

ACORN: Adaptive Consensus of Repeated Clusterings identifies pyramidal neurons and interneurons in the primate motor cortex based on extracellular spike waveforms

SOLTESZ, Zoltan, TENORIO, Viviane, TOHYAMA, Takamichi, BAKER, Anne, BAKER, Stuart and KRASKOV, Alexander

Available from Sheffield Hallam University Research Archive (SHURA) at:

<https://shura.shu.ac.uk/38094/>

This document is the Published Version [VoR]

Citation:

SOLTESZ, Zoltan, TENORIO, Viviane, TOHYAMA, Takamichi, BAKER, Anne, BAKER, Stuart and KRASKOV, Alexander (2026). ACORN: Adaptive Consensus of Repeated Clusterings identifies pyramidal neurons and interneurons in the primate motor cortex based on extracellular spike waveforms. eNeuro. [Article]

Copyright and re-use policy

See <http://shura.shu.ac.uk/information.html>

Research Article: New Research | Sensory and Motor Systems

ACORN: Adaptive Consensus of Repeated Clusterings identifies pyramidal neurons and interneurons in the primate motor cortex based on extracellular spike waveforms

<https://doi.org/10.1523/ENEURO.0394-25.2026>

Received: 14 October 2025

Revised: 10 July 2026

Accepted: 22 September 2026

Copyright © 2026 Soltesz et al.

This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](#), which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

This Early Release article has been peer reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.eneuro.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

Manuscript Title:

ACORN: Adaptive Consensus of Repeated Clusterings identifies pyramidal neurons and interneurons in the primate motor cortex based on extracellular spike waveforms

Abbreviated Title:

Consensus clustering of cortical neurons

Author Names and Affiliations in order as they would appear in the published article:

Zoltan Soltesz, Faculty of Medical Sciences, Newcastle University, Newcastle, NE2 4HH, United Kingdom

Viviane SGM Tenorio, School of Engineering and Built Environment, Sheffield Hallam University, S1 1WB, United Kingdom

Takamichi Tohyama, Department of Rehabilitation Medicine, Fujita Health University, 470-1192, Japan

Anne ME Baker, Faculty of Medical Sciences, Newcastle University, Newcastle, NE2 4HH, United Kingdom

Stuart N Baker, Faculty of Medical Sciences, Newcastle University, Newcastle, NE2 4HH, United Kingdom

Alexander Kraskov, Faculty of Medical Sciences, Newcastle University, Newcastle, NE2 4HH, United Kingdom

Author Contributions:

AK and ZS designed research; VSGMT, TT, AMEB, SNB, AK, ZS collected data; ZS and AK analyzed data; ZS wrote the paper with contributions from coauthors; AK edited the paper

Correspondence should be addressed to:

Dr Alexander Kraskov, Henry Wellcome Building, Faculty of Medical Sciences, Newcastle University, NE2 4HH, United Kingdom, Sasha.Kraskov@newcastle.ac.uk

Number of Figures: 7

Number of Tables: 9 primary (plus 4 extended tables)

Number of Multimedia: 0

Number of words for Abstract: 247

Number of words for Significance Statement: 95

Number of words for Introduction: 748

Number of words for Discussion: 1499

Acknowledgements: This work was supported by the Biotechnology and Biological Sciences Research Council [grant numbers BB/Y000625/1, BB/V00896X/1], the National Institutes of Health [grant number R01NS119319], AIRA [SCNI-PR01-P09] and an EU grant to the In2PrimateBrains consortium [grant number 956669]. The authors are grateful for spike-sorting code shared by Benjamin Dann and for constructive feedback from Roger Lemon.

Conflict of Interest: Authors report no conflict of interests.

Funding sources: BBSRC [BB/Y000625/1, BB/V00896X/1], NIH [R01NS119319], ARIA [SCNI-PR01-P09] and an EU grant to the In2PrimateBrains consortium [956669].

ACORN: Adaptive Consensus of Repeated Clusterings

identifies pyramidal neurons and interneurons in the primate motor cortex
based on extracellular spike waveforms

Abstract

Understanding how neural circuits generate behaviour requires knowledge of both neural activity patterns and the cell types they originate from. While the scale of neuronal recordings has dramatically increased since the advent of dense multicontact probes, accurate identification of cell types remains elusive. Traditional approaches have used spike width to attempt a classification, with narrow spikes assumed to be inhibitory interneurons. However, fast pyramidal tract neurons (PTNs, known to be excitatory) with thin spikes and interneurons with wide spikes challenge these assumptions. Here we accomplish three related objectives. First, we produce a ground truth dataset of extracellular spike shapes from electrophysiologically identified pyramidal neurons from primary (M1) and supplementary motor areas (SMA) and optogenetically identified interneurons in M1 in female macaque monkeys. Next, we show that spike width correlates with antidromic latency not only in M1 but also in the SMA, suggesting that it is a physiologically meaningful measure across the motor cortex. Finally, we introduce our unsupervised ensemble clustering method, "ACORN", and characterise its performance on our neocortical macaque dataset, the C4 cerebellar database, and a juxtacellularly recorded mouse dataset, demonstrating strong out-of-the-box performance across species, brain areas, and recording techniques. ACORN can cluster neuronal waveforms without training or tuning, and provides data-derived confidence values for its predictions. Our ground truth data can be used as a reference to help identify or verify clusters in novel datasets. Together, our method and database provide a streamlined foundation for electrophysiological cell classification without necessitating changes to recording paradigms.

Significance Statement

High-density extracellular recordings are increasingly the norm in neuroscience, but accurate and non-invasive identification of cells remains a challenge. Here we provide a database of action potential spike shapes from pyramidal tract neurons and interneurons of the macaque supplementary motor area, whose identities we verified using nerve stimulation and optogenetics, respectively. We introduce the neuron clustering method “ACORN”, and showcase its robustness by evaluating its performance on a variety of distinct datasets. Our training-free clustering technique is designed to support experimentalists, while our ground truth database can enhance and validate future datasets and method development.

Introduction

Different classes of neurons perform distinct roles within cortical networks. To understand how information is processed, knowledge of activity levels alone may not be sufficient; identifying cell types is a critical part of the puzzle. For example, detecting a high firing rate of local inhibitory interneurons would imply a different type of neural computation than the same pattern observed in excitatory pyramidal neurons that project to distant targets.

The prospect of reliably classifying cells based on their extracellular electrophysiological properties is appealing for several reasons. Firstly, using the same recordings to classify neurons and measure their activity patterns would make experiments more straightforward to perform. Additionally, it would allow building cell-specific computational models of complex animal behaviour based on recordings in awake animals, without the need for additional, often postmortem, techniques required to confirm cell identity. Furthermore, it would facilitate more accurate targeting of cells of interest, by making it possible to adjust electrode position based on live information about cell types. Finally, it would enable real-time adjustments of

electric, optogenetic or pharmacological network manipulations in awake behaving animals as a function of observed cell type-specific activity.

Previous works suggested that neocortical interneurons have short action potentials, which could be used to distinguish them from pyramidal neurons that spike more rarely but with longer durations. Many subsequent studies have relied on action potential durations as a way to distinguish between putative interneurons and excitatory pyramidal cells (for an overview, see Table 2 in Vigneswaran et al., 2011). However, the considerable diversity of cortical interneurons, with multiple morphological subtypes clustering into the same electrophysiological classes (Zaitsev et al., 2009), raised questions about the reliability of such an approach.

The picture is further complicated by different brain regions having distinct compositions of cell types. For example, giant Betz cells are present in the fifth layer of the primary motor cortex (Rivara et al., 2003), but not in secondary areas, while the supplementary motor area (SMA) is thought to contain a higher proportion of small-sized cells (Wiesendanger et al., 2007). As cell size correlates with conduction velocity (Lee et al., 1986), which in turn has been shown to correlate with action potential duration (Vigneswaran et al., 2011), this suggests that the reliability of cell type classification using extracellular spike shapes will depend on the brain region under investigation. For example, in the primary motor cortex (M1) of macaque monkeys, large pyramidal tract neurons have been shown to exhibit spikes with short durations that overlap with those reported for interneurons (Vigneswaran et al., 2011). Additionally, there are significant cross-species differences as well, with the primate cerebrum exhibiting a diverse variety of interneuronal classes (Schmitz et al., 2022).

