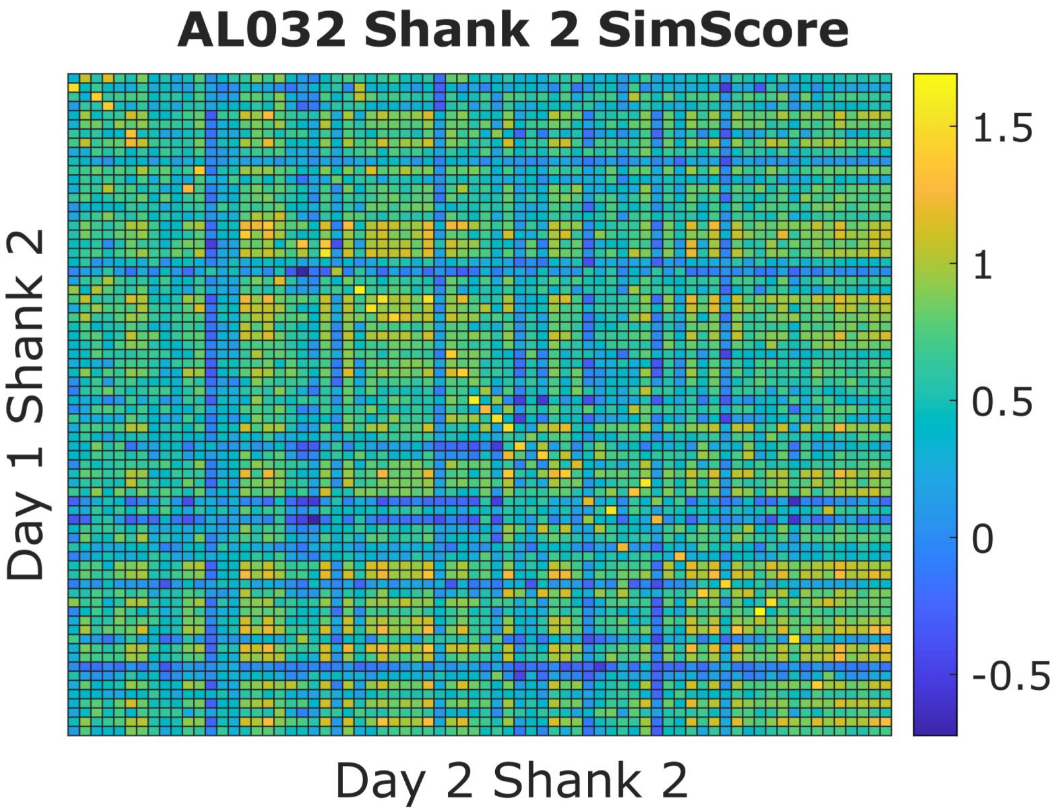

Fig. S6

An example similarity score (vfp + PSTH) heatmap from animal AL032 shank 2 Kilosort-good units between day 1 and 2.

Each small square represents the similarity score (value range from [-2,2]) between one unit from day 1 and one unit from day 2. A warm colored square indicates a higher score. The clusters are ordered by their physical locations on the probe. There is a diagonal line with brighter color blocks, indicating that units with more similar visual responses across days tend to be physically close. This confirms our assumption that neurons are physically stable over time. Also notice that, on each column, there might be more than one bright block in the more distant clusters. We minimize the effect of distant units by constraining the feasible region during selection of reference units. There are also columns without bright yellow blocks; these units do not respond to the stimulus and are not included in the reference set.

8.3 Pre- and post-drift correction reference unit counts

We showed that between-session drift correction improved yield of reference units.

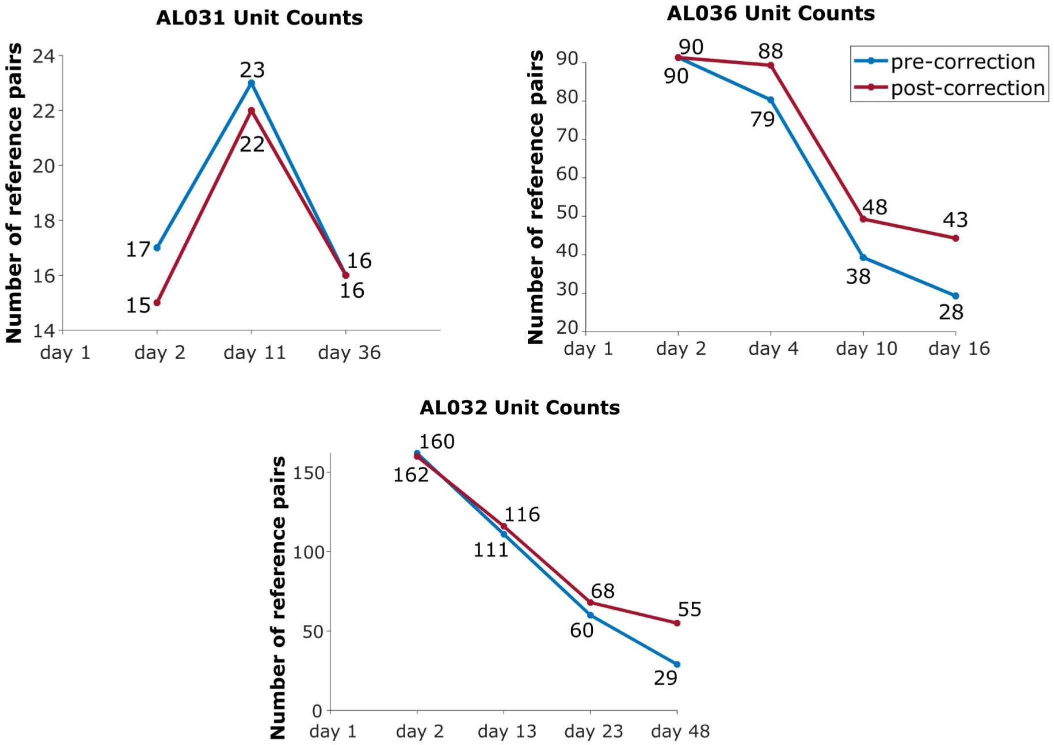

Fig. S7

The effect of drift correction on reference unit yield for all three animals.

Note that drift correction improves the recovery rate for most cases; the degree of improvement is a function of the magnitude of the drift.

8.4 Modeling the z-distance distribution for all units

As shown in ***Figure 4*** a, the z-distance distribution of reference pairs differs significantly from that of all pairs. To estimate the false positive rate for all pairs, we need to account for this difference. We cannot simply extrapolate from the measured false positive rate of the reference units. The difference arises from a bias in the selection of reference units: Because reference units must be detected in two datasets, they must be easily isolated. We created a simple model to determine an appropriate functional form to fit the z-distance distribution of all pairs and estimate the false positive rate.

