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

Augustine(Xiaoran) Yuan[1,2], Jennifer Colonell[1], Anna Lebedeva[3], Michael Okun[4], Adam Charles[2*] and Timothy Harris[1,2*]

[1*]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA.

[2]Department of Biomedical Engineering, Center for Imaging Science, Kavli Neuroscience Discovery Institute, Johns Hopkins University, Baltimore, MD, USA.

[3]Sainsbury Wellcome Centre, University College London, London, UK.

[4]Department of Psychology and Neuroscience Institute, University of Sheffield, Sheffield, UK.

*Corresponding author(s). E-mail(s): adamsc@jhu.edu; harrist@janelia.hhmi.org;
Contributing authors: yuanx@janelia.hhmi.org; colonellj@janelia.hhmi.org; anna.lebedeva.17@ucl.ac.uk; m.okun@sheffield.ac.uk;

Abstract

Accurate tracking of the same neurons across multiple days is crucial for studying changes in neuronal activity during learning and adaptation. 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 iden-tify the same cells independent of firing statistics, which a reused 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.

Keywords: Neuron tracking, Neuropixels, visual cortex, chronic recording

1 Introduction

The ability to longitudinally track neural activity is crucial to understanding central capabilities and changes of neural circuits that operate on long timescales, such as learning and plasticity[1][2][3][4] and motor stability[5][1][6]. 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 for 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][9][10][11][12][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 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, for example,[14] and [15], and template matching methods, for example, Kilosort[16][9][11][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 (Fig. 1a and 1b). 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 lusters may appear and disappear in different recordings[9][20][21][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 to clusters differently, and clusters can be merged or separated differently across runs.

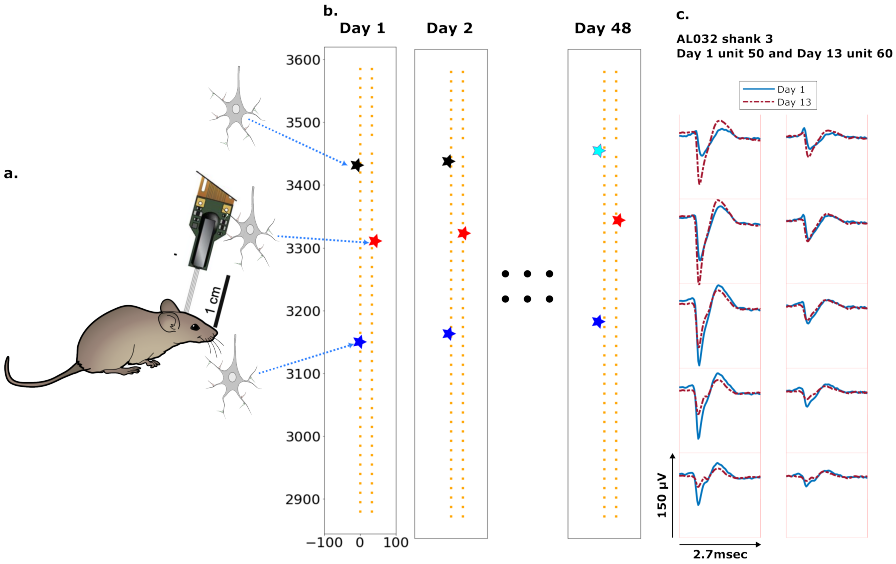

Fig. 1: Schematic depiction of drift: a. Mice were implanted with a 4-shank Neuropixels 2.0 probe in visual cortex area V1. b. Each colored star represents the location of a unit recorded on the probe. In this hypothetical case, the same color indicates unit correspondence across days. The black unit is missing on day 48, while the turquoise star is an example of a new unit. Tracking aims to correctly match the red and blue unit across all datasets and detect that the black units on day 48 is likely undetected. c. Two example spatial-temporal waveforms of units recorded in two datasets that likely represent the same neuron, based on similar visual responses. Each trace is the average waveform on one channel across 2.7 millisecond. The blue traces are waveforms on a peak channel and 9 nearby channels (two rows above, two rows below, and one in the same row) from the first dataset (Day 1). The red traces, similarly selected, are from the second dataset. Waveforms are aligned at the electrodes with peak amplitude, different on the two days.

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][2][3] or representational drift [27], neural activity statistics became less reliable. In this work, we take advantage of the rich spatial-temporal 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 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, spike times, are used as input for our method (post-processed using [29]) (Sec. 4.2). 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. 4.3). KSgood units are mainly determined by the fraction 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: matching of all ensuing sessions to a single session (e.g., the first session) (Sec. 2.2 and Sec. 2.3), or match consecutive pairs of sessions and then tracing matched units through all sessions (Sec. 2.4).

We refer to the subset of KSgood units with strong and distinguishable visual responses in both compared datasets as reference units (See Sec. 4.3 for details). Similar to Steinmetz et al.[7]. we validated our unit matchings of those reference units using visual receptive field similarity. Finally, we showed that trackable units with strong visual response are qualitatively similar to those without (Supplement Fig.S1 to 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 $d_{ik} = d_{location_{ik}} + d_{waveform_{ik}}$ (Fig.2a). The first EMD stage estimates the homogeneous vertical movement of the entire population of KSgood units (Fig. 2b). 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. 4.3).

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 (Fig. 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 units assignments established using visual response (Sec. 4.3).

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

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 number of unit assignments within the z-distance threshold by our method the same as assignments of reference units.

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

We 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 rate, visual responses (see

Supplement Fig.S1 to S5 for details) of all tracked chains, i.e. units that 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 (720 μm) z-range of these recordings. The EMD allows us to extract the homogeneous (rigid) movement of matched units. For datasets with a few units consistently detected across days, this problem is relatively simple (Fig. 2a). In these datasets, we find that only $\approx 60\%$ of units are detected on both 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 (Fig. 2c).

The EMD match of units from the first dataset (Fig. 2b, open circles) and from the dataset recorded the next day (Fig. 2b, closed circles) is indicated by the arrow between them. We added a 12 micron upward drift to the z-coordinate of the units from the second day and use the first stage of the EMD to find the matches using combined distances as described in sections 4.1.2 and 4.1.2. We used the z-distance mode (Mode = 16 μm) of a kernel fit of all the matched units to find the rigid drift (Fig. 2c). The resulting mode is close to the actual imposed drift ($d_i = 12\mu m$).

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. We thus select only the pairs with separation below a threshold. Here we use $\lambda = 15\mu m$, which is chosen to be larger than most of the z-shift in experimental data, and will be refined later by distribution fitting (Fig. S2). All of the sub-threshold (short) distances belong to upward pairs (Fig. 2b and 2c, 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 30 μm) 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.