To address some of these limitations, more sophisticated approaches used not only a single feature, such as action potential duration, but a wider range of information through the unsupervised clustering of spike shape waveforms (Trainito et al., 2019; Lee et al., 2021). While these methods succeeded in identifying distinct cell classes based on neural activity,

scrupulous assessment of the reliability of extracellular waveform-based cell classification would require a ground truth dataset of recordings from neurons with verified identifies.

A recent study used optogenetic and pharmacological tools to obtain a curated library from the cerebellum, which they then used to train and validate their semi-supervised classifier (Beau et al., 2025). The authors subsequently showed that multiple extracellular metrics fed to a deep learning classifier produced results consistent with expert identification of neuron types in the macaque cerebellum (Herzfeld et al., 2025). However, the applicability of the method to brain regions with less well-defined cell types and anatomical layers is unclear. New ground truth datasets are needed to facilitate the investigation of other brain areas.

In this study we establish a validated spike shape dataset of interneurons and pyramidal neurons from M1 and SMA in the macaque cortex, and directly assess the performance of traditional neural classification. We corroborate earlier results showing a correlation between antidromic latency (ADL) and spike duration in M1, and show that the same relationship also holds in SMA. We present a distribution of optogenetically and antidromically verified interneuron and pyramidal neuron spike durations, demonstrating a substantial overlap between the two. To overcome this, we develop and showcase our unsupervised consensus-based clustering method, ACORN. We evaluate its performance on three distinct datasets, and show that the data-driven confidence values it provides are informative. Finally, we discuss promising directions for refining electrophysiological neuron classification, in particular the integration of spatial and temporal firing patterns.

Methods

Recordings were obtained from six adult female purpose-bred pair-housed macaques (Macaca mulatta: BE, BL, WA, YG; Macaca fascicularis: GR, PR). All animal procedures were performed in accordance with the Newcastle University Animal Care Committee's

regulations. Data from awake monkeys analyzed in this study were obtained incidentally as part of other scientific projects that involved recordings of optogenetic excitation or pyramidal tract stimulation. Acute recordings were obtained as a secondary objective following the conclusion of the main experiments on animals terminally anesthetized for other projects.

Anesthesia was induced using a balanced combination of ketamine (10 mg/kg i.m.), medetomidine (3 µg/kg i.m.) and midazolam (0.3 mg/kg i.m.), and supplanted with sevoflurane during surgical preparation prior to the start of the experiment (0-2% in 100% oxygen). As sevoflurane is known to suppress neuronal activity (Aksenov et al., 2019), we maintained general anesthesia using a combination of alfentanil (24-73 µg/kg/h i.v.), ketamine (6-15 mg/kg/h i.v.), and midazolam (0.3-0.4 mg/kg/h i.v.) during our recordings. Hartmann's solution (5 ml/kg/h i.v.) was given to prevent dehydration, and methylprednisolone (5.4 mg/kg/h i.v.) to reduce cerebral oedema. A tracheotomy was made to allow positive-pressure artificial ventilation. Arterial blood pressure was monitored through a central arterial line inserted via the carotid artery. The bladder was catheterized. The animal's temperature was maintained with the help of a warm air blanket and a thermostatic heat pad. Anesthetic monitoring consisted of pulse oximeter, heart rate, arterial blood pressure, core and peripheral temperature, and end-tidal CO₂ measurements.

A multi-contact recording probe (a SiNAPS 1024-electrode NHP probe for monkey BE (Angotzi et al, 2025), and an Atlas 32-contact E32Tri+R-25-S1-L10 NT probe for monkeys GR and PR) was mounted on a Kopf micromanipulator, angled to penetrate the cortical region of interest normal to the surface, and slowly advanced using a computer-controlled NaN drive. The PT stimulation electrodes (tungsten insulated with parylene, Microprobes Inc., part number LF501G with a tip impedance of ~ 10 kΩ) were implanted into the medullary pyramidal tract under physiological control (antidromically evoked cortical potentials). The PT search stimuli consisted of two shocks delivered 10 ms apart (intensity < 1mA; charge-balanced biphasic stimuli, 70 µs per phase).

To calculate the waveforms of putative PTNs, we first applied a 60 Hz high-pass FIR digital filter to 30ms intervals centered around recorded antidromic stimulation timings to remove any residual 50 Hz interference from the signal. Averaging sweeps from the same channels together (**Fig. 1A**), we obtained user-supplied initial starting points for base, trough, and peak timings (**Fig. 1B**), which were subsequently automatically refined to the local extrema within 160 us windows. We selected the channel with the greatest combined difference in baseline-to-trough and peak-to-trough voltages. For 1024-contact SiNAPS probes we averaged together the two channels that showed the strongest responses rather than relying on only a single contact. We interpolated the signal from the selected channel to a rate of 100 KHz using a Modified Akima spline in order to achieve more precise temporal alignment of troughs across the waveforms.

Cells were classified as PTNs when they responded with a consistency of at least 90% across sweeps at a specific antidromic latency and an overall mean absolute amplitude of at least 25 μ V. Pyramidal tract neurons are expected to respond consistently to above-threshold PT stimulation, except in rare cases due to collisions with spontaneous spikes. Consistent with this, we found non-responses to be rare for cells in our PTN dataset, with a median response rate of 97% (**Figure 3-1**). We considered as non-responses deflections that fell below the last negative probability-change prior to the mode in the kernel density estimated amplitude probability distribution function, across sweeps, of the cell in question (**Fig. 1C**). The bandwidth for kernel smoothing was calculated adaptively per cell using the Improved Sheather-Jones plug-in method (Botev et al., 2010). We also considered as non-responses spikes with trough timing jitter in excess of 100 μ s relative to the mean of the included sweeps (**Fig. 1D**).

Cells were classified as interneurons based on their optogenetic responses. An AAV8-mDlx-Chronos-tdTomato vector (8.06×10^{12} vg/ml) was injected at 16 points in M1 over 240 minutes across 2 days for a total volume of 32 μ L (Monkey BL) or injected at 12 points in M1 over 180 minutes across 2 days for a total volume of 48 μ L (Monkeys WA and YG). The

mDlx enhancer is reported to be specific for GABAergic interneurons in the forebrain (Dimidschstein et al., 2016), while Chronos is a highly sensitive general use channelrhodopsin with fast kinetics (Klapoetke et al., 2014). Optogenetic responses were first observed after 9 weeks of protein expression.

Concurrent with extracellular recordings using a 5 channel Mini Matrix Eckhorn drive (Thomas Recordings), blue light (470 nm) was delivered in short intervals (20-100ms) at an irradiance of 0.3 mW/mm². Four optical fibers altogether flanked a single recording electrode in an equidistant cross-like arrangement (0.5 mm spacing), with the electrode lying 0.3 mm deeper than the fiber level. To detect and cluster spikes in optogenetic recordings, we used a modified version of WaveClus (Quiroga et al., 2004), in which wavelet-based feature extraction was substituted for UMAP and superparamagnetic clustering for the Leuven method. We then compared firing rates immediately prior to, and during, optogenetic illumination, across intervals of 50ms, except where optogenetic stimulation lasted shorter than this; then the interval was set to coincide with the duration of the illumination. Where one-tailed Wilcoxon signed-rank tests indicated a significantly higher firing rate during illumination ($p < 0.05$; see **Figure 3-1** for exact p-values), and optogenetic response latency was shorter than 10ms, we concluded that the cell in question was an interneuron.

Two ms waveform snippets were extracted around detected spikes, with troughs aligned to 0.3ms into the spike window as recommended by the WaveMAP protocol (Lee et al., 2023). We calculated signal-to-noise ratios (SNR) for each unit as the mean waveform amplitude divided by twice the standard deviation of the background noise on the channel (Lee et al., 2023), and excluded units with SNRs below 2.7. To normalize the waveforms, we subtracted the median, rather than the mean, of the spike window, because we found the former better-suited to bringing the baseline to 0 μ V in the case of large-amplitude spikes. To evaluate the quality of neuron class predictions, we used Adjusted Mutual Information (AMI) as a measure that compensates for chance effects (Vinh et al., 2010): AMI takes on a value of 0

when the mutual information between two clusterings is at chance level, and 1 at perfect agreement.