Assume the following distributions:

1. The z-distance distribution of all matched neurons, i.e. KSgood unit distribution, ($\Delta > 0$) is $P(\Delta)$
2. The z-distance distribution of matched neurons that are true hits (*H*: correct match/hits) is $P(\Delta \mid H)$
3. The z-distance distribution of false positive matched neurons is $P(\Delta \mid\sim H)$
   Let f be the fraction of units with true hits, then the z-distance distribution for all units is:
   $$P(\Delta) = f * P(\Delta \mid H) + (1 - f) * P(\Delta \mid\sim H) \quad (6)$$

To estimate the distribution of $P\,(\Delta \mid H)$, we assume that drift correction works properly. In this case, the z shift between the two units of a reference pair, or any true hit, is due to the error in measuring the position of the unit. The distribution of Δz, which is the absolute value of the z shift, is expected to be a folded Gaussian with μ = 0, and σ = 2*(error in measured z position).

To estimate the distribution of $P\,(\Delta \mid\sim H)$, we performed a Monte Carlo simulation. In the simulation, the number of units is 150, the average density of subject AL036. A fraction f will have real partners in the second dataset. The unit positions in each dataset have normally distributed errors with σ = 5μm, matching the observed distribution of z-distance in the reference units.

To determine a range of values of f (fraction of true hits) that matches the real data, we can estimate probability of a hit in terms of probability of being a reference neuron $P\,(R)$ using Bayes rule

$$P(H) = P(H \mid R)P(R) + P(H \mid\sim R)P(\sim R)$$

P (H | R) can be estimated from the reference units recovery rate 0.86, and P (R) can be estimated from the ratio of reference units, which is 0.29. P (~ R) = 1 − P (R) = 0.73. Then

$$P(H \mid R)P(R) \leq \quad P(H) \leq \quad P(H \mid R)P(R) + P(\sim R) \tag{7}$$

$$0.25 \leq \quad P(H) \leq \quad 0.96 \tag{8}$$

We modeled the distribution at values of f = 0.23, 0.5, 0.6, 0.7 and 0.96. For each value of f, we generate 500 datasets, and compile the z-distance distributions for H and $\sim H$, from the EMD solution. From these simulations, we learned that the false positive distribution is well fit by an exponential decay. Therefore, the z-distance distribution for all units is the sum of the two, as shown in ***Equation 3*** and Alg. 2.

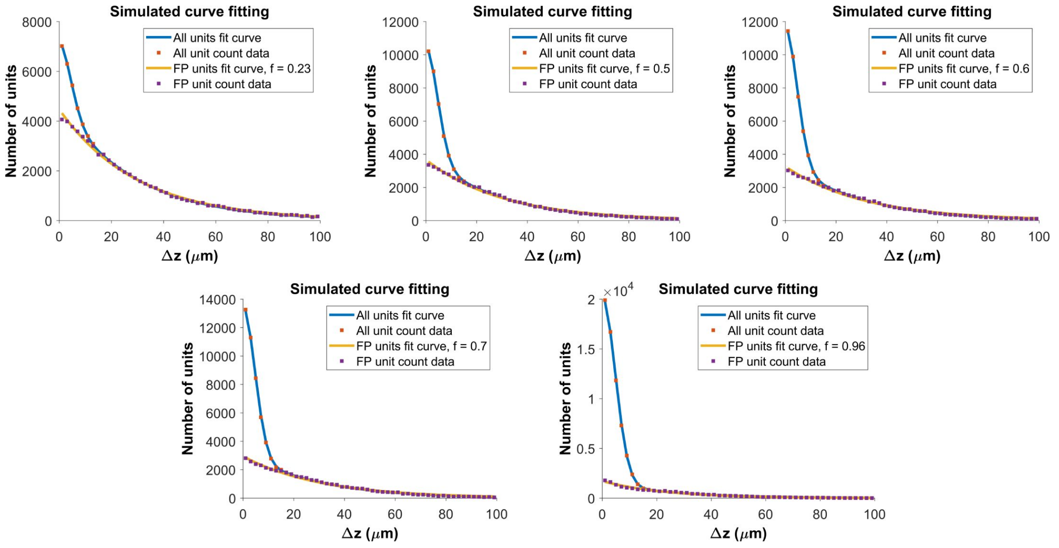

Fig. S8

Fits of z-distance distributions from the Monte Carlo simulations. The five panels correspond to: f = 0.23, 0.5, 0.6, 0.7 and 0.96.

To fit experimental data, we first fit the z-distance distribution of the reference units to obtain the width σ of the folded Gaussian in the first term of ***Equation 3***. With σ fixed, we then fit the z-distance distribution of all KSGood units to ***Equation 3*** to obtain the width of the exponential and f. Then we can estimate the false positive rate by integrating $P\,(\Delta \mid H)$ and $P\,(\Delta \mid\sim H)$ up to the z-distance threshold. The fraction of false positives as a function of z-distance threshold is shown in ***Figure 4***a, in the bottom panel.

Finally, to test model fitting using no information from the reference units, we fit the same z-distance data allowing the width of the folded Gaussian to vary. ***Figure S9***. Panels a and b show the distribution on the same dataset fit with and without fixing the folded Gaussian distribution width. The resulting false positive rate from the no-reference fit at threshold $z = 10\mu m$ is larger than than that from the fit using reference data, so the procedure gives a conservative estimate of the accuracy.

Panel c of ***Figure S9*** shows the model fit to data from an unrelated dataset acquired from mouse prefrontal cortex using a Neuropixels 1.0 probe.$_{35}$ The similar shape of the distribution and a 29% false positive rate suggest that this method can be generalized.

