J Neurophysiol 123: 2217–2234, 2020.
First published May 6, 2020; doi:10.1152/jn.00754.2019.

INNOVATIVE METHODOLOGY

Using deep neural networks to detect complex spikes of cerebellar Purkinje cells

Akshay Markanday,[1,2,3]* Joachim Bellet,[1,2,3]* Marie E. Bellet,[3] Junya Inoue,[1] Ziad M. Hafed,[1,3]* and Peter Thier[1,3]*

[1]*Hertie Institute for Clinical Brain Research, Tübingen, Germany;* [2]*Graduate School of Neural and Behavioral Sciences, International Max Planck Research School, Tübingen University, Tübingen, Germany; and* [3]*Werner Reichardt Centre for Integrative Neuroscience (CIN), Tübingen, Germany*

Submitted 16 December 2019; accepted in final form 4 May 2020

Markanday A, Bellet J, Bellet ME, Inoue J, Hafed ZM, Thier P. Using deep neural networks to detect complex spikes of cerebellar Purkinje cells. *J Neurophysiol* 123: 2217–2234, 2020. First published May 6, 2020; doi:10.1152/jn.00754.2019.—One of the most powerful excitatory synapses in the brain is formed by cerebellar climbing fibers, originating from neurons in the inferior olive, that wrap around the proximal dendrites of cerebellar Purkinje cells. The activation of a single olivary neuron is capable of generating a large electrical event, called "complex spike," at the level of the postsynaptic Purkinje cell, comprising of an initial large-amplitude spike followed by a long polyphasic tail of small-amplitude spikelets. Several ideas discussing the role of the cerebellum in motor control are centered on these complex spike events. However, these events, only occurring one to two times per second, are extremely rare relative to Purkinje cell "simple spikes" (standard sodium-potassium action potentials). As a result, drawing conclusions about their functional role has been very challenging. In fact, because standard spike sorting approaches cannot fully handle the polyphasic shape of complex spike waveforms, the only safe way to avoid omissions and false detections has been to rely on visual inspection by experts, which is both tedious and, because of attentional fluctuations, error prone. Here we present a deep learning algorithm for rapidly and reliably detecting complex spikes. Our algorithm, utilizing both action potential and local field potential signals, not only detects complex spikes much faster than human experts, but it also reliably provides complex spike duration measures similar to those of the experts. A quantitative comparison of our algorithm's performance to both classic and novel published approaches addressing the same problem reveals that it clearly outperforms these approaches.

NEW & NOTEWORTHY Purkinje cell "complex spikes", fired at perplexingly low rates, play a crucial role in cerebellum-based motor learning. Careful interpretations of these spikes require manually detecting them, since conventional online or offline spike sorting algorithms are optimized for classifying much simpler waveform morphologies. We present a novel deep learning approach for identifying complex spikes, which also measures additional relevant neurophysiological features, with an accuracy level matching that of human experts yet with very little time expenditure.

action potentials; cerebellum; complex spikes; convolutional neural networks; local field potentials; simple spikes

INTRODUCTION

The Purkinje cell (PC) output, the sole output of the cerebellar cortex, is driven by two distinct types of responses (Fig. 1*A*), the simple spike (SS) and the complex spike (CS) (Eccles et al. 1966; Thach 1967, 1968). SSs are ordinary sodium-potassium spikes with a simple bi- or triphasic shape in extracellular recordings (Fig. 1*B*). These spikes, lasting only a fraction of a millisecond and firing up to several hundred times per second, reflect the concerted impact of mossy fiber input, mediated via the granule cell-parallel fiber system, as well as inhibitory interneurons. On the other hand, an individual CS (Fig. 1*C*), elicited by a single climbing fiber originating from the inferior olivary nucleus and pervading the proximal dendrites of a PC, is characterized by a polyphasic somatic spike consisting of a first back propagated axonal spike component followed by a series of spikelets riding on a long-lasting, calcium-dependent depolarization (Davie et al. 2008; Eccles and Szentágothai 1967; Fujita 1968; Llinás and Sugimori 1980; Stuart and Häusser 1994; Thach 1968). In addition to an exceptional morphology, CSs also exhibit a perplexingly low firing rate of at most two spikes per second (Fig. 1*A*), which is much lower than the rate of SSs recorded from the same cells. What could these infrequent, yet unique, events possibly tell us about their purpose, and what might be the best statistical tool allowing us to unravel the full extent of information carried by them? These are questions that have kept researchers busy until today.