For some of the PTNs, antidromic spike shapes might be distorted by activity synchronous with PT stimuli (early PT volley or other PTNs; see e.g. **Fig. 3B**). Because this activity is time-locked to the PT stimulus, it will not be averaged out across sweeps. To address this issue, we employed “pattern-matched noise imputation”. In brief, we examined the signal outside the temporal extent of the spike waveforms, and upon finding “breakpoints” that indicated such distortions, we replaced those intervals with noise derived from the same sweep in a way that does not introduce sharp boundaries. A snippet of the code performing this procedure (“cleanspike”) is shared as supplementary material (**Extended Data 1**). For consistency, we applied this procedure to all our units: both PTNs and interneurons.

To calculate spike shape characteristics for a unit, we used the waveform obtained by averaging across valid sweeps. We measured the spike width as the period between the trough and the first subsequent datapoint at which the time-derivative of the voltage signal turns negative. We calculated spike amplitudes as the trough voltage relative to the median of the spike window. For PTNs, we measured ADL as the delay between a nerve stimulation event and the start of the consequent negative deflection detected in the cortex (“base timing” in **Fig. 1B**), averaged across valid sweeps.

We developed an unsupervised clustering method, ACORN. We started by dimension-reducing the mean spike waveforms using an optimized form of UMAP (McInnes et al., 2018) that has been shown to improve downstream clustering performance on a variety of datasets: MST-min with Path Neighbors (Dalmia and Sia, 2021). To generate the initial set of constituent clusterings, we employed HDBSCAN (McInnes et al., 2017). We sampled a large region of the parameter-space, generating $3*5*5*3*5 = 1125$ individual clusterings, one for each set of sample values (**Table 1**). We verified that overall results remained robust within this parameter space (**Fig. 7-1**). All other parameters were kept at their default values.

280 We then applied our SEED (Sequential Election of Exemplar Delegates) algorithm to turn
 281 multiple HDBSCAN clusterings into a single consensus result. Based on the above, we
 282 obtained a matrix of N neurons and M constituent clusterings. Notably, HDBSCAN does not
 283 pre-set the number of clusters it generates, so different constituent clusterings can end up
 284 with varying numbers of clusters. Additionally, HDBSCAN can designate potential outliers as
 285 “unassigned” instead of necessarily including them in one of the clusters (**Fig. 2**). SEED is
 286 based on the idea that representative members (exemplars) of each final cluster should be
 287 similar to members of their own cluster *but dissimilar to other exemplars*. To identify
 288 exemplars, we use the following iterative algorithm (see the accompanying code in
 289 **Extended Data 1** for the exact implementation).

290

291 For each neuron i , we calculate a “seed metric”, S_i , accumulating information across each
 292 clustering m . Let $C_{i,m}$ be a cluster into which neuron i was assigned in the m -th constituent
 293 clustering. For each neuron i and clustering m , we calculate $X_{i,m}$, the number of already-
 294 identified exemplars that belong to cluster $C_{i,m}$. If $X_{i,m}$ is zero (as by definition would be for
 295 the first exemplar), then we increase the value of S_i by the cluster size of $C_{i,m}$ (i.e. number of
 296 other neurons assigned to the same cluster as neuron i in clustering m). This is the *similarity*
 297 part of the metric (**Table 2**).

298

299 If $X_{i,m}$ is greater than zero, then a *dissimilarity* term (a negative component) is applied to S_i ,
 300 to ensure that exemplars are distinct from each other. In that case, we subtract from S_i a
 301 value of $X_{i,m} N / K / 2$, where K is the maximum number of clusters in constituent clusterings.
 302 We then continue this procedure over each neuron i and constituent clustering m , adding to
 303 and subtracting from S_i as described (**Table 3**). We thus calculate the S metric for each
 304 neuron in the dataset. If there are positive S values, the neuron with the highest value is
 305 chosen as an exemplar. We then recalculate the S values for each neuron, but with the new
 306 exemplar also influencing the results. We repeat this procedure until no positive S value is
 307 obtained (**Table 4**).

308

309 These exemplar neurons are the seeds of the final clusters. While constituent clusterings
310 may vary in the number and pattern of cluster labels they assign, the same neurons remain
311 the designated exemplars regardless of the constituent clustering under consideration.

312 Intuitively, the *similarity* component of the SEED algorithm promotes an arrangement where
313 exemplars are part of large clusters. At the same time, the *dissimilarity* component
314 encourages each exemplar to belong to a distinct cluster in each constituent clustering.

315 However, in practice, in some constituent clusterings, exemplar neurons may happen to be
316 unclassified or two or more exemplars may sometimes belong to the same cluster.

317

318 Each remaining neuron needs to be distributed into one of the final clusters corresponding to
319 one the L exemplars, or alternatively to be designated as “uncertain” (cluster $L+1$). We start
320 with V , an N -by- $(L+1)$ matrix initialized with zeros tallying the votes. For each neuron i and
321 each exemplar neuron k , we iterate over each constituent clustering m . Let w be the number
322 of exemplars that belong to the same cluster to which neuron i belongs in constituent
323 clustering m . If $w \geq 1$, we increment $V_{i,l}$ by $1/w$. If $w=0$, we increment $V_{i,L+1}$ by 0.5. The
324 motivation for the lower increment in the latter case is to ensure that designations of
325 “uncertain” have a lower impact than clashing cluster designations when deriving confidence
326 values from the matrix V in the following step.

327

328 For each neuron i , we then examine its final vote count $V_{i,l}$ across the $L+1$ clusters (**Table 5**).
329 Neuron i is assigned to cluster l with the highest $V_{i,l}$ score. If the highest $V_{i,l}$ score is at $l=L+1$,
330 then neuron the classification of the neuron is “uncertain”. The confidence value associated
331 with the assignment is the ratio between the highest and second-highest $V_{i,l}$ scores. Neurons
332 whose assignment confidence values exceed the threshold parameter (2.0 unless otherwise
333 specified) form the overall consensus clustering result.

334

335

336 **Code Accessibility**

338 A snippet of the code performing this procedure (“cleanspike”) is shared as supplementary
339 material (<https://github.com/zolts/ACORN>). For consistency, we applied this procedure to
340 all our units: both PTNs and interneurons.

341 To identify exemplars, we use the following iterative algorithm (see the accompanying code
342 at <https://github.com/zolts/ACORN> for the exact implementation).

343 The ACORN code, as well as an implementation of the “pattern-matched noise imputation”
344 used in this paper, are freely available (<https://github.com/zolts/ACORN>). The code is not
345 OS-specific. Here, it was run on a Dell Precision 7920 PC with a Microsoft Windows 10
346 Enterprise operating system, using Python 3.8 (Python Software Foundation) and MATLAB
347 R2025b (MathWorks).

348

349 **Results**

350

351 To establish ground truths, we identified PTNs based on antidromic stimulation (**Fig. 3A**).
352 Multi-contact probes allowed us to distinguish between spikes from spatially separated cells
353 within recordings (**Fig. 3B**). We used the shapes of reliably evoked antidromic spikes (**Fig.**
354 **3C**) for neuronal classification. We identified interneurons through optogenetics. We first
355 spike sorted all recorded neurons on a given electrode and then verified the responses of
356 individual clusters to optical stimulation (**Fig. 4**).

357

358 We confirmed the finding that spike width correlates with antidromic latency (ADL) in M1,
359 and found that the same correlation also holds in SMA (**Fig. 5**). This suggests that trough-to-
360 peak spike width is a physiologically meaningful measure. We observed longer trough-to-
361 peak PTN widths in M1 than some previous works (Vigneswaran et al., 2011). This may be

due to differences in recording sites, macaque species, or the effects of general anesthesia, which can increase repolarization times (Jia et al., 2024).