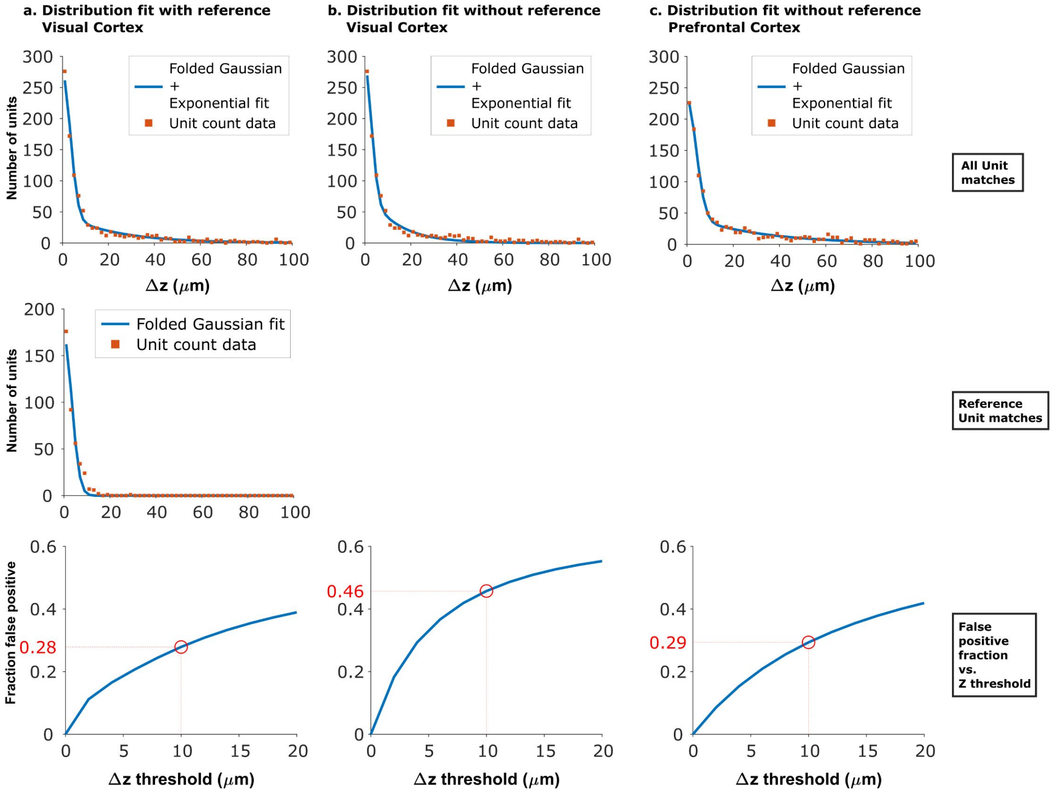

Fig. S9

z-distance distribution fit comparison: a. Distribution fit with 3 parameters, where the z-distribution for true hits is estimated from the reference units. The same as **figure 4a**. b. Distribution fit with 4 parameters, using no reference information. c. Distribution fit of a dataset in prefrontal cortex using Neuropixels 1.0, using no reference information.35

8.5 Recovery rate vs. time between recordings

Fig. S10

The reference unit recovery rate for recordings spanning durations.

Each triangle represents the matching results of two datasets. Animal AL031 has 6 sets of matched units, with one outlier removed. Animal AL032 has 24 sets of matched units. Animal AL036 has 60 sets of matching. The recovery rate is lower for longer durations.

8.6 Reference unit count and the EMD cost matrix

In animal AL036, there is a large decrease in the number of reference units after the second dataset, likely due to a large physical shift of the probe relative to the tissue. It is important to be able to detect such discontinuities to eliminate datasets from consideration. We find that the discontinuity can be detected in the EMD mean cost, location mean cost and waveform mean cost. The pairwise values for the costs are shown in ***Figure S11*** .

To show that days 1-2 (first two rows) are significantly different from days 3-9, we use the Mann-Whitney U Test. All three cost values show significant differences between the groups (EMD mean cost, reject H0, $p = 6 \times 10^{-7}$; location mean cost, reject H0, $p = 6 \times 10^{-5}$; waveform mean cost, reject H0, $p = 5 \times 10^{-7}$)). To show that days 3-9 come from the same distribution, we compare odd and and even rows using the same test. All three cost values show no significant difference between odd and even days (accept H0, p = 0.92).

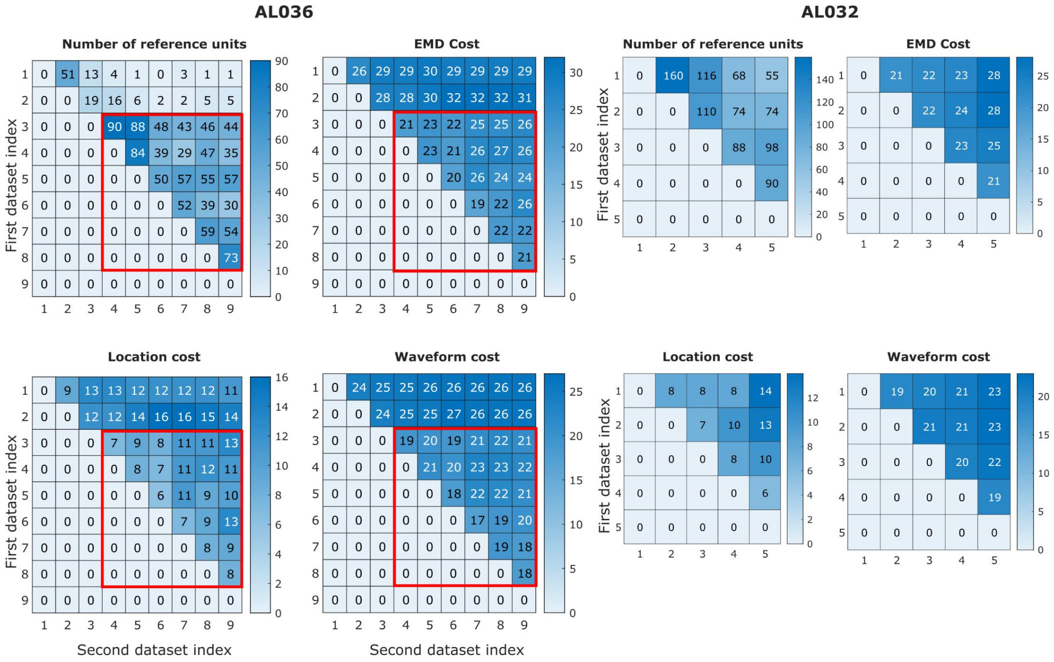

Fig. S11

Reference unit counts and normalized EMD cost for each pair of datasets recorded by the same shank.

For animal AL036 (left), we excluded the first two datasets and all of their matching results (first two rows of each matrix on the left) based on the low reference unit counts. Following analysis on their matching EMD cost, location-only cost and waveform-only cost suggest a significant difference compared to the following days (datasets in the red rectangles). We infer that the first two datatsets were recorded from a different population than later days. The other matrices show similar information for animal AL032 for reference. To show the relative magnitude of EMD cost in related datasets versus unrelated datasets, we calculated the cost between unrelated datasets with similar unit count (AL032 shank 1 and AL036 shank 1: EMD cost = 78, location cost = 67, and waveform cost = 32). The EMD cost is between 70-80, much larger than those between related datasets (between 20-30).

Because days 1-2 are significantly different from 3-9, we eliminated them from our analysis.