CSs have originally been hypothesized to play a crucial role in either motor timing (Leznik and Llinás 2005; Llinás 1974) or performance-error based motor learning (Albus 1971; Ito 1972; Marr 1969). While many follow-up experiments seemed to support the latter idea (Herzfeld et al. 2015, 2018; Kitazawa et al. 1998; Medina and Lisberger 2008; Oscarsson 1980), not all findings have been fully compatible with this so-called Marr-Albus-Ito hypothesis, at least not in its original form (Catz et al. 2005; Dash et al. 2010; Junker et al. 2018; Kostadinov et al. 2019; Ohmae and Medina 2015; Streng et al. 2017). As a result, reaching consensus on the diverse views of CS functions would be substantially facilitated by more data on these sparse neural events, collected in conjunction with advanced behavioral paradigms. Yet, it is exactly their unique properties of rarity combined with complex and highly idio-

* A. Markanday and J. Bellet contributed equally to this work. Z. M. Hafed and P. Thier contributed equally to this work.
Correspondence: P. Thier (thier@uni-tuebingen.de).

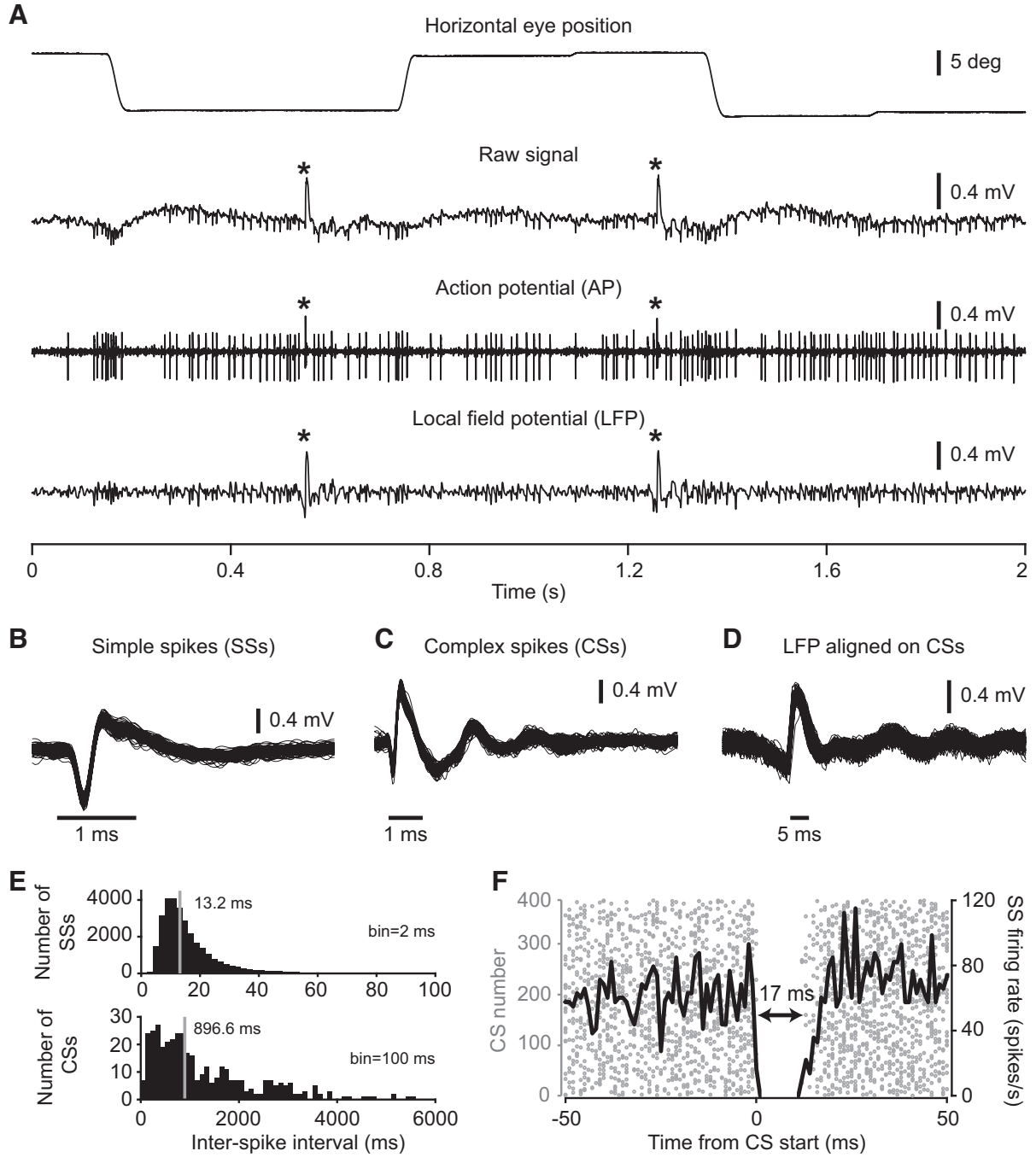


Fig. 1. Characteristics of an exemplary Purkinje cell. *A*; raw signal (wide-band, *2nd row*), action potential (AP; high band-passed, 300 Hz to 3 KHz, *3rd row*), and local field potential (LFP; low passed, 30–400 Hz, *4th row*) activity in relation to horizontal eye movements (*1st row*). *Complex spikes (CSs). *B*: a subset of isolated simple spike (SS) waveforms aligned on SS start. *C*: a subset of isolated CS waveforms aligned on CS start. *D*: a subset of LFP responses aligned to CS start. *E*: histogram of interspike intervals of SSs (*top*) and CSs (*bottom*). Solid gray line depicts the median value. *F*: raster plot showing a 17-ms pause in SS activity caused by the occurrence of a CS. Solid black line represents the mean SS firing rate aligned to CS start.

syncratic spike morphology that have hampered progress. In fact, CS morphology not only differs between individual PCs, but it also often changes over the course of a single recording session from the same PC. This is why standard spike sorting software, which may work well in detecting the much "simpler" and more frequent SSs, turns out to be highly error prone when adapted for detecting CSs. Critically, given the rarity of CSs in relation to their SS counterparts recorded from the same cells, even a few missing or erroneously detected CS events will have a profound impact on conclusions drawn about CS functional roles. Consequently, most researchers either completely shy away from dealing with this problem at all, or they do so by resorting to meticulously labeling CSs manually, possibly prepared by prior coarse predetection by conventional spike-sorting approaches. The manual approach is highly valuable, but it is also exhausting and constrains the amount of experimental data that can be processed. The net result is that in both cases (i.e., either not