We found a substantial overlap in spike widths between PTNs and interneurons (**Fig. 6A**), raising questions about the reliability of spike width as a distinguishing feature between the two classes. Many previous studies used a spike width value between 0.2 - 0.4 ms as the threshold for discrimination (reviewed in Vigneswaran et al., 2011). Careful choice of spike width threshold could affect classification performance. Hence, to determine how well such an approach would perform on our dataset, we calculated the Adjusted Mutual Information between predicted and true binary classifications with thresholds covering the width range used in the literature (0.2-0.4ms). This served as a blind test in the sense that no information about the true identities of the units was used in picking the exact threshold. We observed highly variable performance using the spike width-based approach ($AMI=0.15\pm0.07$ mean $\pm$ std), confirming that threshold choice strongly affects the results. This prompted us to seek a more robust approach to cell type classification.

Trough-to-peak duration is a salient feature, but not the only piece of information a spike waveform contains. We therefore asked if a feature-agnostic method could discriminate between these two classes of neurons more accurately than a naïve spike width-based approach did. We applied WaveMAP (Lee et al., 2023) to the normalized average waveforms of our units, clustering them using the Louvain method based on UMAP-derived graph-representation. (**Fig. 6B**). While WaveMAP succeeded in identifying a cluster heavily enriched for PTNs (false discovery rate of only 7% with our dataset), it had low specificity for interneurons, with a false discovery rate of 56% (**Table 6**). Despite the unbiased nature of the approach, we found a strong correlation between the spike width measure and the dimension-reduced forms of normalized waveforms represented by UMAP values (**Fig. 6C,D**).

390 While the unbiased approach taken by WaveMAP showed promise, we identified a number
391 of aspects that could be improved: first, the Louvain method employed has been shown to
392 suffer from a “major defect” (Traag et al., 2019) that can result in arbitrarily badly connected
393 communities. Second, WaveMAP draws hard boundaries, assigning each neuron
394 exhaustively as part of a cluster. This approach may work well with datasets containing
395 distinct and well-represented clusters, but could be sensitive to the inclusion of outliers or
396 neurons whose waveforms are ambiguous or intermediate. Third, while it clusters all units, it
397 provides no confidence values for those assignments: archetypal members of a group and
398 best guesses will be equally part of the same cluster.

399

400 We aimed to provide a method that would allow researchers to cluster neurons based on
401 their waveforms robustly and reliably, without training or parameter-tuning. An approach that
402 produces an overall consensus result that incorporates information from an ensemble of
403 constituent clusterings over the same dataset has several advantages:

- 404 1. The consensus may be a better fit than any of its input clusterings, as it can piece
405 together the overall picture from identified "building blocks". Differences in the input
406 clusterings can be a positive factor in this context, akin to looking at a scene from different
407 angles.
- 408 2. Information from clusterings can be integrated robustly even when they contain different
409 numbers of clusters.
- 410 3. Confidence in neurons belonging to their assigned cluster can be quantified. This way, a
411 researcher can choose a confidence threshold according to whether minimizing false
412 inclusions or false exclusions is a greater priority in the paradigm.

413

UMAP and HDBSCAN feature a number of parameters that can affect their results. Rather than picking a particular set of parameter values tuned to our dataset, we employed an unbiased approach: we iterated over a large region of the parameter-space, producing a variety of different clusterings. Our method, ACORN (Adaptive Consensus of Repeated Clusterings), was designed to find a robust consensus of these constituent clusterings (see Methods). We evaluated cluster predictions against the ground truth along two axes: the proportion of neurons labelled, and the accuracy of those labels. The results of the ACORN consensus were better than the average constituent clustering along both axes (**Fig. 7A**). The consensus prediction achieved an AMI of 0.26 while labelling 83% of our units (**Table 7**).

To test the generalizability of our method, we applied ACORN to pre-processed data accompanying the recent PhysMAP paper (Lee et al., 2026). In particular, we used the juxtacellular ground truth dataset recorded from the mouse barrel cortex (Yu et al., 2019). This dataset labels excitatory (E), presumed parvalbumin-expressing (PV⁺), somatostatin-expressing (SOM⁺), and vasoactive intestinal polypeptide-expressing (VIP⁺) interneurons as separate classes. Using renormalized average waveforms, ACORN achieved an AMI of 0.52 while labelling 79% of the neurons (**Table 8, Fig. 7B**). Treating different layers as different classes, ACORN achieved an AMI of 0.36 while labelling the same 79% of neurons (**Table 8-1**).

To compare with PhysMAP, we calculated the Modified Adjusted Rand Indices (MARI) for our clustering results as well. This was 0.38 for the 8-class scenario at the default confidence threshold, and 0.31 when ACORN was forced to classify all units regardless of confidence levels (**Table 8-2**). On the same dataset, PhysMAP achieved a MARI of around 0.32 when using full multimodal information, or around 0.28 when using only the waveforms in an equivalent comparison (Fig. 2C in Lee et al, 2026).

To further test the robustness of our clustering technique, we applied it to data from the recently published C4 database (Beau et al, 2025). This dataset also derives from a different brain area, namely the cerebellum, and includes ground truth labels of five distinct cell types. By processing the waveform data in the same way and clustering them using the same sets of parameters as before, our method achieved an AMI of 0.66, labelling 46% of the dataset (**Table 9, Fig. 7C**).

Both high AMI and a high proportion of neurons labelled are desirable in clustering results. By adjusting the confidence threshold, it is possible to trade off the two measures against each other. To investigate this relation, we introduced a “performance” metric defined as a simple product of the AMI and proportion labelled values. Across the three different datasets, we observed that this performance score was stable around our default confidence threshold of 2.0, but declined at extreme values on the C4 dataset (**Fig. 7D**).

To examine the relation between our blindly obtained confidence values and ground truth agreement, we calculated the AMI values sequentially for the n neurons whose cluster assignments were most confident in the C4 dataset. We expected a group of high-confidence assignments to produce high AMI, and that adding increasingly lower confidence predictions to the set would gradually reduce the AMI. This is indeed what we observed (**Fig. 7E**). Notably, the 39 highest-confidence assignments achieved a perfect AMI of 1.00 on this dataset. Taken together, these results indicate that our data-driven confidence metric, derived without reference to the ground truth labels, meaningfully relates to clustering performance.

Discussion

A single parameter, such as spike width, cannot capture the totality of the electrophysiological properties of a cell, so it is unsurprising that some overlap can be

observed among different classes of neurons along what is effectively a one-dimensional projection of a complex feature space. What is more remarkable is that we find this overlap in the macaque cortex to be more significant than previously reported, with typical spike width thresholds used in previous works unable to select interneurons reliably.

We observed a notable difference in PTN and interneuron spike width distributions, with a corresponding asymmetry in classification performance. A majority of our PTNs exhibited long spike widths that make them easily distinguishable from interneurons, which tend to have shorter widths. However, the presence of “fast PTNs” made reliable identification of interneurons challenging, as the former can have similarly “thin spikes” as the latter. Here we extend this observation to the supplementary motor area (SMA), suggesting it as a more general feature of the motor cortex rather than a peculiarity of M1.

While we used an optogenetic approach to target GABAergic interneurons for our ground truth dataset, the group likely contains a mixture of both parvalbumin- (PV) and somatostatin-expressing (SOM) interneurons, which may have different waveform and spiking properties. In particular, a recent study suggests that SOM subtypes can exhibit spike durations ranging from less than 0.4 ms to over 2 ms (Morales et al., 2021), which challenges the idea that a singular spike width threshold, however well chosen, would allow for neuron subtypes to be reliably distinguished from each other.

What might affect the extent of the overlap observed in different studies? One important consideration is inter-species differences in gene expression and therefore in neuronal ion channel composition. While in rodents the fast potassium channel Kv3.1b is expressed in interneurons but only rarely and weakly in PTNs, in the macaque the same channel is present on both classes of neurons (Soares et al., 2017), resulting in more similar spike widths between the two. Relatedly, there is a large diversity of interneuron subtypes in primates (Krimer et al., 2005). A greater variety of interneuronal subtypes could occupy a

more substantial volume in electrophysiological feature-space, which could contribute to the greater overlap in PTN and interneuron spike widths observed in primates.