8.7 Recovery rate vs. the EMD cost

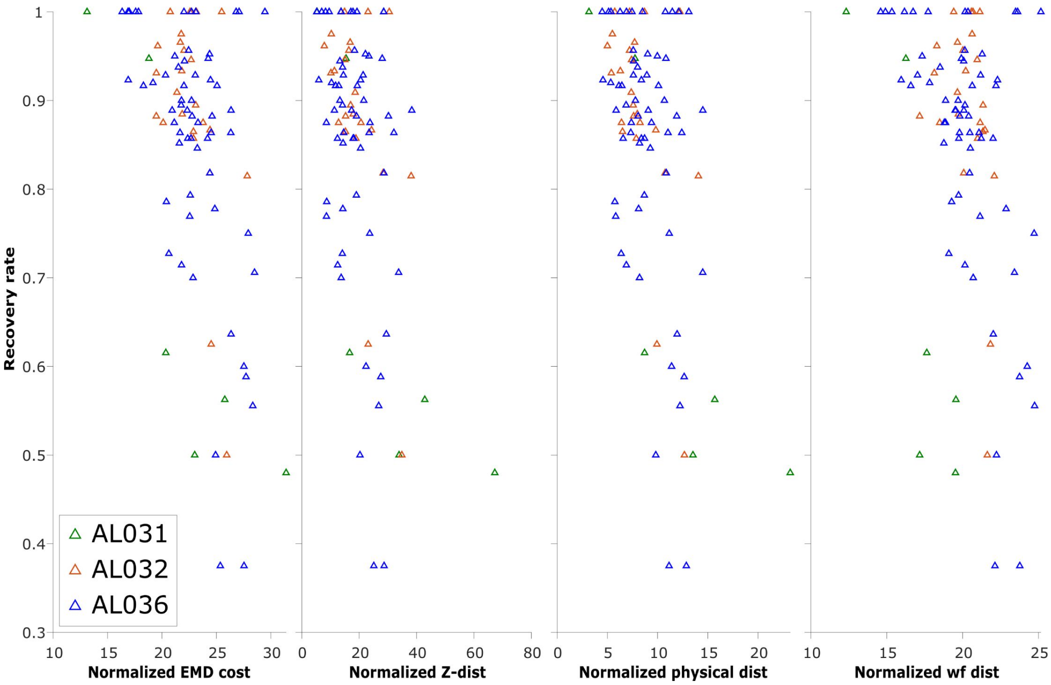

Fig. S12

The normalized EMD cost (unitless), z distance (*μm*), physical distance (*μm*), and waveform distance (unitless) and the corresponding recovery rate in pairwise matches of all to all pairs of recordings, on each shank. Each triangle represents the recovery rate in a pair of datasets. Animal AL031 has 6 sets of matching, with one outlier removed. Animal AL032 has 24 sets of matching. Animal AL036 has 60 sets of matched units. Overall, most of the datatsets with high recovery rates have per-unit EMD cost in the range 20-30. Note that the EMD cost is not predictive of recovery rate.

8.8 Reference unit ratio

Fig. S13

The ratio of number of reference units to number of KSgood units decreases for pairs of datasets with larger time intervals. However, the variability of the number of reference units is generally large for all time intervals.

8.9 Parameter tuning: L2-weight vs. Recovery rate

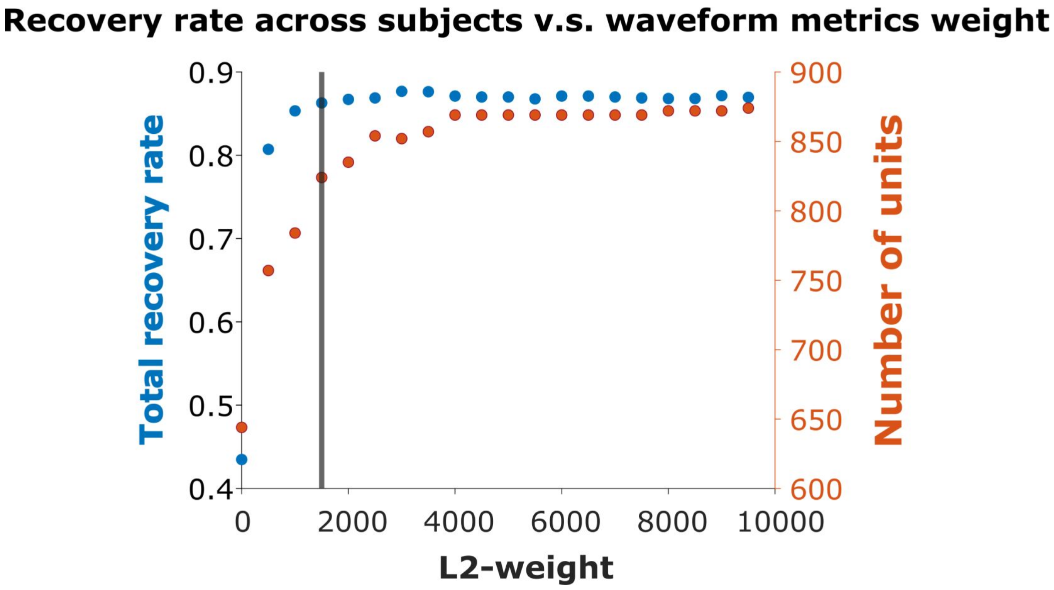

Fig. S14

We varied the weight ω in ***Equation 4*** used to combine the physical and waveform distances in increments of 500. The vertical line indicates weight = 1500, where the overall recovery rate = 86.29%. The maximum recovery rate = 87.68% occurs at weight = 3000. We chose weight = 1500 for all subsequent analysis.

8.10 Reference unit counts

The number of KSgood units in each datatset and number of reference units between a later dataset and the first dataset in animals AL031 and AL032 are shown here.

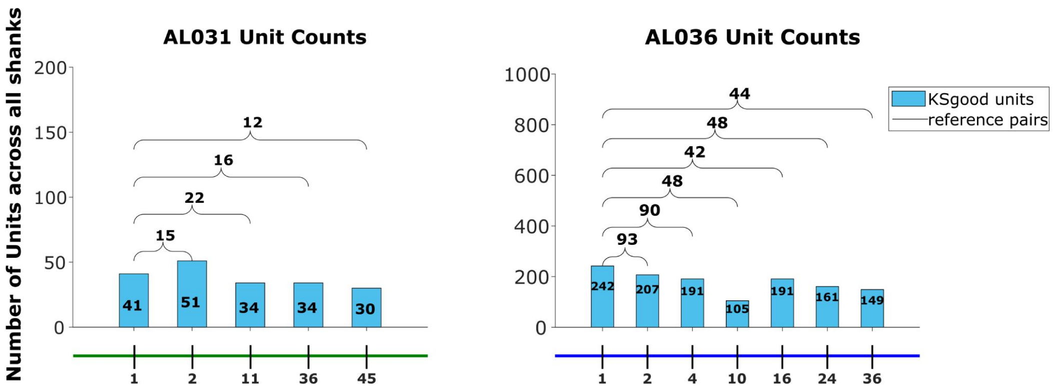

Fig. S15

The Kilosort-good and reference unit counts for the animals AL031 and AL036, as shown for animal AL032 in **Figure 5**.