addressing CS functions at all, or meticulously labeling CSs using heavy manual loads), the general pace of development in the field is compromised.

In this paper, we present a study of the performance of a convolutional neural network (CNN) developed to dramatically reduce the burden of investigators in identifying CSs. We show that our algorithm is able to learn fast and that it easily matches the performance of an experienced human expert in detecting CSs. We also demonstrate that our approach outperforms an algorithm based on principal component analysis (PCA), which was recently suggested to detect CSs (Zur and Joshua 2019), as well as a commonly used online sorting solution that researchers in the field typically use. Finally, we additionally show that our algorithm is the first, to our knowledge, to provide an accurate estimate of CS duration, a parameter that is supposed to contain critical information for motor learning (Yang and Lisberger 2014).

MATERIALS AND METHODS

Animals, Preparation, Surgical Procedures, and Recording Methods

Two adult male rhesus macaques (*Macaca mulatta*) of age 10 (*monkey K*) and 8 (*monkey E*) yr, purchased from the German Primate Center, Göttingen, were subjects in this study. Initial training of all animals required them to voluntarily enter an individually customized primate chair and get accustomed to the setup environment, a procedure that could last for up to 3 mo. Following initial training, they underwent the first major surgical procedure in which foundations of all implants were fixed to the skull using titanium bone screws and then were allowed to rest for a period of ~3–4 mo to improve the long-term stability of the implant foundations. Then, a titanium-based hexagonal tube-shaped head post was attached to the implanted head holder base to painlessly immobilize the head during experiments, and scleral search coils were implanted to record eye positions using electromagnetic induction (Bechert and Koenig 1996; Judge et al. 1980). Within 2–3 wk of recovery from the eye-coil implantation procedure, the monkeys quickly recapitulated the already learned chair-training protocol and were trained further on their respective behavioral paradigms. Once fully trained, a cylindrical titanium recording chamber, whose position and orientation were carefully planned and confirmed based on pre- and postsurgical MRIs, was finally mounted on the implanted chamber base, tilting backward by an angle of 30° with respect to the frontal plane, right above the midline of the cerebellum. A part of the skull within the chamber was removed to allow precise electrode access to our region of interest, the oculomotor vermis (OMV, lobuli VIc/VIIa), for electrophysiological recordings. All surgical procedures were carried out under aseptic conditions using general anesthesia, and postsurgical analgesics were delivered until full recovery. See Prsa et al. (2010) for full details.