The performance of spike width-based classification heavily depends on the threshold value chosen, typically between 0.2-0.4ms. We characterized it by calculating the average Adjusted Mutual Information (AMI) of the range, representing a "typical" performance in the absence of guiding information. Setting this as a bar, we assessed the performance of the unsupervised classification technique WaveMAP on our dataset, and found it to be similar, as measured by AMI, as spike width-based classification. Furthermore, low-dimensional UMAP representations of the units strongly correlated with their spike width measures, underscoring that spike width is a salient piece of information represented in the spike waveform snippets in the context of binary classification.

Adjusted Mutual Information has several desirable properties for comparing clusterings. It corrects for chance effects: the expected AMI of a random clustering compared against the ground truth is zero regardless of the number of classes and neurons. AMI uses an information-theoretical approach that works well even on unbalanced datasets. However, one downside is its nonmetricity: a set of three clusterings can violate the triangle inequality if AMI is used as a distance measure. The Modified Adjusted Rand Index (MARI) also corrects for chance effects, but also violates the triangle inequality. The main difference between the two is that MARI is based on pairwise element comparisons, whereas AMI measures overall information overlap.

Normalized Information Similarity (NIS) is a related measure that obeys the triangle inequality; however, NIS violates the constant baseline property: random clusterings will have different expected NIS values against a ground truth dataset depending on the number of clusters. Overall, we consider AMI the most useful measure for assessing different clusterings against the *same* ground truth dataset – but with the notable caveat that AMI values across *different* datasets are not numerically comparable.

529 For any new method aiming to classify cells based on measurements of their
530 electrophysiological properties, a solid ground truth dataset is crucial to serve as its
531 foundation and to validate its reliability. Importantly, such a dataset should contain a good
532 representation of the cell types under investigation, to enable training and assessing suitable
533 classification models. In this study, we exploited antidromic stimulation and optogenetic
534 interventions to establish our validated dataset. Critically, this included both fast and slow
535 PTNs from two cortical different areas, with ADLs ranging from 1.0 ms to 8.6 ms in the SMA.

536

537 With a larger number of units in a dataset, more subtle differences beyond that of spike
538 width may become significant. In particular, the developers of WaveMAP recommend “300
539 high-quality units” for reliable classification (Lee et al., 2023). Thus, while we used the same
540 87-unit ground truth dataset as a test bed to compare different classification approaches,
541 WaveMAP may have performed better if supplied with a larger dataset. Another important
542 factor is the SNR cutoff imposed on the dataset: WaveMAP recommends a default value of
543 3.0, but suggests that the threshold be relaxed when investigating low-amplitude spiking
544 units (Lee et al., 2023). As our recordings included many well-isolated but low-amplitude
545 units, we settled on an SNR threshold of 2.7.

546

547 Given that reliably ascertaining neuron identities in vivo is a challenging task, and so verified
548 ground truth datasets are likely to remain relatively small in the foreseeable future, we
549 decided to explore steps we could take to improve classification results even with a limited
550 dataset. We introduced “ACORN”, an ensemble clustering technique acting on normalized
551 waveform snippets that uses UMAP for dimension reduction and HDBSCAN for
552 classification. We used the consensus of 1125 clustering iterations with different sets of
553 parameters to obtain the ultimate classification results. This had the added benefit of
554 providing a confidence measure of each unit’s group assignment.

555

ACORN outperformed both WaveMAP and classical spike width-based approaches on our dataset, as measured by AMI, suggesting that our method effectively incorporated more relevant information than just the spike width measure into the classification decisions. Notably, we did not fine-tune the parameters for our dataset, but covered a large region of the parameter space; this enabled our method to perform well out-of-the-box on substantially different datasets (Yu et al, 2019, Beau et al, 2025, Lee et al, 2026).

While in this study we applied ACORN to normalized spike waveforms to cluster neuronal cell types, as is likely to be the most common use case in electrophysiology, the method itself is more generally applicable. There is no requirement that the input data be waveforms nor indeed that they derive from neurons. The consensus step (SEED) of the pipeline is universal and can be applied to any set of constituent clusterings to produce a consensus result with concomitant confidence values. Importantly, the number of final clusters is algorithmically derived from the data rather than from user input, underlining the blind and unbiased nature of the method.

A promising next step in building a more advanced classification method would be to incorporate information not just from the spike waveforms of individual contacts, but also from the temporal (Lee et al., 2026; Yu et al., 2025) and spatial characteristics of the units. The former can be effectively investigated by means of “three-dimensional autocorrelograms” (Beau et al., 2025), while the latter through examining multiple contacts across a multielectrode probe. This could provide additional clues not only about the spatial extent of the neuron (Kraskov et al., 2020; Ye et al., 2025), but potentially also about its interactions with the rest of the network. For example, correlations in spike timing patterns could indicate excitatory or inhibitory connections between two units captured simultaneously in the same recording, which, in turn, could help produce a verified ground truth dataset (Ren et al., 2020).

Future neuron classification models should incorporate a wider range of information beyond the traditionally used spike width measures, and their performance be evaluated using solid ground truth datasets. The database we provide can support future research in two main ways. First, it features two distinct neuronal types: PTNs and interneurons, which any classification algorithm should accordingly separate into different clusters – this can be a way to assess the accuracy of classification techniques. Second, it can supplement new datasets obtained from unidentified neurons with a known ground truth dataset: the identity of the cells in our database can be used to label the clusters obtained by unsupervised classification algorithms.

Our unsupervised ensemble clustering technique, ACORN, can be used to label cell types blindly based on waveforms alone without training or parameter tuning. Notably, it provides data-driven confidence values for each cluster assignment it makes, offering researchers a choice between exhaustiveness or a smaller but cleaner subset labeled with high confidence. Its strong out-of-the-box performance on datasets obtained by different labs using a variety of probes in distinct brain areas showcases its robustness, providing an effective tool for electrophysiological cell classification.

References

- Aksenov DP, Miller MJ, Dixon CJ, Wyrwicz AM (2019) The effect of sevoflurane and isoflurane anesthesia on single unit and local field potentials. *Experimental brain research* 237:1521–1529.
- Angotzi GN, Baker AME, Orban G, Vincenzi M, Ribeiro J, Tenorio V, Berdondini L, Baker SN (2025) High-density, Identified Cell Recordings from Motor Cortex of Awake Behaving Macaques using 1024-channel SiNAPS Probes. *bioRxiv* 2025.07.22.665434.

612 Beau M, Herzfeld DJ, Naveros F, Hemelt ME, D'Agostino F, Oostland M, Sánchez-López A,
 613 Chung YY, Maibach M, Kyranakis S (2025) A deep learning strategy to identify cell types
 614 across species from high-density extracellular recordings. *Cell* 188:2218–2234.

615 Botev ZI, Grotowski JF, Kroese DP (2010) Kernel density estimation via diffusion.

616 Dalmia A, Sia S (2021) Clustering with UMAP: Why and how connectivity matters. *arXiv*
 617 preprint arXiv:210805525.

618 Dimidschstein J, Chen Q, Tremblay R, Rogers SL, Saldi G-A, Guo L, Xu Q, Liu R, Lu C, Chu
 619 J (2016) A viral strategy for targeting and manipulating interneurons across vertebrate
 620 species. *Nature Neuroscience* 19:1743–1749.

621 Herzfeld DJ, Hall NJ, Lisberger SG (2025) Strategies to decipher neuron identity from
 622 extracellular recordings in the cerebellum of behaving non-human primates. *bioRxiv*
 623 2025.01.29.634860.

624 Jia Q, Duan Y, Liu Y, Liu J, Luo J, Song Y, Xu Z, Zhang K, Shan J, Mo F (2024) High-
 625 performance bidirectional microelectrode array for assessing sevoflurane anesthesia effects
 626 and in situ electrical stimulation in deep brain regions. *ACS sensors* 9:2877–2887.

627 Klapoetke NC, Murata Y, Kim SS, Pulver SR, Birdsey-Benson A, Cho YK, Morimoto TK,
 628 Chuong AS, Carpenter EJ, Tian Z (2014) Independent optical excitation of distinct neural
 629 populations. *Nature methods* 11:338–346.

630 Kraskov A, Soteropoulos D, Glover I, Lemon R, Baker S (2020) Slowly-conducting pyramidal
 631 tract neurons in macaque and rat. *Cerebral Cortex* 30:3403–3418.