8.11 Example reference and putative chains

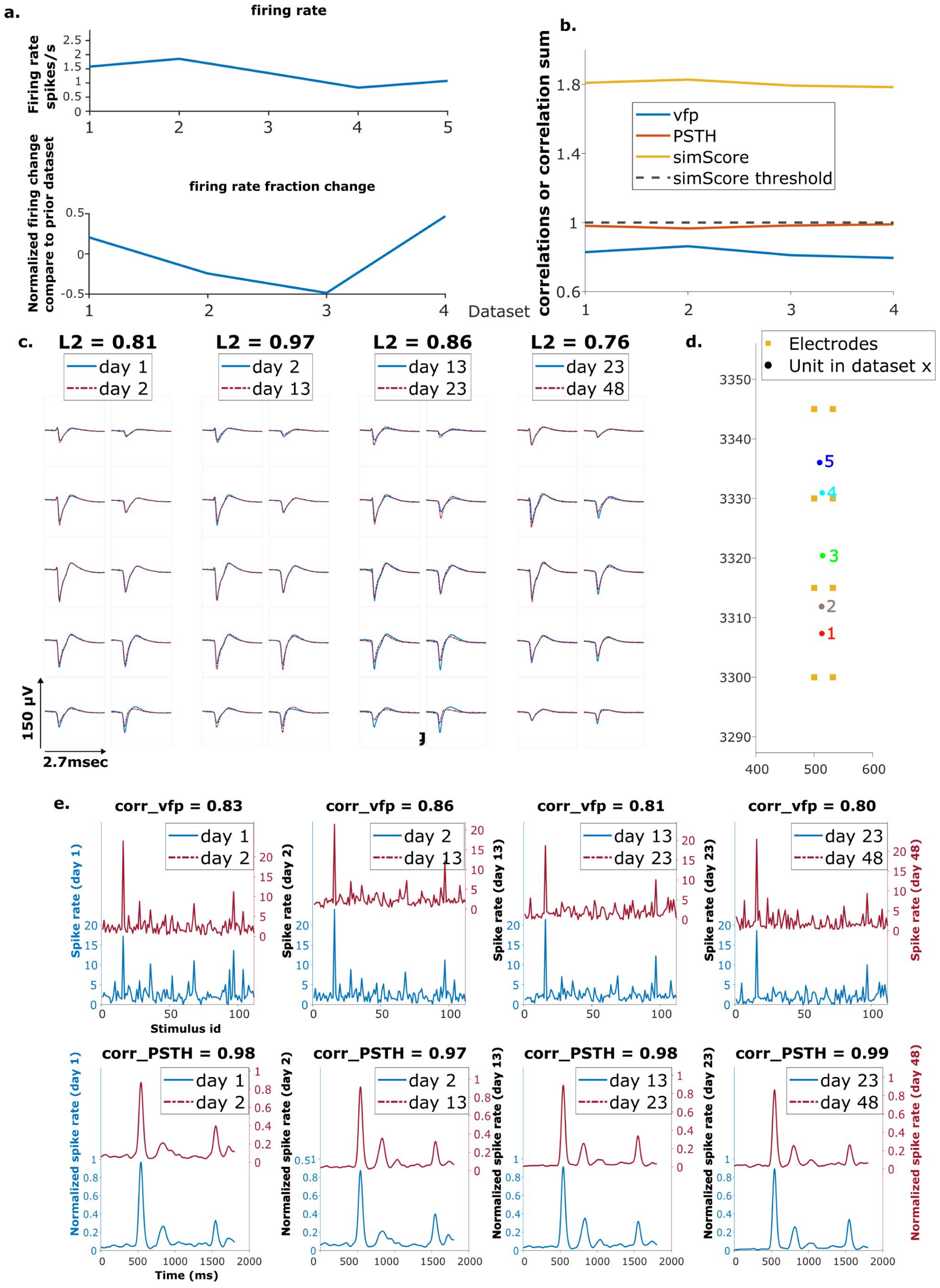

Fig. S16

An example of reference chain.

a. Above: Firing rates of this neuron on each day. Below: Firing rate fractional change compared to the previous day. b. Visual response similarity (yellow line), PSTH correlation (orange line), and visual fingerprint correlation (blue line). The similarity score is the sum of vfp and PSTH. The dashed black line shows the threshold to be considered a reference unit. c. Spatial-temporal waveform of a trackable unit. Each pair of traces represent the waveform on a single channel. d. Estimated location of this unit on different days. Each colored dot represents a unit on one day. The orange squares represent the electrodes. e. The pairwise vfp and PSTH traces of this unit.