All experiments and surgical procedures were approved by the local animal care authority (Regierungspräsidium Tübingen) and complied with German and European law as well as the National Institutes of Health's *Guide for the Care and Use of Laboratory Animals*. All procedures were carefully monitored by the veterinary service of Tübingen University.

Behavioral Tasks

We collected data from two monkeys generating visually guided saccades, which are known to be associated with CS occurrence (Fig. 1*A*). Each trial started with a red fixation dot (diameter: 0.2°) displayed at the center of a CRT monitor placed 38 cm in front of the monkey. After a short and variable fixation period (400–600 ms from trial onset), the fixation dot disappeared and at the same time, a target, having the same features as the fixation dot, appeared on the horizontal axis at an eccentricity of 15°. In a given session, the target was presented consistently either on the left or right of the central fixation dot. The maximum number of trials (>200) per session depended on the willingness of the monkey to cooperate and on the duration for which a PC could be kept well isolated. Each trial lasted for 1,200 ms. At the end of every correct trial, the monkeys were rewarded with a drop of water.

Electrophysiological Recordings

We recorded extracellular activity with commercially available glass-coated tungsten microelectrodes (impedance: 1–2 MΩ; Alpha Omega Engineering, Nazareth, Israel). Electrode position was controlled using a modular multielectrode manipulator (Electrode Positioning System and Multi-Channel Processor, Alpha Omega Engineering). We targeted the OMV based on the implanted position and orientation of the recording chamber that we used, and we also identified the OMV region based on the characteristic saccade-related modulation of an intense background activity, reflecting multiunit granule cell activity. The wide-band raw signal picked up by our electrode was also clearly modulated by saccadic eye movements (Fig. 1*A*, *2nd row*). The raw signal, sampled at 25 KHz, was band-pass filtered online between 30 Hz and 3 KHz to enable online spike sorting of SSs and CSs based on spike waveform shapes.

Multi Spike Detector: the Online Spike Sorting Algorithm

Single PC units were identified, online, by the presence of a high-frequency SS discharge accompanied by the signatory, low-frequency CS discharge. We used a real-time spike sorter, the Alpha Omega Engineering Multi Spike Detector (MSD), for online unit detection. The MSD, designed for detecting sharp waveforms, uses a template matching algorithm developed by Wörgötter et al. (1986), sorting waveforms according to their shape. The algorithm employs a continuous comparison of the electrode signal against an eight-point template defined by the experimenter to approximate the shape of the spike of interest. The sum of squares of the difference between the template and electrode signal is used as a statistical criterion for the goodness of fit. Whenever the goodness of fit exceeds a threshold, the detection of a spike is reported. The eight-point template can be adjusted manually or, alternatively, run in an adaptive mode that allows it to keep track of waveforms that may gradually change over time.

Online Identification of Simple Spikes and Complex Spikes in Purkinje Cells

As opposed to short duration SSs (Fig. 1*B*), characterized by short interspike intervals (Fig. 1*E*, *top*), the long duration CSs (Fig. 1*C*) are much rarer (Fig. 1*E*, *bottom*, note the different *x*-axis range from the *top*). In addition to the 10- to 20-ms-long pause in SS firing rate following the occurrence of a CS (e.g., Fig. 1*F*, Eccles et al. 1966; Bell and Grimm 1969; Latham and Paul 1971; McDevitt et al. 1982; Thach 1967), the presence of a CS is also indicated by a massive deflection of the local field potential (LFP) signal (30–400 Hz, constructed using a second order Butterworth filter with a sampling frequency of 25 KHz) lasting for the whole duration of the CS (Fig. 1*D*).