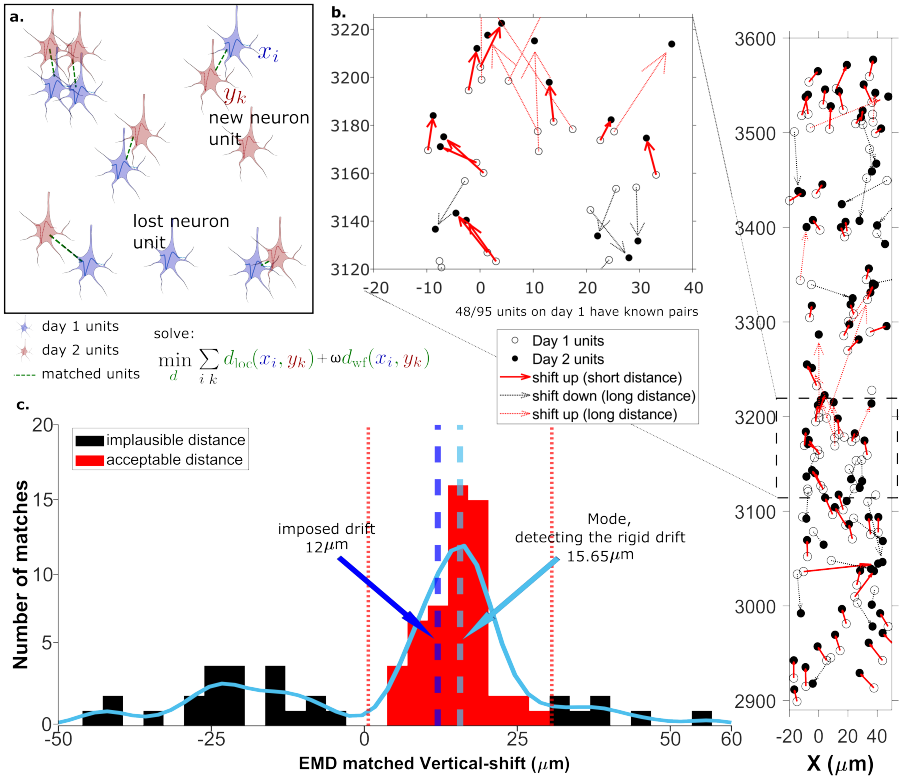

Fig. 2: The EMD can detect the displacement of single units: a. Schematic of EMD unit matching. Each blue unit in day 1 is matched to a red unit in day 2. Dashed lines indicate the matches to be found by minimizing the weighted sum of physical and waveform distances. b. Open and filled circles show positions of units on days 1 and 2, respectively. Arrows indicate matching using EMD. The arrow color represents the match direction; upward matches found with the EMD are in red and downward in black. Solid lines indicate a z-match distance within $15\mu m$, while a dashed line indicates a z distance $> 15\mu m$. An expand segment is shown for the probe section from 3120 to $3220\mu m$. c. The match distance histogram (black and red bars) and kernel fit (light blue solid curve). The light blue dashed line shows the mode ($d_m = 15.65\mu m$). The dark blue dashed line shows the imposed drift ($d_i = 12\mu m$). The red region shows the matches within $15\mu m$ of the mode. The EMD needs to detect the homogeneous movement against the background, i.e. units in the black region that are unlikely to be the real match due to biological constraints.

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 ROC curve in figure 3 shows the percentage of reference units recovered and the number of KSgood preserved as a function of z-distance threshold. We want to determine the best threshold that maximizes the overall accuracy in the reference units (Fig. 3, blue curve) while including as many KSgood units as possible (Fig. 3, red curve).

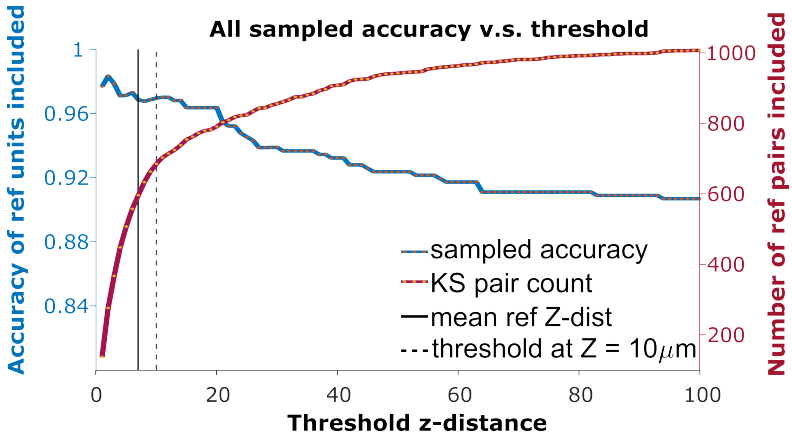

Fig. 3: The ROC curve of matching accuracy vs. distance. The blue curve indicates the recovery rate of reference units. The red line indicates the number of reference units included. The solid vertical line indicates the average z distance across all reference pairs in all animals ($z = 6.96\mu m$). The dashed vertical black line indicates a z-distance threshold at z = $10\mu m$.

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 (Fig.4a, top and middle panel). While

both distributions may be fit to exponentials, 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 distribution of KSgood correct and incorrect pairs vs. z-distance of the matched units.

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, Eqn. 5).

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 (Fig. 4a, middle panel) and an exponential distribution for false positive units. The KSgood distribution is a weighted combination of the folded Gaussian and an exponential. (Fig. 4a, top panel, see Sec. 8.4 for details).

Based on the the estimated false positive rate (Fig. 4a, bottom panel), we used the threshold of $10\mu m$ (Fig. 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 (Fig. 4b, green bars).

Note that this threshold eliminates most of the known false positive matches of reference pairs (Fig. 4b, red fraction) at the cost of recovering fewer correct pairs (Fig. 4b, green bars). The recovery rate varies from day to day; datasets separated by longer times tend to have higher tracking uncertainty (Fig. S9).

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

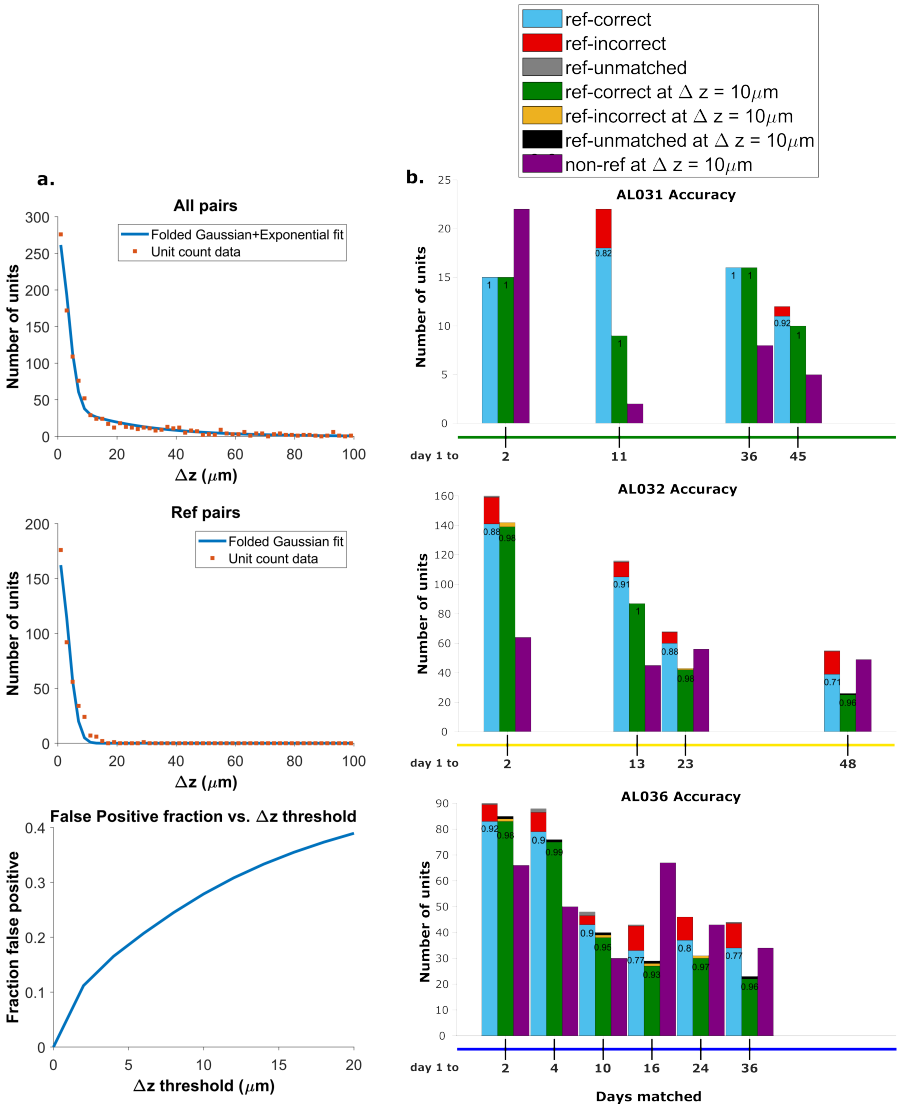

Fig. 4: Recovery rate, accuracy and putative pairs: a. The histogram distribution fit for all KSgood units (top) and reference units alone (middle). False positives for reference units are defined as units matched by EMD but not matched when using receptive fields. The false positive fraction for the set of all KSgood units is obtained by integration. z = $10\mu m$ 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 those matches using receptive field data that were not recovered using EMD without receptive field information. Similarly, the green bars show matching accuracy with distance under threshold z at $10\mu m$. 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 = $10\mu m$.

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 S11, and figure S10, respectively.