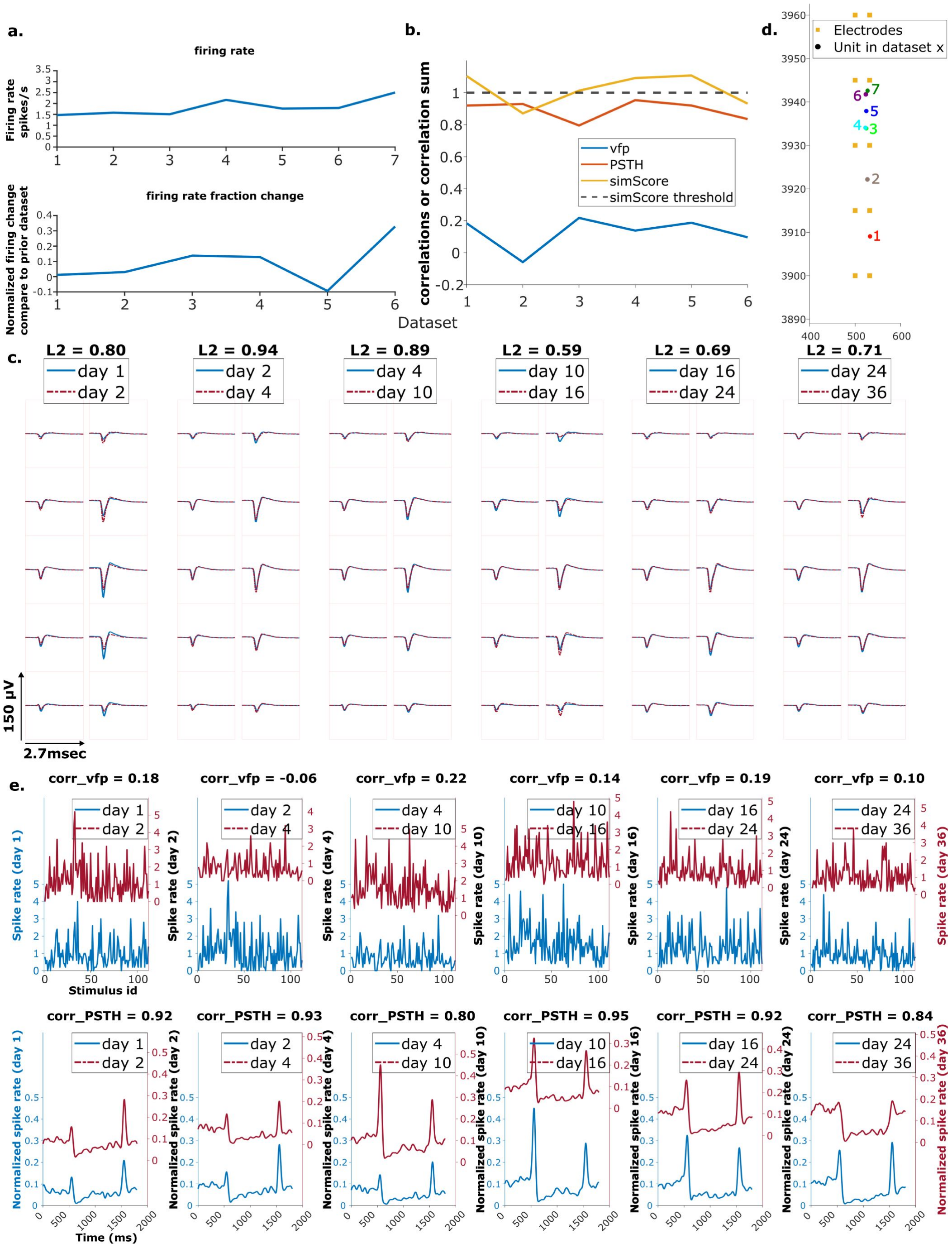

Fig. S17

An example of putative chain.

Order is the same as above.

References

[1] Carmena JM, Lebedev MA, Henriquez CS, Nicolelis MAL. (2005) **Stable Ensemble Performance with Single-Neuron Variability during Reaching Movements in Primates** *J Neurosci* **25**:10712–10716 https://doi.org/10.1523/JNEUROSCI.2772-05.2005

[2] Huber D, Gutnisky DA, Peron S, O'Connor DH, Wiegert JS, Tian L, et al. (2012) **Multiple dynamic representations in the motor cortex during sensorimotor learning** *Nature* **484**:473–478 https://doi.org/10.1038/nature11039

[3] Liberti WA, Markowitz JE, Perkins LN, Liberti DC, Leman DP, Guitchounts G, et al. (2016) **Unstable neurons underlie a stable learned behavior** *Nat Neurosci* **19**:1665–1671 https://doi.org/10.1038/nn.4405

[4] Clopath C, Bonhoeffer T, Hübener M, Rose T. (2017) **Variance and invariance of neuronal long-term representations** *Phil Trans R Soc* **372** https://doi.org/10.1098/rstb.2016.0161

[5] Dhawale AK, Poddar R, Wolff SB, Normand VA, Kopelowitz E, Ölveczky BP. (2017) **Automated long-term recording and analysis of neural activity in behaving animals** *eLife* **6** https://doi.org/10.7554/eLife.27702

[6] Jensen KT, Harpaz NK, Dhawale AK, Wolff SBE, Ölveczky BP. (2022) **Long-term stability of single neuron activity in the motor system** *Nat Neurosci* **25**:1664–1674 https://doi.org/10.1038/s41593-022-01194-3

[7] Steinmetz NA, Aydin C, Lebedeva A, Okun M, Pachitariu M, Bauza M, et al. (2021) **Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings** *Science* **372** https://doi.org/10.1126/science.abf4588

[8] Luo TZ, Bondy AG, Gupta D, Elliott VA, Kopec CD, Brody CD. (2020) **An approach for long-term, multi-probe Neuropixels recordings in unrestrained rats** *eLife* **9** https://doi.org/10.7554/eLife.59716

[9] Harris KD, Quiroga RQ, Freeman J, Smith SL. (2016) **Improving data quality in neuronal population recordings** *Nature Neuroscience* **19**:1165–1174 https://doi.org/10.1038/nn.4365

[10] Buzsáki G. (2004) **Large-scale recording of neuronal ensembles** *Nature Neuroscience* **7**:446–451 https://doi.org/10.1038/nn1233

[11] Brown EN, Kass RE, Mitra PP. (2004) **Multiple neural spike train data analysis: state-of-the-art and future challenges** *Nature Neuroscience* **7**:456–461 https://doi.org/10.1038/nn1228

[12] Quiroga RQ, Panzeri S. (2009) **Extracting information from neuronal populations: information theory and decoding approaches** *Nature Reviews Neuroscience* **10**:173–185 https://doi.org/10.1038/nrn2578

[13] Harris KD. (2005) **Neural signatures of cell assembly organization** *Nature Reviews Neuroscience* **6**:399–407 https://doi.org/10.1038/nrn1669

[14] Quiroga RQ, Nadasdy Z, Ben-Shaul Y. (2004) **Unsupervised Spike Detection and Sorting with Wavelets and Superparamagnetic Clustering** *Neural Computation* **16**:1661–1687 https://doi.org/10.1162/089976604774201631

[15] Chah E, Hok V, Della-Chiesa A, Miller JJH, O'Mara SM, et al (2011) **Automated spike sorting algo-rithmbased on Laplacian eigenmaps and k -means clustering** *J Neural Eng* **8** https://doi.org/10.1088/1741-2560/8/1/016006

[16] Pachitariu M, Steinmetz N, Kadir S, Carandini M, Harris KD. **Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels**

[17] Carlson D, Carin L. (2019) **Continuing progress of spike sorting in the era of big data** *Current Opinion in Neurobiology* **55**:90–96 https://doi.org/10.1016/j.conb.2019.02.007

[18] Jun JJ, Steinmetz NA, Siegle JH, Denman DJ, Bauza M, Barbarits B, et al. (2017) **Fully integrated silicon probes for high-density recording of neural activity** *Nature* **551**:232–236 https://doi.org/10.1038/nature24636