Complex Spike Detection Using a Convolutional Neural Network

Overview of our algorithm. Inspired by the architecture of a convolutional neural network (CNN) that was originally designed to segment images ("U-Net"; Ronneberger et al. 2015), our network was initially developed to detect eye movement events in one-dimensional

eye position signals ("U'n'Eye"; Bellet et al. 2019; for the network architecture, see Fig. 2 in that paper, as well as the source code on https://github.com/berenslab/uneye). In the present work, we extended the horizon of our state-of-the-art eye movement algorithm toward the detection of more complex electrophysiological events, such as CSs. The main idea of our approach is to train a classifier to extract relevant features from electrophysiological recordings of PCs and to use these features for identifying CSs. Therefore, we repurposed our existing algorithm to use the LFPs within a frequency band of 30 Hz to 400 Hz (Fig. 1*A*, *4th row*) and high-pass filtered action potential signals (30 Hz to 3 KHz; Fig. 1*A*, *3rd row*), sampled at the same frequency of 25 KHz, as inputs to the network (Fig. 2*A*, *top*). We chose these two inputs because human experts achieve consensus on the presence or absence of a CS, more easily and reliably, if both action potentials and LFPs are simultaneously available. Raw unfiltered signals (potentially containing enough information to detect CSs) are often not analyzed directly, because of a general interest in identifying spiking activity; hence, we focused on using the most commonly used filtered electrophysiological signals in the field. This approach is also consistent with that used by a recent algorithm for CS detection, which we benchmark our algorithm against in the present study (Zur and Joshua 2019). Note that since the fast Fourier transform (FFT) of a CS signal suggests that most of the low-frequency power in CSs lies around ~700 Hz, a wider LFP band than what we used here would have potentially been more optimal. However, the choice of LFP band should not strongly affect the performance of our algorithm with proper training data, and, more importantly, our choice of the LFP band (30–400 Hz) followed Zur and Joshua's recommendation, which was important to allow a fair comparison between the performance of our algorithm and that of these authors.

Because factors such as electrode impedance and distance of the electrode relative to the cell body may potentially result in amplitude