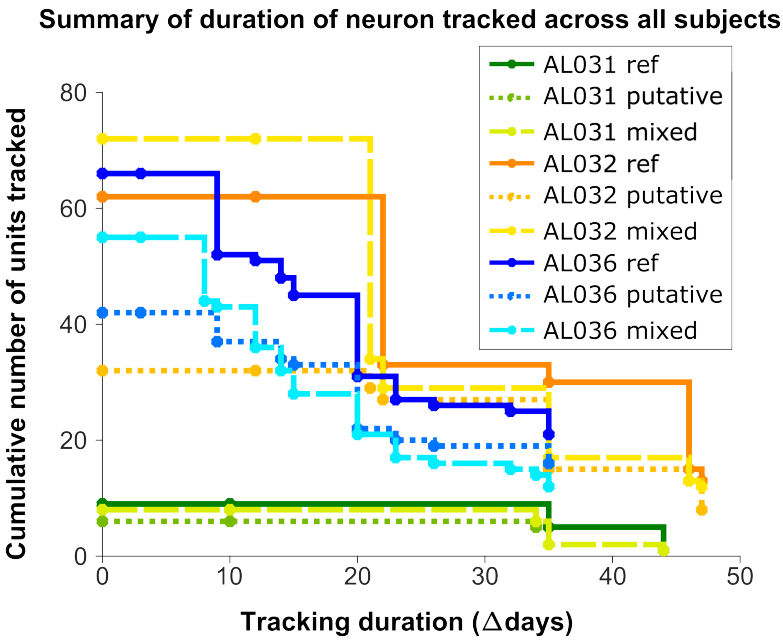

Fig. 5: Number of reference units (deep blue, dark orange and green for different subjects), putative (medium green, medium orange and blue) units, and mixed units (light green, yellow, and light blue) tracked for different duration of days. The decrease rate is similar for different chain types in the same subject.

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, respectively.

The spatial-temporal waveform similarity is measured by the L2 distances between waveforms (Sec. 4.1.2). A Kruskal Wallis test is performed on the amount of L2 change between two matched waveforms among the three groups. There is no statistical difference in the waveform similarity in reference, putative, and mixed units ($H = 0.59$, $p = 0.75$) (Fig.S1). There is no significant difference in the physical distances of units per dataset ($H = 1.31$, $p = 0.52$) (Fig.S2, bottom panel), and in the location change per units ($H = 0.23$, $p = 0.89$) (Fig.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 ($H = 1$, $p = 0.6$) (Fig.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 similarity score, the sum of the two correlations (Sec. 4.3). Vfp change between matched units is similar among the three groups ($H = 2.23$, $p = 0.33$). Similarly, PSTH change is not different among the three groups ($H = 1.61$, $p = 0.45$) (Fig. S4).

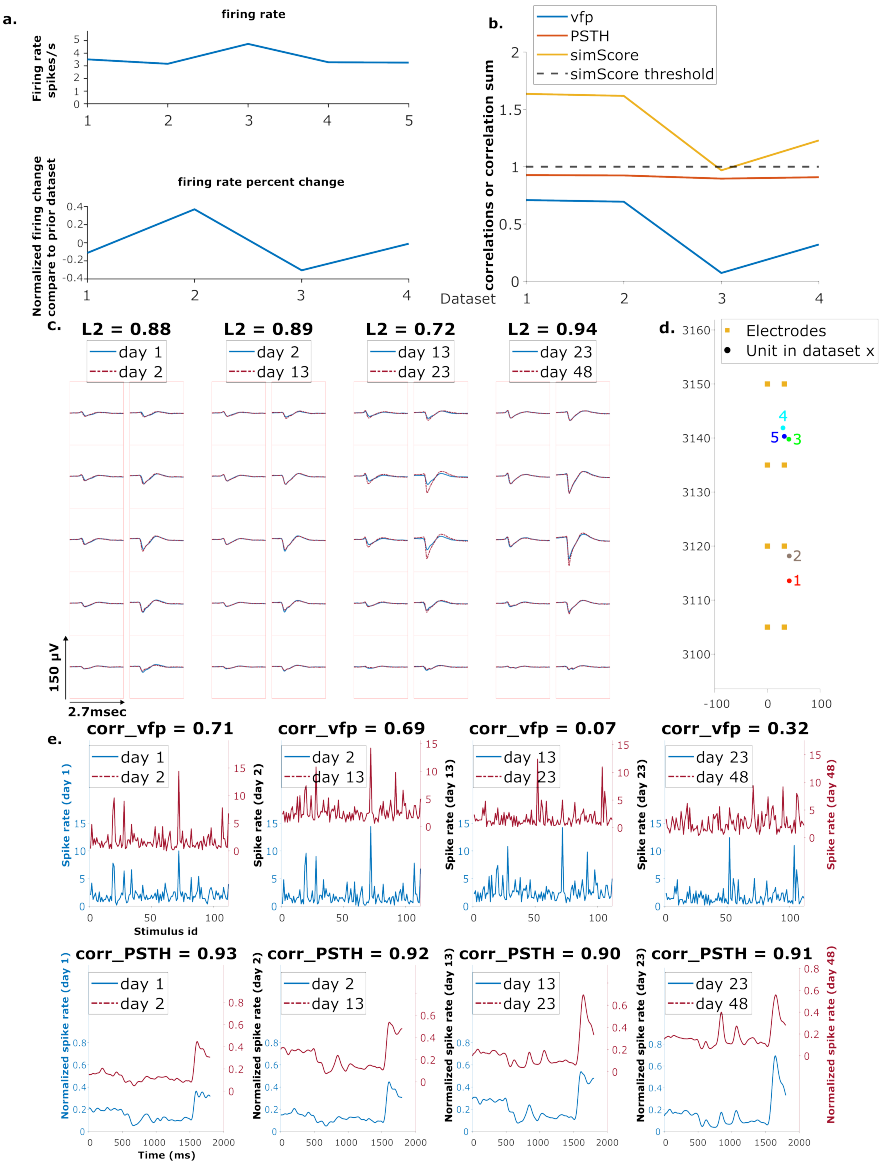

Fig. 6: Example mixed chain: a. Above: Firing rates for this neuron on each day (Day 1, 2, 13, 23, 48). Below: Firing rate percent change compared to the previous day. b. Visual response similarity (yellow line), PSTH similarity correlation (orange line), and visual fingerprint similarity correlation (blue line). The similarity score is the sum of vfp and PSTH. The dashed black line represent 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 represent a unit in a day. The orange squares represent the electrodes. e. The pairwise vfp and PSTH traces of this unit.

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 fields 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 (Fig. 4b, blue bars). We also achieved 97% accuracy up to one week apart and 95% five to seven weeks apart, when applying a threshold of 10 μm (Fig. 4b, 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 generalizable with existing sorting methods, we chose Kilosort, one of the most widely used spike sorting methods (e.g., [33][34]). We show the capability of our method to track neurons with no specific tuning preference (Fig. S10).

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. 8.7). For example, dataset 1 and 2 in animal AL036 have very few reference units compared to other datasets (see Fig. S12, AL036). This observation is consistent with the overall relationship between the EMD cost and recovery rate (Fig. S13). 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 (Fig. 6, S12), 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 $30\mu m$. This contributed to inaccurate assignment of reference units and irregularity in trackable unit analysis. In summary, 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. To account for the possibility that the most similar unit within the neighborhood might not be the real match, we redefined the reference units as any neuron in the neighborhood with a similarity score ≥ 1, rather than the one with the highest similarity score. This reassignment results in the loss of 162 out of 934 (17%) reference units across all comparisons, while the overall recovery rate decreases by only 0.14%.

We note that the ratio of reference units over KSgood units decreases as recordings are further separated in time (Fig. S14). This reduction in KSgood units might due in part 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 ([7] supplement).

As suggested in Dhawale et al.[5], the false positive rate is not ideal for discontinuous recordings. Improving spike sorting and restricting the analysis to reliably sorted units will help to decrease the false positive rate. Current spike sorting methods involves fitting of many parameters. The stochastic nature of template initialization makes only around 60% to 70% of the same units for two independently executed analysis passes. Additionally, the existence of unpaired units in the neighborhood will deteriorate the matching accuracy. Future users may consider limiting their analysis to the most reliably detected units for tracking. 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' without receptive field information: 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. 4.3). 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 (Fig. 4a). '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

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 $\in U_2 : z_{corr} = z - z_{mode}$

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 $\sum d_{EMD}$, 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 one to day 2, by solving a minimum overall effort, the total distances required for the movement[30][31].

$$\begin{aligned} \min_{d_F} \quad & \sum_{i\,k} D(x_i, y_k), where\ D = d_{loc} + \omega d_{wf} \\ \text{subject to} \quad & f_{ik} \in [0,1]\ \forall i, k \\ & \sum_k (f_k) \leq length(Y) \\ & \sum_i (f_i) \leq length(X) \\ & \sum (F) = \min(\sum X, \sum Y) \end{aligned} \tag{3}$$

in which $d_{loc} \in \mathcal{D}^3$ is the three-dimensional physical distance between a unit from the first dataset x_i, and a unit from the second dataset y_k. $d_{wf} \in \mathcal{D}^1$ 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 indicates the matched objects between the two datasets (See Fig.S15 for details about selecting weight).