632 Krimer LS, Zaitsev AV, Czanner G, Kroner S, González-Burgos G, Povysheva NV, Iyengar
 633 S, Barrionuevo G, Lewis DA (2005) Cluster analysis–based physiological classification and
 634 morphological properties of inhibitory neurons in layers 2–3 of monkey dorsolateral
 635 prefrontal cortex. *Journal of neurophysiology* 94:3009–3022.

636 Lee EK, Balasubramanian H, Tsoalias A, Anakwe SU, Medalla M, Shenoy KV,
 637 Chandrasekaran C (2021) Non-linear dimensionality reduction on extracellular waveforms
 638 reveals cell type diversity in premotor cortex. *Elife* 10:e67490.

639 Lee EK, Gül AE, Heller G, Lakunina A, Yu H, Shelton A, Olsen S, Steinmetz NA, Hurwitz C,
 640 Jaramillo S, Przytycki PF, Chandrasekaran C (2026) A multimodal approach for visualization
 641 and identification of electrophysiological cell types in vivo. *Nature Communications* 17:5241.

642 Lee K, Carr N, Perliss A, Chandrasekaran C (2023) WaveMAP for identifying putative cell
 643 types from in vivo electrophysiology. *STAR protocols* 4:102320.

644 Lee KH, Chung K, Chung JM, Coggeshall RE (1986) Correlation of cell body size, axon size,
 645 and signal conduction velocity for individually labelled dorsal root ganglion cells in the cat.
 646 *Journal of Comparative Neurology* 243:335–346.

647 McInnes L, Healy J, Astels S (2017) hdbscan: Hierarchical density based clustering. *J Open*
 648 *Source Softw* 2:205.

649 McInnes L, Healy J, Melville J (2018) Umap: Uniform manifold approximation and projection
 650 for dimension reduction. *arXiv preprint arXiv:180203426*.

651 Morales C, Morici JF, Espinosa N, Sacson A, Lara-Vasquez A, García-Pérez M,
 652 Bekinschtein P, Weisstaub NV, Fuentealba P (2021) Dentate gyrus somatostatin cells are
 653 required for contextual discrimination during episodic memory encoding. *Cerebral Cortex*
 654 31:1046–1059.

655 Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with
 656 wavelets and superparamagnetic clustering. *Neural Computation* 16:1661–1687.

657 Ren N, Ito S, Hafizi H, Beggs JM, Stevenson IH (2020) Model-based detection of putative
 658 synaptic connections from spike recordings with latency and type constraints. *Journal of*
 659 *Neurophysiology* 124:1588–1604.

660 Rivara C, Sherwood CC, Bouras C, Hof PR (2003) Stereologic characterization and spatial
 661 distribution patterns of Betz cells in the human primary motor cortex. *The Anatomical Record*

662 Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology: An Official Publication
 663 of the American Association of Anatomists 270:137–151.

664 Schmitz MT, Sandoval K, Chen CP, Mostajo-Radji MA, Seeley WW, Nowakowski TJ, Ye CJ,
 665 Paredes MF, Pollen AA (2022) The development and evolution of inhibitory neurons in
 666 primate cerebrum. *Nature* 603:871–877.

667 Soares D, Goldrick I, Lemon RN, Kraskov A, Greensmith L, Kalmar B (2017) Expression of
 668 Kv3. 1b potassium channel is widespread in macaque motor cortex pyramidal cells: A
 669 histological comparison between rat and macaque. *Journal of Comparative Neurology*
 670 525:2164–2174.

671 Traag VA, Waltman L, Van Eck NJ (2019) From Louvain to Leiden: guaranteeing well-
 672 connected communities. *Scientific Reports* 9:1–12.

673 Trainito C, von Nicolai C, Miller EK, Siegel M (2019) Extracellular spike waveform
 674 dissociates four functionally distinct cell classes in primate cortex. *Current Biology* 29:2973–
 675 2982.

676 Vigneswaran G, Kraskov A, Lemon RN (2011) Large identified pyramidal cells in macaque
 677 motor and premotor cortex exhibit “thin spikes”: implications for cell type classification.
 678 *Journal of Neuroscience* 31:14235–14242.

679 Vinh NX, Epps J, Bailey J (2010) Information Theoretic Measures for Clusterings
 680 Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of*
 681 *Machine Learning Research* 11:2837–2854.

682 Wiesendanger M, Hummelsheim H, Bianchetti M, Chen D, Hyland B, Maier V,
 683 Wiesendanger R (2007) Input and output organization of the supplementary motor area,
 684 Presented at the Ciba Foundation Symposium 132-Motor Areas of the Cerebral Cortex:
 685 Motor Areas of The Cerebral Cortex: Ciba Foundation Symposium 132 pp40–62. Wiley
 686 Online Library.

Ye Z et al. (2025) Ultra-high-density Neuropixels probes improve detection and identification in neuronal recordings. *Neuron* 113(23):3966-3982.e12

Yu H, Lyu H, Xu EY, Windolf C, Lee EK, Yang F, Shelton AM, Olsen S, Minavi S, Winter O, International Brain Laboratory, Dyer EL, Chandrasekaran C, Steinmetz NA, Paninski L, Hurwitz C (2025) In vivo cell-type and brain region classification via multimodal contrastive learning. *bioRxiv* 2024.11.05.622159.

Yu J, Hu H, Agmon A, Svoboda K (2019) Recruitment of GABAergic Interneurons in the Barrel Cortex during Active Tactile Behavior. *Neuron* 104(2):p412-427.e4

Zaitsev AV, Povysheva NV, Gonzalez-Burgos G, Rotaru D, Fish KN, Krimer LS, Lewis DA (2009) Interneuron Diversity in Layers 2–3 of Monkey Prefrontal Cortex. *Cerebral Cortex* 19:1597–1615.

Legends

Figure 1. Analysis outline. **A**, Sweeps are aligned to the timing of antidromic stimulation at $t=0$. **B**, The best channel for each spike is selected based on voltage-differences at base, trough and peak timepoints. These timepoints are automatically refined to the local extrema from user-supplied starting values. **C**, Spike amplitudes that fall below the left tail of the distribution centered around the primary mode get excluded. The probability density function is estimated adaptively using normal kernels (see Methods). **D**, Sweeps with trough timing jitter in excess of 100 μ s relative to the expected are excluded. The expected trough timing is the mean of sweeps whose amplitudes were found acceptable in the previous step.

Figure 2. An illustrative example of constituent clusterings in ACORN. Circled numbers represent neurons. Across clusterings, identical neurons are represented with the same number. Colored shapes enclosing sets of neurons represent clusters. Across clusterings,

there is no direct relation between clusters, hence each cluster is represented using a unique color. This arrangement of neurons and clusterings is used throughout **Tables 2, 3, 4, and 5** to illustrate the method using a simple example.

Figure 3. Ground truths for PTNs. **A**, Dual pulse antidromic stimulation, 10ms apart, evokes characteristic spikes at defined latencies. Traces are channel-averages of sweeps. **B**, Spikes with similar timings can be distinguished due to spatial separation across the contacts. The colors of the traces correspond to the physical location of the recording channel as indicated on the inset picture of the electrode tip (based on an image supplied by Atlas Neuro). In this example, the earlier spike (red) localizes to superficial channels, while the later spike (blue) to near the tip of the probe. Grey circles represent contacts from which no data were obtained. **C**, Spontaneous spikes generated prior to antidromic stimulation can suppress the expected PTN response, as the two action potentials collide along the axon. We provide a dataset with spike waveforms for all identified neurons: 65 PTNs and 22 interneurons (**Figure 3-1**).

Figure 4. Ground truth for interneurons. **A**, Individual spike waveforms (grey traces), and their mean and standard deviation (magenta and cyan traces, respectively). **B,C**, Cells are considered interneurons if they respond to optogenetic stimulation (**B**) elicited through a channelrhodopsin driven by an interneuron-specific promoter (see Methods) with a burst of activity (**C**). We provide a ground truth dataset (**Figure 3-1**).

Figure 5. Spike width correlates with antidromic latency in anesthetized animals in both M1 and SMA. Marker diameters relate linearly to spike amplitudes. Linear regression models yielded R² of 0.45 and 0.65, and p-values of 0.034 and 1.32x10⁻¹³, in M1 and SMA, respectively.