[19] Hall NJ, Herzfeld DJ, Lisberger SG. (2021) **Evaluation and resolution of many challenges of neural spike sorting: a new sorter** *Journal of Neurophysiology* **126**:2065–2090 https://doi.org/10.1152/jn.00047.2021

[20] Tolias AS, Ecker AS, Siapas AG, Hoenselaar A, Keliris GA, Logothetis NK. (2007) **Recording Chronically From the Same Neurons in Awake, Behaving Primates** *Journal of Neurophysiology* **98**:3780–3790 https://doi.org/10.1152/jn.00260.2007

[21] Swindale NV, Spacek MA. (2014) **Spike sorting for polytrodes: a divide and conquer approach** *Frontiers in Systems Neuroscience* **8** https://doi.org/10.3389/fnsys.2014.00006

[22] Bar-Hillel A, Spiro A, Stark E. (2006) **Spike sorting: Bayesian clustering of non-stationary data** *Journal of Neuroscience Methods* **157**:303–316 https://doi.org/10.1016/j.jneumeth.2006.04.023

[23] Lee J, Mitelut C, Shokri H, Kinsella I, Dethe N, Wu S, et al. (2020) **YASS: Yet Another Spike Sorter applied to large-scale multi-electrode array recordings in primate retina** :10712–10716 https://doi.org/10.1101/2020.03.18.997924

[24] Chung JE, Magland JF, Barnett AH, Tolosa VM, Tooker AC, Lee KY, et al. (2017) **A Fully Automated Approach to Spike Sorting** *Neuron* **95**:1381–1394 https://doi.org/10.1016/j.neuron.2017.08.030

[25] Chung JE, Joo HR, Fan JL, Liu DF, Barnett AH, Chen S, et al. (2019) **High-Density, Long-Lasting, and Multiregion Electrophysiological Recordings Using Polymer Electrode Arrays** *Neuron* **101**:21–31 https://doi.org/10.1016/j.neuron.2018.11.002

[26] Vasil'eva LN, Badakva AM, Miller NV, Zobova LN, Roshchin VY, Bondar IV. (2016) **Long-Term Recording of Single Neurons and Criteria for Assessment** *Neuroscience and Behavioral Physiology* **46**:264–269 https://doi.org/10.1007/s11055-016-0227-8

[27] Rokni U, Richardson AG, Bizzi E, Seung HS. (2007) **Motor Learning with Unstable Neural Representations** *Neuron* **54**:653–666 https://doi.org/10.1016/j.neuron.2007.04.030

[28] Lewicki MS. (1998) **A review of methods for spike sorting: the detection and classification of neural action potentials Michael S Lewicki** *Network* **9**:R53–78 https://doi.org/10.1088/0954-898X/9/4/001

[29] Colonell J. **ecephys spike sorting**

[30] Cohen S. (1999) **FINDING COLOR AND SHAPE PATTERNS IN IMAGES** *1999*

[31] Bertrand NP, Charles AS, Lee J, Dunn PB, Rozell CJ. (2020) **Efficient Tracking of Sparse Signals via an Earth Mover's Distance Dynamics Regularizer** *IEEE* **27**:1120–1124 https://doi.org/10.1109/LSP.2020.3001760

[32] Boussard J, Varol E, Lee HD, Dethe N, Paninski L. (2021) **Three-dimensional spike localization and improved motion correction for Neuropixels recordings** *NeurIPS Proceedings* https://doi.org/10.1101/2021.11.05.467503

[33] Sauerbrei BA, Guo JZ, Cohen JD, Mischiati M, Guo W, Kabra M, et al. (2020) **Cortical pattern generation during dexterous movement is input-driven** *Nature* **577**:386–391 https://doi.org/10.1038/s41586-019-1869-9

[34] Stringer C, Pachitariu M, Steinmetz N, Carandini M, Harris KD. (2019) **High-dimensional geometry of population responses in visual cortex** *Nature* **571**:361–365 https://doi.org/10.1038/s41586-019-1346-5

[35] Böhm C, Lee AK. **Functional specialization and structured representations for space and time in prefrontal cortex**

Article and author information

Augustine(Xiaoran) Yuan

Janelia Research Campus, Howard Hughes Medical Institute, USA, Department of Biomedical Engineering, Center for Imaging Science Institute, Kavli Neuroscience Discovery Institute, Johns Hopkins University, USA

Jennifer Colonell

Janelia Research Campus, Howard Hughes Medical Institute, USA

Anna Lebedeva

Sainsbury Wellcome Centre, University of Sheffield, UK

Michael Okun

Department of Psychology and Neuroscience Institute, Howard Hughes Medical Institute, USA

Adam S. Charles

Department of Biomedical Engineering, Center for Imaging Science Institute, Kavli Neuroscience Discovery Institute, Johns Hopkins University, USA
For correspondence: adamsc@jhu.edu

Timothy D. Harris
Janelia Research Campus, Howard Hughes Medical Institute, USA, Department of Biomedical Engineering, Center for Imaging Science Institute, Kavli Neuroscience Discovery Institute, Johns Hopkins University, USA
For correspondence: harrist@janelia.hhmi.org
ORCID iD: 0000-0002-6289-4439

Copyright

Editors

Reviewing Editor
Adrien Peyrache
McGill University, Montreal, Canada

Senior Editor
Panayiota Poirazi
FORTH Institute of Molecular Biology and Biotechnology, Heraklion, Greece

Reviewer #1 (Public Review):

Neurons are not static-their activity patterns change as the result of learning, aging, and disease. Reliable tracking of activity from individual neurons across long time periods would enable studies of these important dynamics. For this reason, the authors' efforts to track electrophysiological activity across days without relying on matching neural receptive fields (which can change due to learning, aging, and disease) is very important.

By utilizing the tightly-spaced electrodes on Neuropixels probes, they are able to measure the physical distance and the waveform shape 'distance' between sorted units recorded on different days. To tune the matching algorithm and to validate the results, they used the visual receptive fields of neurons in the mouse visual cortex (which tend to change little over time) as ground truth. Their approach performs quite well, with a high proportion of neurons accurately matched across multiple weeks.

This suggests that the method may be useable in other cases where the receptive fields can't be used as ground truth to validate the tracking. This potential extendibility to tougher applications is where this approach holds the most promise. However, the study only looks at one brain area (visual cortex), in one species (mouse), using one type of spike sorter (Kilosort), and one type of behavioral prep (head-fixed). While the authors suggest methods to generalize their technique to other experimental conditions, no validation of those generalizations was done using data from different experimental conditions. Anyone using this method under different conditions would therefore need to perform such validation themselves.

https://doi.org/10.7554/eLife.92495.2.sa1

Reviewer #2 (Public Review):

The manuscript presents a method for tracking neurons recorded with neuropixels across days, based on the matching of cells' spatial layouts and spike waveforms at the population level. The method is tested on neuropixel recordings of the visual cortex carried over 47 days, with the similarity in visual receptive fields used to verify the matches in cell identity.