The EMD has three benefits:

- It allows combing different information into 'distance matrix' to characterize the features of units.
- The EMD can detect homogeneous movement of units (Fig. 2c), thus providing means 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 property of the data, it can make matches to the lower number of the two sets. We thus added a threshold over the permissible z-distance to select physical plausible matches. Supplement Fig. S15 shows the recovery rate change for using different weight parameters to combine neuron location and waveform metrics into a distance matrix.

4.1.2 Distance 1: Unit Location Estimation

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

$$\min \sum_{c \in C_{u_n}} \left(V_{PTP_c} - \frac{1}{\sqrt{(x-x_c)^2+(z-z_c)^2+y^2}} \right)^2 \tag{4}$$

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

Distance 2: Waveform Similarity Metrics We want to describe the waveform characteristics of each unit with its spatial-temporal waveform at the channels capturing the greatest differentiability. 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 1c 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 Dataset

The data used in this work are two chronically implanted NP 2.0 four-shank probes and one chronically implanted one-shank NP 2.0 probe recordings collected in the visual cortex of three head fixed mice (Fig. 7b, see [7] for experiment details). The recordings were taken while 112 visual stimuli were shown from three surrounding screens (data from [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 on 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 (Fig. 7a), 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 (Fig. 7d and S16).

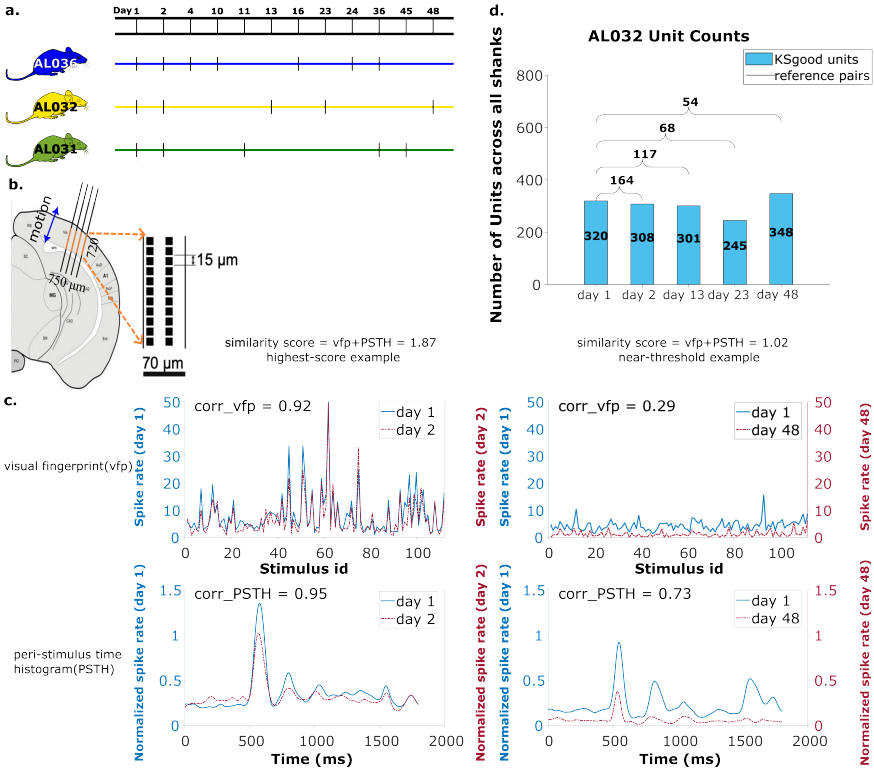

Fig. 7: Summary of dataset: a. The recording intervals for each animal. A black dash indicates one recording on that day. b. All animals were recorded from visual cortex V1 with a 720 μm section of the probe containing 96 recording sites. The blue arrow indicate main drift direction. c. Examples of visual fingerprint(vfp) and peri-stimulus time histogram(PSTH) from a high correlation (left column) and a just-above-threshold (right column) correlation unit. Both vfp and PSTH values vary from [0,1]. d. Kilosort-good and reference unit counts for animal AL032, including units from all four shanks.

4.3 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 (15 μm 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 ([7], Fig. 4). In addition, even though mean fingerprint similarity decreases for recordings separated by more than 16 days, there is roughly 60% of mean fingerprint similarity for the same unit recorded from 40 days apart (see [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 [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 400 msec after the stimulus presentation. The PSTH is calculated by 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 across the 2 sessions (0-1) and the correlation of the vfp across those same sessions (0-1), ranging from 0 to 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 30 μm. We chose 30 μm on both pre- and post-correction data because the drift is relatively small in our case, and we can reduce false positives by constrain 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), similar to the case of putative units in our case, which do not have reference visual information. 2) A Kruskal-Wallis test is applied on all trials of the vfps to make sure the response firing pattern triggered by 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 the sorted clusters based on ISI violation [16]. This step attempts to ensure the 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 reference units capture two representations from different sessions of the neurons' visual response to the stimulus. The portion of units with qualified visual response ranges from 5% to 61%, depending on the time gap between datatets (Fig.S14). Overall, these reference units made up 29% of all KSgood units (Fig. S16) across all three animals in our dataset. Figure 7c shows examples of visual responses from a high similarity and a just-above-threshold similarity reference units.

5 Code sharing

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

6 Acknowledgments

We thank M. Okun from University College London for his assistance in data pre-processing. This work was supported supported in part by NIH grant U01 NS115587.

7 Declaration of interests

The authors declare no competing interests.

8 Supplements

8.1 Trackable units statistics

In order to make sure trackable reference, putative, and mixed units are qualitatively similar, We summarized the median, maximum and minimum amount of change on firing rate, visual receptive field, and location in the box plots below. Kruskal Wallis test performed in each feature suggested no difference among the three groups (see Sec. 2.4 for details).

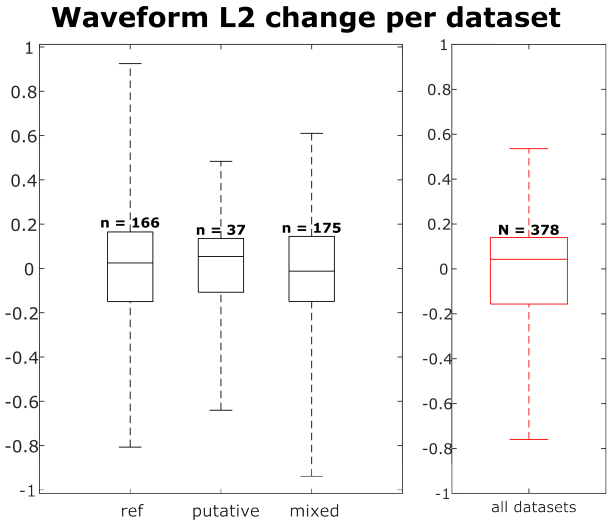

Fig. S1: The waveform L2 similarity change distribution per dataset by neuron groups and across all neurons. Box plots indicate 25% percentile, medians, and 75% percentile. Whiskers at the ends of the box plot show maximum and minimum values. n and N represents the number of unit comparisons, i.e. number of units times (number of datatset - 1).

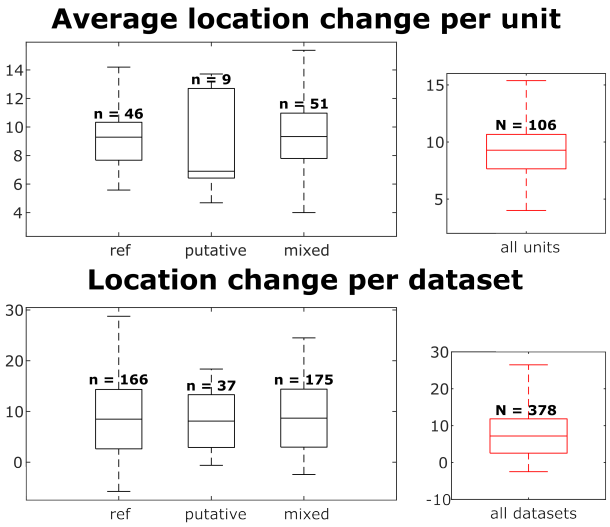

Fig. S2: The plots show location change distribution per units and the bottom plot show the change per dataset by neuron groups and across all neurons. Box plots indicate 25% percentile, medians, and 75% percentile. Whiskers at the ends of the box plot show maximum and minimum values. n and N represents the number of units (above plot), and the number of unit comparisons, i.e. number of units times (number of datatset - 1) (lower plot).