Figure 6. While broad-width PTNs are largely distinguishable from interneurons, narrow-width PTNs and interneurons are not readily separable using width-based approaches. **A**, Spike widths of PTNs and interneurons overlap substantially, especially around 300 μ s. While on average PTNs have longer action potential durations than interneurons, both short-width PTNs and long-width interneurons exist. **B**, The WaveMAP unsupervised approach produced two clusters: one enriched for PTNs (cluster 1) and another that is a roughly equal mixture of the two neuron classes (cluster 0). **C,D**, Linear regressions indicate that both UMAP dimensions correlate significantly with spike width (F-tests against constant models: $p=5.49 \times 10^{-21}$ and $p=6.13 \times 10^{-25}$ for UMAP1 and UMAP2, respectively).

Figure 7. Clustering performance of ACORN evaluated across three datasets. Clustering results for all constituent clusterings (1125 parameter combinations) are represented in terms of classification performance (AMI) evaluated against the ground truth and the proportion of neurons labelled. The results are shown as a 2D normalized histogram in grey scale. Darker shades represent a higher proportion of constituent clusterings falling within the corresponding ranges along the two axes. Colored crosses with numbers ranging from 1.0 to 4.0 denote consensus clustering results at the specified confidence thresholds. High confidence thresholds produced high AMI at the cost of exhaustiveness, while low confidence thresholds labelled a large proportion of neurons at the cost of AMI. **A**, ACORN results on our macaque ground truth dataset. **B**, ACORN results on the Yu et al dataset. **C**, ACORN results on the C4 dataset. **D**, The performance metric (AMI x proportion labelled) plotted as a function of the confidence threshold. As the confidence threshold increases, the proportion of labelled neurons decreases, yet the performance metric remains largely stable within the parameter range of 1.5-3.0 in the three datasets. **E**, Evaluation of the data-driven confidence metric on the C4 dataset. Including increasingly lower-confidence neuronal assignments in the results produces, on average, worse clustering classification, as measured by AMI. Plotted are the confidence metric of the assignment of each particular neuron (blue trace), and the AMI result of considering all neurons with at least that much

confidence in their clustering assignments (brown trace). In this dataset, the 17 highest-confidence assignments were all Purkinje cell complex spikes correctly placed in the same cluster by our algorithm. AMI values are defined where there are at least two distinct neuronal cluster labels to evaluate against the ground truth ($N \geq 18$). Further analysis of the parameter-space is provided in **Figure 7-1**.

Table 1. The region of parameter space used to generate constituent clusterings. Each combination of parameters is used to generate a single clustering; their consensus produces the final result. N denotes the number of neurons in the dataset. $n_neighbors$ sets the size of the local neighborhood when learning the manifold structure (a trade-off between interpreting local vs global data structure). $n_components$ sets the number of dimensions the waveform data gets embedded into (a trade-off between the density and the amount of information). min_dist sets the minimum distance between neurons in the low-dimensional embedding (a trade-off between representing local vs global topological structure). $min_cluster_size$ sets the minimum number of neurons that must be present in a group for it to be considered a cluster. $min_samples$ sets the minimum number of neighbors for a neuron to be considered core (a trade-off between exhaustiveness and rigor). We set $min_samples$ as a fraction of the $min_cluster_size$ parameter as shown in the table. Rounding is done using the "half-to-even" tie-breaking rule (IEEE 754), which is the default mode in Python.

Table 2. The first exemplar selection iteration on the example dataset. Numbers correspond to terms of the S-metric for each neuron as the SEED algorithm identifies exemplars. The positive term is the number of neurons assigned to the same constituent cluster. Neuron 2, achieving the highest S-value, is chosen as the first exemplar.

Table 3. The second exemplar selection iteration on the example dataset. Numbers correspond to terms of the S-metric for each neuron as the SEED algorithm identifies exemplars. The negative term applies instead of the positive term when sharing a constituent cluster with an already-assigned exemplar. The negative term is $-N/K/2 = -6/3/2 = -1$ in this example, where N is the number of neurons and K is the maximum number of clusters in constituent clusterings. Neuron 5, achieving the highest S-value, is chosen as the second exemplar.

Table 4. The third exemplar selection iteration on the example dataset. Numbers correspond to terms of the S-metric for each neuron as the SEED algorithm identifies exemplars. The negative term is scaled according to the number of already-assigned exemplars in a neuron's constituent cluster. Thus, in constituent clustering 3, where two exemplars are in the same constituent cluster, the negative term is -2. With no neuron achieving a positive S-value in this iteration, the exemplar selection stage ends.

Table 5. Consensus cluster assignments on the example dataset. Each neuron, once per constituent clustering, votes for the exemplar with which it shares its cluster. If a neuron shares its constituent cluster with multiple exemplars, its vote is split equally between them. If a neuron's constituent cluster is "unassigned" (-1), it votes for "uncertain" with half strength. If a neuron neither shares its constituent cluster with any exemplars, nor is it unassigned in that constituent cluster, then it abstains. Finally, each neuron whose confidence is above the threshold is assigned to the consensus cluster built around the exemplar that won its votes. If "uncertain" won a neuron's votes, or if the neuron's confidence value is less than the threshold, then the consensus result is "uncertain" for that neuron.

Table 6. Confusion matrix of the WaveMAP method applied to our units. In each row (cluster), the highest value is underlined. In each column (ground truth class), the highest value is highlighted in bold if it is underlined.

Table 7. Confusion matrix of ACORN applied to our units, using the default confidence threshold of 2.0. Cluster predictions are blind; their textual descriptions (in quotation marks within parentheses) are just for convenience to aid interpretation. In each row (cluster), the highest value is underlined. In each column (ground truth class), the highest value is highlighted in bold if it is underlined.

Table 8. Confusion matrix of ACORN applied to the Yu et al dataset, using the default confidence threshold of 2.0. Cluster predictions are blind; their textual descriptions (in quotation marks within parentheses) are just for convenience to aid interpretation. In each row (cluster), the highest value is underlined. In each column (ground truth class), the highest value is highlighted in bold if it is underlined. Here we treated each cell type as its own class, regardless of its layer, given that layer information was not used in our clusterings. For a closer comparison with Lee et al, see **Tables 8-1** and **8-2**, where, still without layer information, ACORN is set to cluster cells types and layers both.

Table 9. Confusion matrix of ACORN applied to the C4 dataset, using the default confidence threshold of 2.0. Cluster predictions are blind; their textual descriptions (in quotation marks within parentheses) are just for convenience to aid interpretation. In each row (cluster), the highest value is underlined. In each column (ground truth class), the highest value is highlighted in bold if it is underlined.

eNeuro Accepted Manuscript

Parameter	Sampled parameter values
UMAP n_neighbors	round(N/3, N/4, N/6)
UMAP n_components	4, 8, 16, 32, 64
UMAP min_dist	0.3, 0.4, 0.5, 0.6, 0.7
HDBSCAN min_cluster	12, 16, 20
HDBSCAN min_samples	round(min_cluster / [1.5, 2, 2.5, 3, 4])

847

848

Iteration 1	Neuron 1	Neuron 2	Neuron 3	Neuron 4	Neuron 5	Neuron 6
Clustering 1	3	3	3	0	2	2
Clustering 2	2	2	2	2	2	2
Clustering 3	2	4	2	4	4	4
Clustering 4	3	3	0	3	2	2
Clustering 5	0	2	2	2	2	0
S-Metric (Sum)	10	14	9	11	12	10

849

850

Iteration 2	Neuron 1	Neuron 2 (A)	Neuron 3	Neuron 4	Neuron 5	Neuron 6
Clustering 1	-1	-1	-1	0	2	2
Clustering 2	-1	-1	2	2	2	2
Clustering 3	2	-1	2	-1	-1	-1
Clustering 4	-1	-1	0	-1	2	2
Clustering 5	0	-1	-1	2	2	0
S-Metric (Sum)	-1	-5	2	2	7	5