This is an important tool as electrophysiological recordings have been notoriously limited in terms of tracking individual neuron's fate over time, unlike imaging approaches. The method is generally sound and properly tested but I think some clarifications would be helpful regarding the implementation of the method and some of the results.

(1) Page 6: I am not sure I understand the point of the imposed drift and how the value of 12µm is chosen.
Is it that various values of imposed drift are tried, the EMDs computed to produce histograms as in Fig2c, values of rigid drifts estimated based on the histogram modes, and then the value associated with minimum cost selected? The corresponding manuscript section would need some clarification regarding this aspect.

(2) The EMD is based on the linear sum, with identical weight, of cell distance and waveform similarity measures. How performance is affected from using a different weighting of the 2 measures (for instance, using only cell distance and no waveform similarity)? It is common that spike waveforms associated to a given neuron appear different on different channels of silicon probes (i.e. the spike waveform changes depending the position of recording sites relative to the neuron), so I wonder if that feature is helping or potentially impeding the tracking.

(3) Fig.5: I assume the dots are representing time gaps for which cell tracking is estimated. The 3 different groups of colors correspond to the 3 mice used. For a given mouse, I would expect to always see 3 dots (for ref, putative and mixed) for a given tracking gap. However, for mouse AL036 for instance, at tracking duration of 8 days, a dot is visible for mixed but not for ref and putative. How come this is happening?

(4) Matched visual responses are measured by the sum of correlation of visual fingerprints, which are vectors of cells' average firing rate across visual stimuli, and correlation of PSTHs, which are implemented over all visual stimuli combined. I believe that some information is lost from combining all stimuli in the implementation of PSTHs (assuming that PSTHs show specificity to individual visual stimuli). The authors might consider, as alternative measure of matched visual responses, a correlation of the vector concatenations of all stimulus PSTHs. Such simpler measure would contain both visual fingerprint and PSTH information, and would not lose the information of PSTH specificity across visual stimuli.

2nd revision

(1) From reading the authors' response, I could understand several of the points I had previously missed. I still think that some part of the results are not straightforward to understand, the way it is written. Adding a few introductory sentences to the paragraphs (for instance the one related to my previous point #1) would really help the reader comprehend this important work.

(2) Following on my point #2, the w value used is 1500 and the recovery rate doesn't seems to reach a peak but rather a plateau for larger w values. From such large w value and the absence of a downward trend for increasing values, it would seem that only the 'waveform distance' matter and that the 'location distance' doesn't contribute much to the EMD distance. Is this correct?

https://doi.org/10.7554/eLife.92495.2.sa0

Author Response

The following is the authors’ response to the original reviews.

Reviewer #1, in both the public review and recommendations to authors, raises the important question of generalizability of the new technique to other brain areas, to analysis with sorters other than Kilosort, and in the absence of reference data. Specifically, how can experimenters working in brain areas other than visual cortex understand if the tracking is functioning, and set the parameters in the tracking pipeline.

We agree that generalizability of the tracking procedure is a serious issue, especially with respect to other brain areas with varying degrees of measured waveform preservation over time. As the number of potential recording conditions is combinatorial to experimentally test, we instead address these issues in the manuscript by providing a general prescription for interpreting the distribution of vertical distances of matched pairs that can be used for data from any recording using any spike-sorter (Methods section 4.2, Supplement section 8.4, figure S9, paragraphs 7-10 of the Discussion section). This extension of the method allows users to estimate the matching success in the context of their own data, even in the absence of reference data. To address the concern of overfitting, we have also added discussion covering adjustment of the two parameters in the procedure (the relative weight of waveform distance vs. physical distance, and the threshold for accepting matches as real) to the Discussion section.

Reviewer #2 suggested clarification of the following points in the public review. We answer those here and have also clarified these points in the main text where appropriate.

> *(1) What is the purpose of testing the drift correction with imposed drift (Figure 2, page 6 in the original manuscript), and how the value was chosen?*

To test the ability of EMD to detect substantial drift, we need examples that resemble experimental data, including error in fit unit positions and units with no correct matches. We chose to create these examples by taking waveform and position sets from real data with modest drift, and adding a fixed shift to one dataset. The value of 12 um in the figure is arbitrary, simply an example in the range of real drift. These tests allow us to demonstrate the success of EMD for detection of drift in real data.

> *(2) How is performance affected by using a different weighting of the 2 measures (physical distance and waveform distance) in the EMD?*

Recovery rate (number of reference units successfully matched in EMD) vs weighting of the waveform distance is shown in Supplement section 8.10. Recovery rate increases with low values of waveform weighting, leveling off at a value of 1500. We selected that inflection point for the analysis in this paper, to avoid coincidental matching of physically distant units with similar waveforms.

> *(3) Should the intervals measured in the survival plot in Figure 5 be identical for the three different classes of tracked neurons?*

The plot includes all chains of tracked neurons, which can start on arbitrary days in the set of all recordings (see the definition of chains in section 2.4). As a result, the gaps between days, which determine where there is a point on the plot, can be different for the three sets of neurons (reference, putative, and mixed). We have added a comment to the Figure 5 caption to ensure this is clear.

> *(4) Would other metrics of the similarity of visual responses work better?*

The similarity metric we use was adopted from the original paper using this data (reference 7). We chose to use the same metric both to take advantage of the original authors' expertise about the data and allow for reasonable comparison of the new technique to theirs. It is correct that this similarity metric alone does not allow for unique matching (see Discussion and Supplement section 8.2). However, the agreement of EMD with reference pairs determined from the combination of position and visual response similarity is very high, suggesting there are few incorrect reference pairs. Any incorrect reference pairs cause an underestimate of the tracking accuracy.

> *(5) Add a definition of ROC.*

Added this definition to the text.

> ***Reviewer #1 Recommendation to authors:***
>
> *The main text needs proofreading.*

We agree that the manuscript needed more thorough proofreading, and we have made corrections of typos and minor language errors throughout.

Additional comment from the authors:

Since the posting of this manuscript, another method for tracking neurons has been introduced:

Enny H. van Beest, Célian Bimbard, Julie M. J. Fabre, Flóra Takács, Philip Coen, Anna Lebedeva, Kenneth Harris, Matteo Carandini, Tracking neurons across days with high-density probes, bioRxiv 2023.10.12.562040; doi: https://doi.org/10.1101/2023.10.12.562040