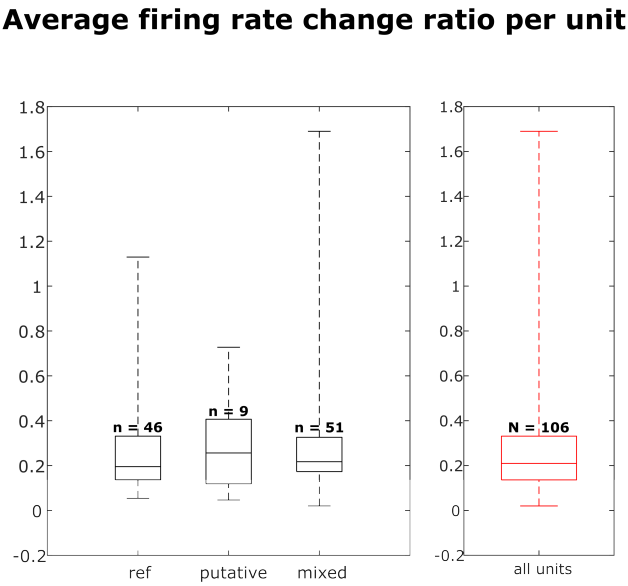

Fig. S3: The average firing rate fold change per dataset by neuron groups and across all neurons. Box plots indicate 25% percentile, medians, and 75% percentile. Whiskers at the ends of the box plot show maximum and minimum values. n and N represents the number of units.

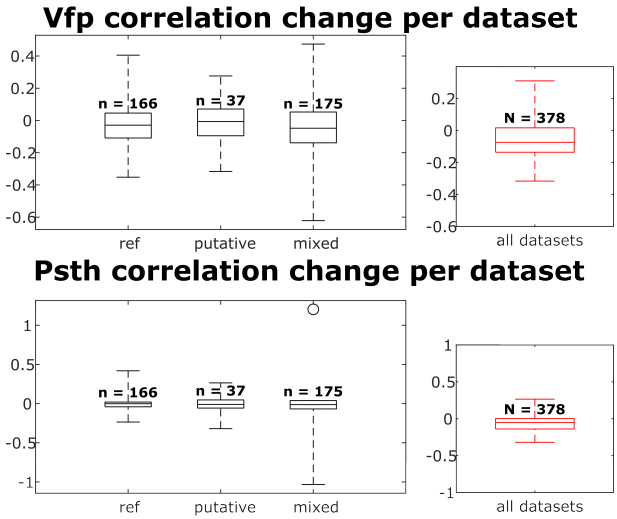

Fig. S4: The visual fingerprint and PSTH change distribution per dataset by neuron groups and across all neurons. Box plots indicate 25% percentile, medians, and 75% percentile. Whiskers at the ends of the box plot show maximum and minimum values. n and N represents the number of unit comparisons, i.e. number of units times (number of datatset - 1).

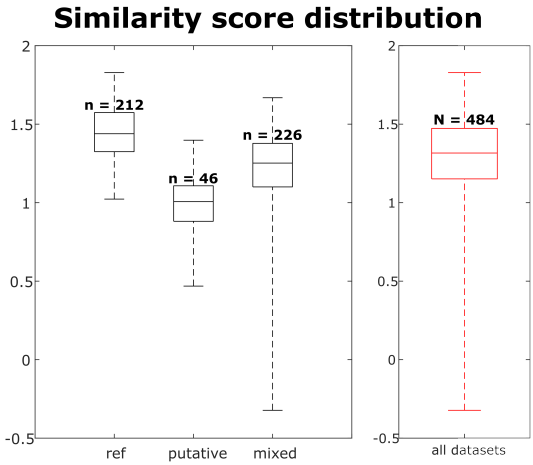

Fig. S5: The similarity score distribution per dataset by neuron groups and across all neurons.Box plots indicate 25% percentile, medians, and 75% percentile. Whiskers at the ends of the box plot show maximum and minimum values. n and N represents the number of occurrence of units, i.e. number of units times number of datasets this unit have.

8.2 Similarity score heatmap

We identify reference pairs as units that are close in space (peak channels separated by $< 30\mu m$) 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 $30\mu m$ distance threshold.

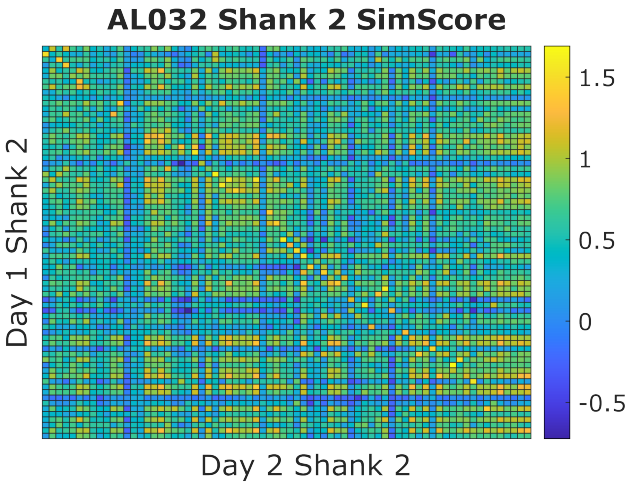

Fig. S6: An example similarity score (vfp + PSTH) heatmap from animal AL032 shank 2 Kilosort-good units between day 1 and 2. Each small square represents the similarity score (value range from [0,2]) between one unit from day 1 and one unit from day 2. A warm colored square indicates a higher score. All clusters are sequenced by their physical locations on the probe. There is a diagonal line with brightest color blocks, indicating that units with more similar firing responses across days tend to be physically close. This confirms our assumption that neurons are physically stable overtime. Also notice that, on each column, there might be more than one bright blocks in the more distance clusters. We minimizes the effect of the distant units by constraining the feasible region during reference units selection. There are also columns without bright yellow blocks. This happens because some units do not respond to the stimulus and those units are not included in the reference set.

8.3 Pre- and post-drift correction Unit counts

We showed that between-session drift correction retrieved more reference units.

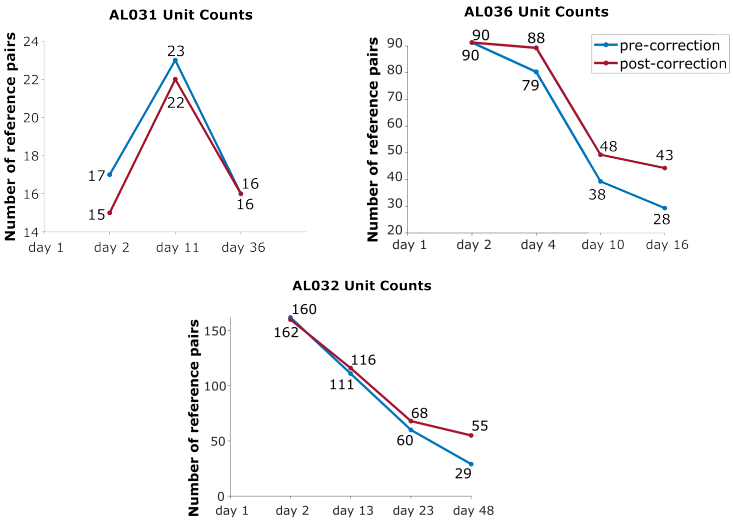

Fig. S7: The effect of drift correction in finding reference units for all three animals. Note that drift correction improves the recovery rate for most cases; the degree of improvement is a function of the magnitude of the drift.

8.4 ref and KSgood distribution

As shown in Fig.4a, the z distance distribution of reference differs significantly from that of all pairs. To estimate the false positive rate (False positives are defined as units matched by EMD but not matched by using receptive fields) 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 selection bias in the reference units, which must be easily detectable units in both datasets to meet the requirements of similar visual fingerprint. We created a simple model to determine an appropriate functional form to fit the z distribution of all units and estimate the false positive.

Assume the following distributions:

1. The distance distribution of matched neurons, i.e. KSgood unit distribution, ($\Delta > 0$) is
$$P(\Delta)$$
2. The distance distribution of matched neurons that are true hits, i.e. the correctly-matched reference unit distribution (H: correct match/hits) is
$$P(\Delta \mid H)$$
3. The distance distribution of false positive matched neurons, i.e. false positive distribution, 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:

$$P(\Delta) = f * P(\Delta \mid H) + (1 - f) * P(\Delta \mid\sim H) \tag{5}$$

To estimate the distribution of $P(\Delta \mid H)$, we assume that drift correction works properly. The z shift between the two units of a reference pair 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 mean = 0, and sigma = 2*(error in measured z position).