851

852

Iteration 3	Neuron 1	Neuron 2 (A)	Neuron 3	Neuron 4	Neuron 5 (B)	Neuron 6
Clustering 1	-1	-1	-1	0	-1	-1
Clustering 2	-1	-1	-1	2	-1	2
Clustering 3	2	-2	2	-2	-2	-2
Clustering 4	-1	-1	0	-1	-1	-1
Clustering 5	0	-1	-1	-1	-1	0
S-Metric (Sum)	-1	-6	-1	-2	-6	-2

853

854

Voting	Neuron 1	Neuron 2 (A)	Neuron 3	Neuron 4	Neuron 5 (B)	Neuron 6
Clustering 1	A	A	A	Uncertain	B	B
Clustering 2	A	A	B	Abstain	B	Abstain
Clustering 3	Abstain	0.5 for A 0.5 for B	Abstain	0.5 for A 0.5 for B	0.5 for A 0.5 for B	0.5 for A 0.5 for B
Clustering 4	A	A	Uncertain	A	B	B
Clustering 5	Uncertain	A	A	B	B	Uncertain
Exemplar A (Sum)	3	4.5	2	1.5	0	0.5
Exemplar B (Sum)	0	0.5	1	1.5	4.5	2.5
Uncertain (Sum/2)	0.5	0	0.5	0.5	0	0.5
Winner	A	A	A	Tie	B	B
Confidence	6.00	9.00	2.00	1.00	9.00	5.00
Consensus	Cluster A	Cluster A	Cluster A	Uncertain	Cluster B	Cluster B

855

856

WaveMAP , our dataset (AMI: 0.14)	Actual Interneuron	Actual PTN
Cluster 1 ("Interneuron")	19	<u>24</u>
Cluster 2 ("PTN")	3	<u>41</u>

857

858

ACORN , our dataset (AMI: 0.26)	Actual Interneuron	Actual PTN
Cluster 1 ("Interneuron")	<u>19</u>	14
Cluster 2 ("PTN")	2	<u>37</u>
Uncertain	1	14

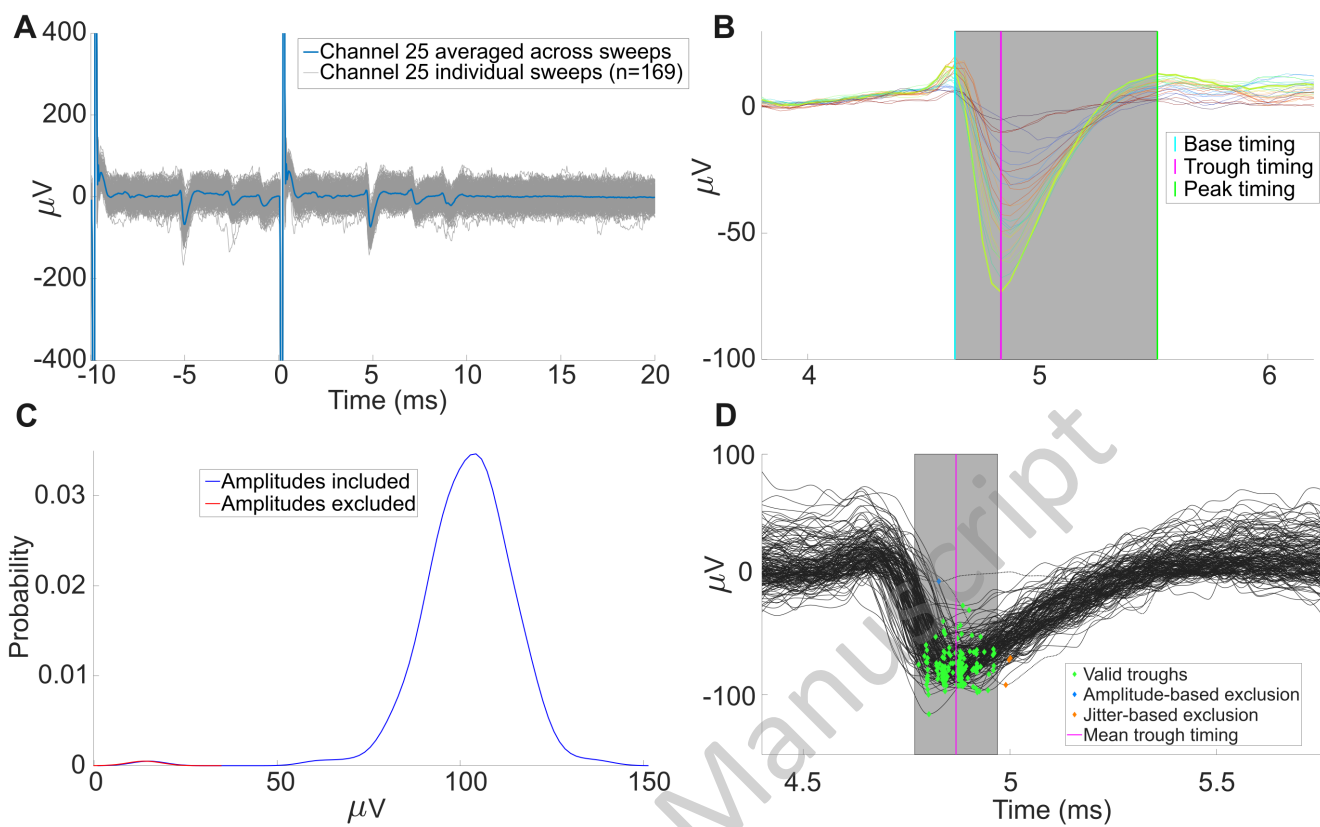
859

860

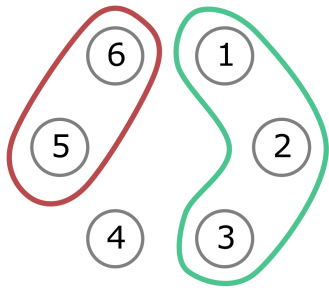
861

ACORN , Yu et al dataset (AMI: 0.52)	Actual Excitatory	Actual PV ⁺	Actual SOM ⁺	Actual VIP ⁺
Cluster 1 ("Excitatory")	<u>79</u>	0	7	4
Cluster 2 ("PV ⁺ ")	0	<u>34</u>	0	0
Cluster 3 ("SOM ⁺ ")	6	21	<u>42</u>	1
Uncertain	22	6	20	4

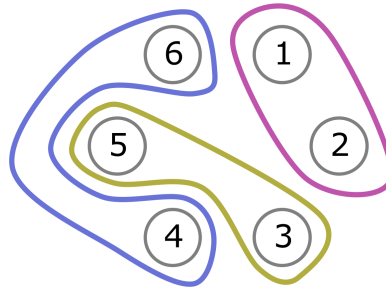
ACORN, C4 dataset (AMI: 0.66)	Actual GoC	Actual MFB	Actual MLI	Actual PkCcs	Actual PkCss
Cluster 1 ("MFB")	0	<u>25</u>	1	0	6
Cluster 2 ("PkCcs")	0	0	0	<u>20</u>	0
Cluster 3 ("PkCss")	0	0	4	1	<u>38</u>
Uncertain	18	5	23	38	27



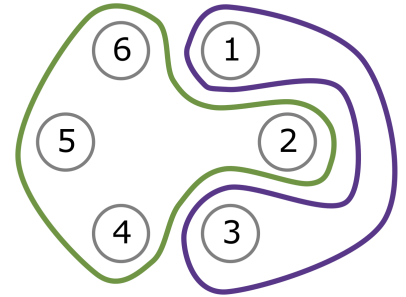
Clustering 1



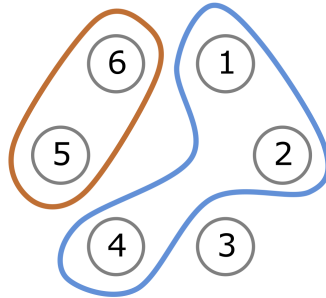
Clustering 2



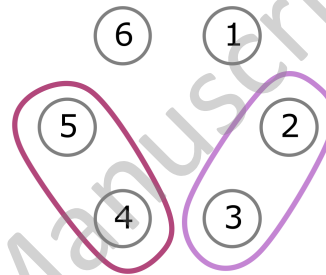
Clustering 3



Clustering 4



Clustering 5



eNeuro Accepted Manuscript