To estimate the distribution of $P(\Delta \mid\sim H)$, we perform a Monte Carlo model of simulated units. We chose the number of units to be 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 sigma = $5\mu m$. This error is set to match observed distribution of z-distance in the reference units.

To determine a range 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

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

$P(H \mid R)$ 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. $P(\sim R) = 1 - P(R) = 0.73$. Then

$$P(H \mid R)P(R) \leq \quad P(H) \leq \quad P(H \mid R)P(R) + P(\sim R) \tag{6}$$

$$0.25 \leq \quad P(H) \leq \quad 0.96 \tag{7}$$

We estimated the conditions with 25% to 96% of hits, which includes units with or without reference information.

$$P(AllUnits) = f * P(FoldedGaussian) + (1 - f) * P(Exponential) \tag{8}$$

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 $\sim H$, both from the EMD solution that 'mispair' some of the true hits due to unit crowding.

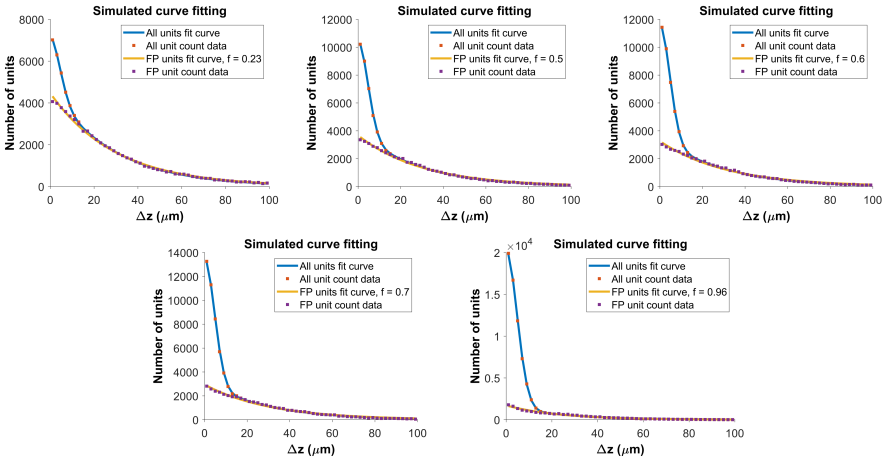

Fig. S8: The distribution fitting for KSgood and correct reference units with simulated data at f = 0.23, 0.5, 0.6, 0.7, 0.96, respectively.

We fit the distance distribution of the reference units to obtain sigma, the width of the folded gaussian in the first term of Eqn.8. With sigma fixed, we then fit the distance distribution of all KSGood units to Eqn.8 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 positivies as a function of z distance threshold is shown in Fig.4a, in the bottom panel.

8.5 Recovery rate vs. time between recordings

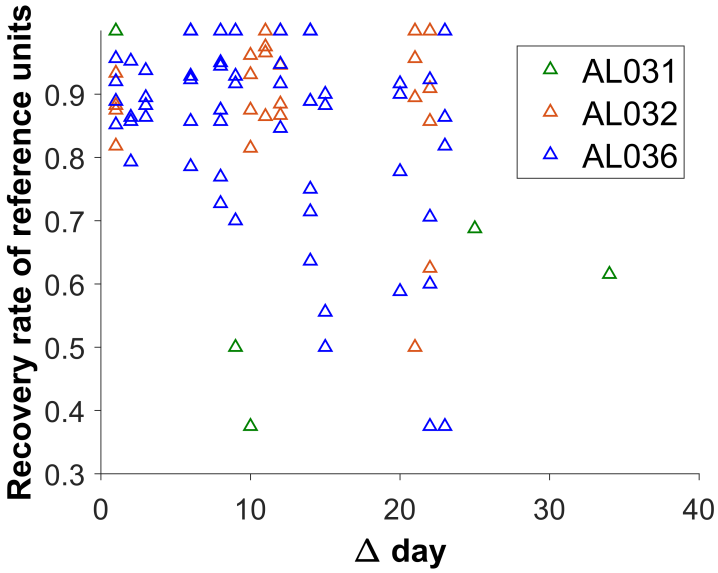

Fig. S9: The reference units recovery rate for spike sorted recordings spanning different days. Each triangle represents the matching results of two datasets. Animal AL031 has 6 sets of matching, with one outlier removed. Animal AL032 has 24 sets of matching. Animal AL036 has 60 sets of matching. The recovery quality becomes lower as datasets spans longer time.

8.6 Example reference and putative chains

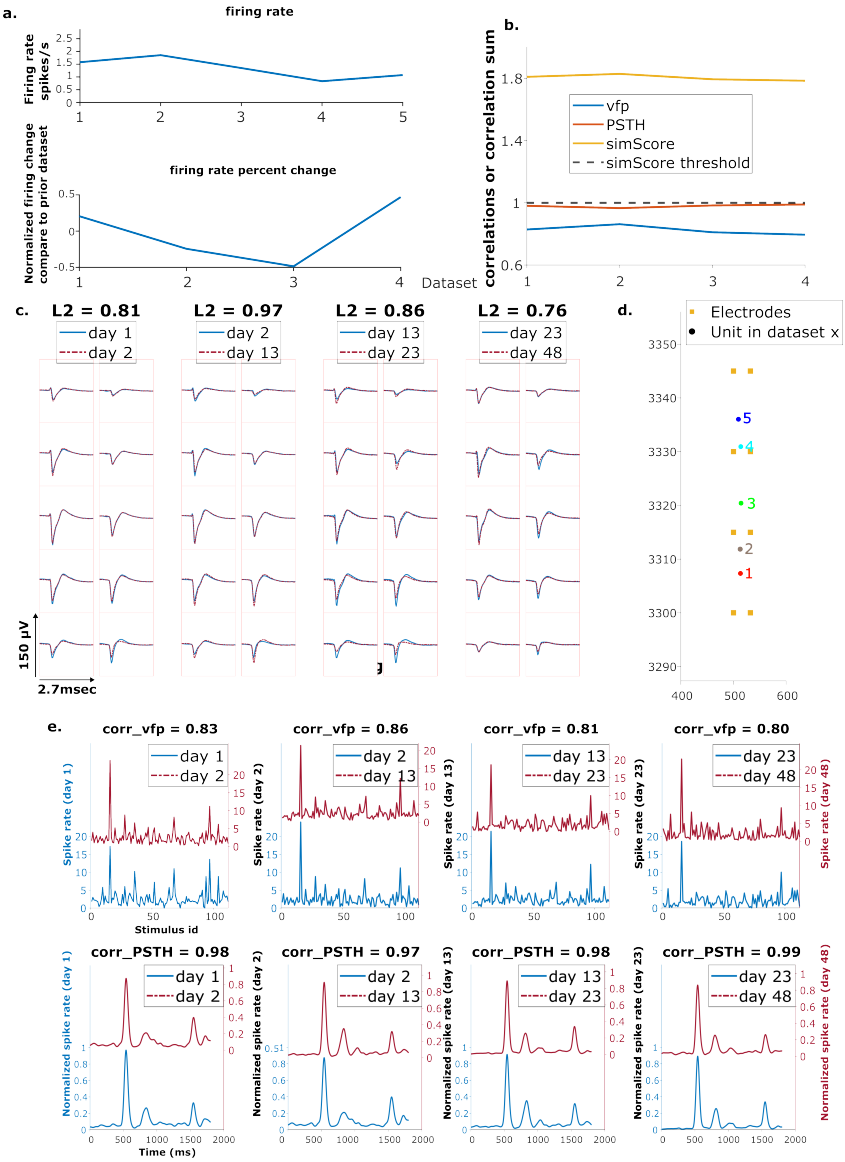

Fig. S10: An example of reference chain. a. Above: Firing rates of this neuron on each day. Below: Firing rate percent change compared to the previous day. b. Visual response similarity (yellow line), PSTH similarity correlation (orange line), and visual fingerprint similarity correlation (blue line). The similarity score is the sum of vfp and PSTH. The dashed black line represent 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 represent a unit in a day. The orange squares represent the electrodes. e. The pairwise vfp and PSTH traces of this unit.

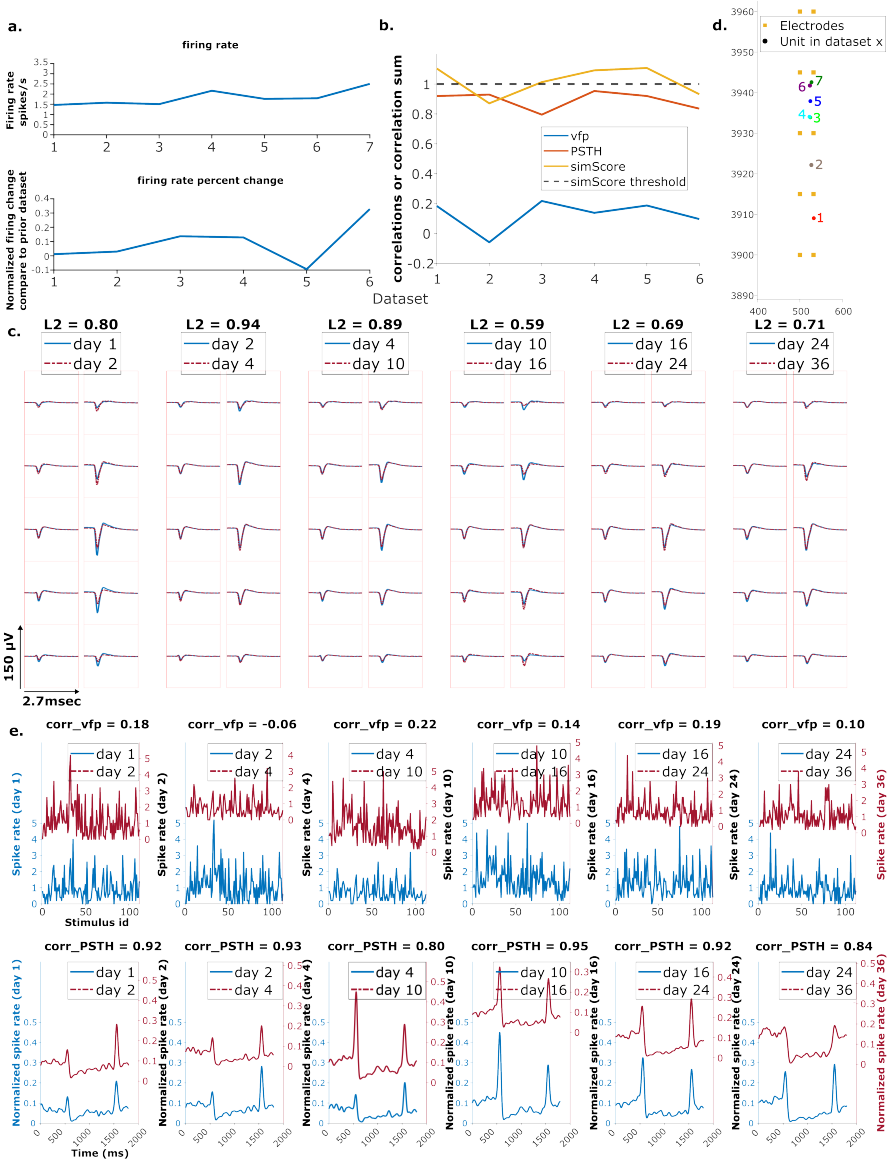

Fig. S11: An example of putative chain. Order is the same as above.

8.7 Reference unit counts 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 show in Fig.S12.

To show that days 1-2 (first two rows) are significantly different from days 3-9, we use 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, $p = 0.92$).

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

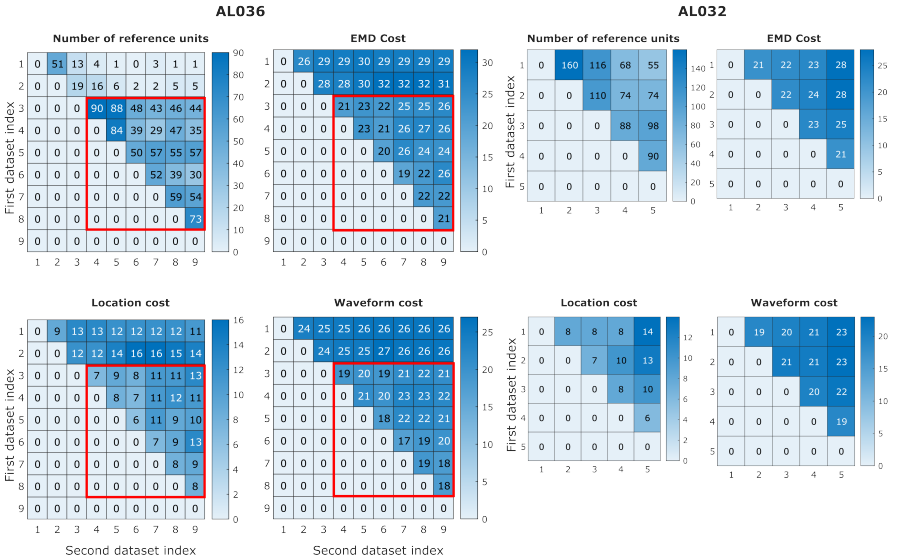

Fig. S12: Reference unit counts and normalized EMD cost for each pair of datasets recorded by the same shank. For animal AL036 (left), we excluded the first two datasets and all of their matching results (first two rows of each matrix on the left) based on the reference unit counts. Following analysis on their matching EMD cost, location-only cost and waveform-only cost suggest a big difference compared to the following days (datasets in the red rectangles). We infer that the first two datatsets have recorded from a different population from the later days. The other matrices show the similar information for animal AL032 for reference. To show the relative size of EMD cost in related datasets versus unrelated datasets, we calculated the cost between unrelated datasets with similar number of units (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).

8.8 Recovery rate vs. the EMD cost

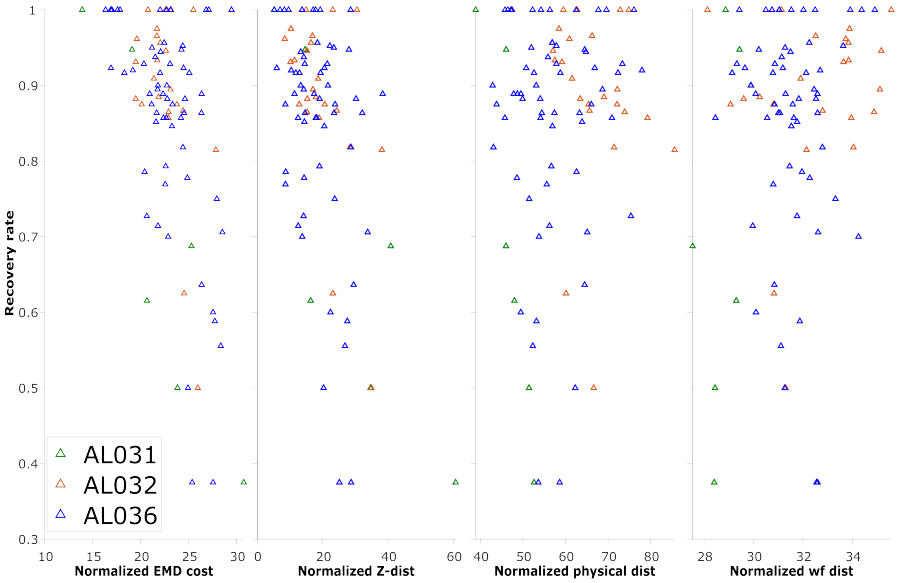

Fig. S13: The normalized EMD cost (unitless), z distance (μm), physical distance (μm), and waveform distance (unitless) and the corresponding recovery rate in pairwise match of all recording to all other recordings, shank by shank. Each triangle represents the recovery rate of two datasets. Animal AL031 has 6 sets of matching, with one outlier removed. Animal AL032 has 24 sets of matching. Animal AL036 has 60 sets of matching. Overall, most of the datatset with high recovery rate has per-unit EMD cost around 50. Note that the EMD cost is not predictive of recovery rate.

8.9 Reference unit ratio

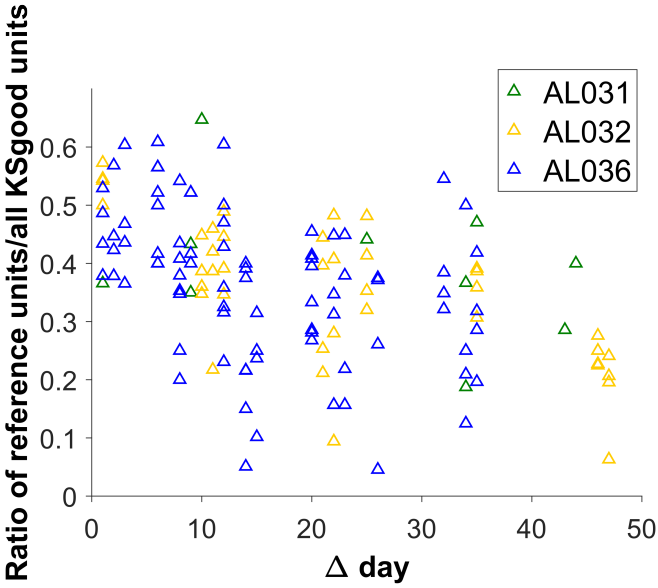

Fig. S14: The reference units to KSgood units ratio decreases for datasets with larger time interval. But the variability of the number of reference units is generally large for datasets with the same time interval.

8.10 Parameter tuning: L2-weight vs. Recovery rate

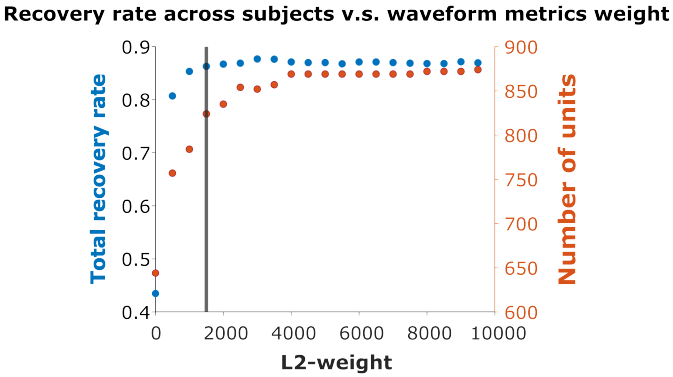

Fig. S15: We varied weight ω in equation 3 used to combine physical and waveform distances at an increment of 500. The vertical line indicate weight = 1500, where the overall recovery rate = 86.29%. The maximum recovery rate = 87.68% occurs at weight = 3000. We chose weight = 1500 for all subsequent analysis.

8.11 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 animal AL021 and AL032 is shown here.

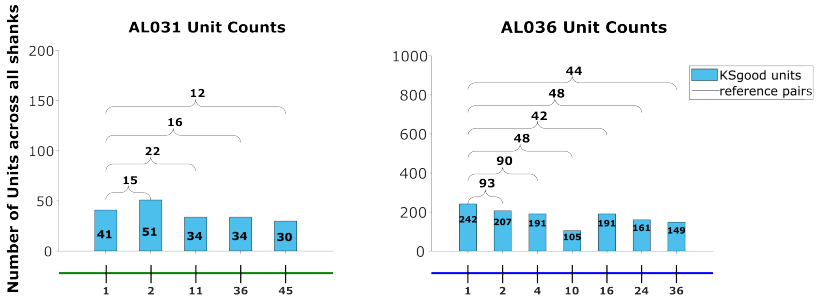

Fig. S16: The Kilosort-good and reference unit counts for the animals. AL031 and AL036 as shown for animal AL032 in Figure 5.

References

[1] J. M. Carmena, C. S. H., M. A. Lebedev & Nicolelis, M. A. L. Stable ensemble performance with single-neuron variability during reaching movements in primates. *J. Neurosci.* **25**, 10712–10716 (2005). https://doi.org/10.1523/JNEUROSCI.2772-05.2005 .

[2] D. Huber, S. P. D. H. O. J. S. W. L. T. T. G. O. L. L. L. . K. S., D. A. Gutnisky. Multiple dynamic representations in the motor cortex during sensorimotor learning. *Nature* **484**, 473–478 (2012). https://doi.org/10.1038/nature11039 .

[3] William A Liberti III, L. N. P. D. C. L. D. P. L. G. G. T. V. D. N. K. C. L. . T. J. G., Jeffrey E Markowitz. Unstable neurons underlie a stable learned behavior. *Nat Neurosci* **19**, 1665–1671 (2016). https://doi.org/10.1038/nn.4405 .

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

[5] A. K. Dhawale, S. B. W. V. A. N. E. K., R. Poddar & Ölveczky, B. P. Automated long-term recording and analysis of neural activity in behaving animals. *eLife* **6**, e27702 (2017). https://doi.org/10.7554/eLife.27702 .

[6] K. T. Jensen, A. K. D. S. B. E. W., N. Kadmon Harpaz & Ölveczky, B. P. Long-term stability of single neuron activity in the motor system. *Nat Neurosci* **25**, 1664–1674 (2022). https://doi.org/10.1038/s41593-022-01194-3 .

[7] Steinmetz, N. A. *et al.* Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science* **372**, eabf4588 (2021). https://doi.org/10.1126/science.abf4588 .

[8] Luo, T. Z. *et al.* An approach for long-term, multi-probe neuropixels recordings in unrestrained rats. *eLife* **9** (2020). https://doi.org/10.7554/eLife.59716 .

[9] Harris, K. D., Quiroga, R. Q., Freeman, J. & Smith, S. L. Improving data quality in neuronal population recordings. *Nature Neuroscience* **19**, 1165–1174 (2016). https://doi.org/10.1038/nn.4365 .

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

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

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

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

[14] R. Q. Quiroga, Z. N. & Ben-Shaul, Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation* **16**, 1661–1687 (2004). https://doi.org/10.1162/089976604774201631 .

[15] E. Chah, A. D.-C. J. J. H. M. S. M. O., V. Hok & Reilly, R. B. Automated spike sorting algorithmbased on laplacian eigenmaps and k -means clustering. *J. Neural Eng.* **8**, 016006 (2011). https://doi.org/10.1088/1741-2560/8/1/016006 .

[16] Pachitariu, M., Steinmetz, N., Kadir, S., Carandini, M. & Kenneth D., H. Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels (2016).

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

[18] Jun, J. J. *et al.* Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017). https://doi.org/10.1038/nature24636 .

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

[20] Tolias, A. S. *et al.* Recording chronically from the same neurons in awake, behaving primates. *Journal of Neurophysiology* **98**, 3780–3790 (2007). https://doi.org/10.1152/jn.00260.2007 .

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

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

[23] JinHyung Lee, H. S.-I. K. N. D. S. W. K. L. E. B. R. D. T. E. B. Y. J. K. N. B. A. K. G. G. E. C. D. C. L. P., Catalin Mitelut. Yass: Yet another spike sorter applied to large-scale multi-electrode array recordings in primate retina 10712–10716 (2020). https://doi.org/10.1101/2020.03.18.997924 .

[24] Jason E Chung, A. H. B. V. M. T. A. C. T. K. Y. L. K. G. S. S. H. F. L. M. F. L. F. G., Jeremy F Magland. A fully automated approach to spike sorting. *Neuron* **95**, 1381–1394.e6 (2017). https://doi.org/10.1016/j.neuron.2017.08.030 .

[25] Chung, J. E. *et al.* High-density, long-lasting, and multi-region electrophysiological recordings using polymer electrode arrays. *Neuron* **101**, 21–31.e5 (2019). https://doi.org/10.1016/j.neuron.2018.11.002 .

[26] Vasil'eva, L. N. *et al.* Long-term recording of single neurons and criteria for assessment. *Neuroscience and Behavioral Physiology* **46**, 264–269 (2016). https://doi.org/10.1007/s11055-016-0227-8 .

[27] U. Rokni, E. B., A. G. Richardson & Seung, H. S. Motor learning with unstable neural representations. *Neuron* **54**, 653–666 (2007). https://doi.org/10.1016/j.neuron.2007.04.030 .

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

[29] Colonell, J. ecephys spike sorting. https://github.com/jenniferColonell/ecephys_spike_sorting (2018).

[30] Cohen, S. Finding color and shape patterns in images (stanford university, palo alto, 1999) (1999) .

[31] N. P. Bertrand, J. L. P. B. D., A. S. Charles & Rozell, C. J. Efficient tracking of sparse signals via an earth mover's distance dynamics regularizer. *IEEE* **27**, 1120–1124 (2020). https://doi.org/10.1109/LSP.2020.3001760 .

[32] Boussard, J., Varol, E., Lee, H. D., Dethe, N. & Paninski, L. Three-dimensional spike localization and improved motion correction for neuropixels recordings. *NeurIPS Proceedings* (2021) .

[33] Britton A. Sauerbrei, J. D. C. M. M. W. G. M. K. N. V. B. M. K. B. . A. W. H., Jian-Zhong Guo. Cortical pattern generation during dexterous movement is input-driven. *Nature* **577**, 386–391 (2020). https://doi.org/