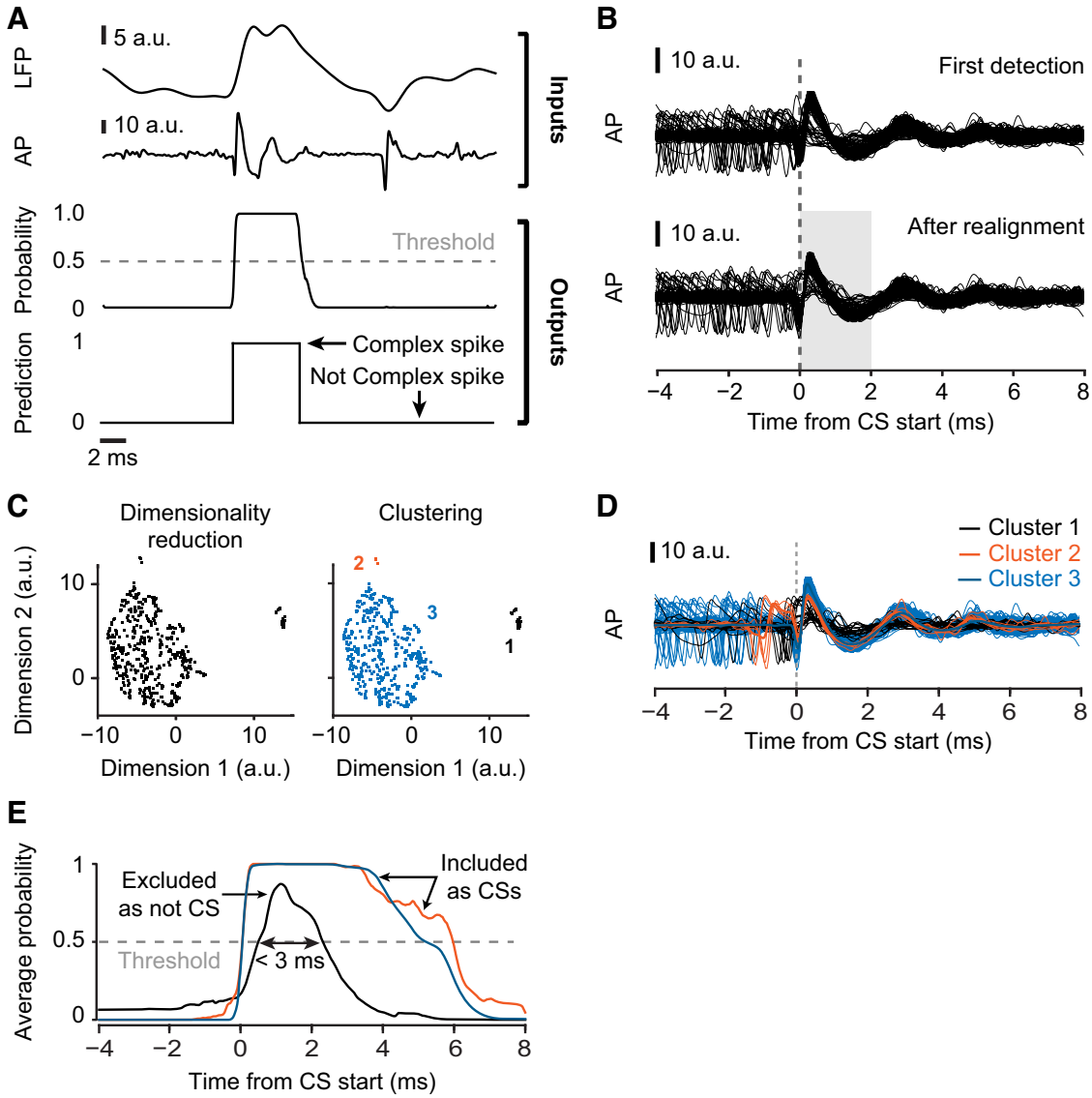


Fig. 2. Pipeline for complex spike (CS) detection. *A*: input to the network [normalized local field potential and action potential signal, labeled as LFP and AP, respectively.] as well as its output (bin-wise predictive probability CSs occurrence and binary CS classification). Dashed gray line signifies the 0.5 probability threshold value; a.u., artificial units. *B*: waveforms aligned to the first detection of start times of all CSs detected by the network (*top*) used for computing an average waveform that served as a template for realigning the waveforms of all detected CS events (*bottom*). *C*: projection of the waveforms during the 2-ms time interval (gray shaded region in *B*, *bottom*) onto a 2-dimensional plane and identification of clusters in this space. Different colors indicate distinct clusters. *D*: waveforms of the clusters in *C*. Note how *cluster 1* clearly violates well-known CS waveform shapes. *E*: average predictive probability output of the network for the events in each cluster. Clusters, whose probability output exceeds the classification threshold of 0.5 (dashed gray line) for <3 ms, are excluded as not representing CSs (*cluster 1*).

scaling of the recorded signals across different recording sessions, we first normalized the LFP and action potential signals by dividing them by the median of the nonnegative signal components. These normalized signals were input to the CNN (Fig. 2*A*, *top*) and combined by the first convolutional layer. The network classified the inputs at every given time point as being either "CS" or "Not CS" (Fig. 2*A*, *bottom*). The classification was achieved by applying a threshold to the result of the final output neuron of the CNN, which may be thought of as providing a predictive probability of the presence of a CS (Fig. 2*A*, *middle*). We picked a threshold of 0.5 on this predictive probability. The prediction for each time bin depended on an interval in the input signal whose size was determined by the size of the max-pooling and convolutional kernels of the CNN. The output was a bin-wise predictive probability of CS occurrence that further underwent a subsequent postprocessing step (described below) to exclude potential false alarms. The details of the network and the steps involved in post processing are explained in more detail in the following sections.

Convolutional neural network details. Our network uses convolutional and max-pooling operations to extract temporal features relevant for distinguishing CSs from the surrounding signal. Max-pooling is an operation that down-samples the input signal to reduce the dimensionality of its representation in the network. It filters the input with a certain window size and extracts only the maximum value. It then steps further on the input, repeating the same operation on the next time window. Convolutional layers extract relevant features of the input signal by learning the parameters of its convolutional kernel during training. The total number of parameters used in the model was 31,482. We chose the size of the max-pooling (mp) and convolutional kernels (c) as seven and nine bins, respectively. These influence the signal interval (SI) taken into account for labeling one time bin in the output, as described by the formula

$$\mathrm{SI} = \frac{(c+1)mp^2 + (c-1)mp + 2(c-2)}{2}$$

The formula for this SI was determined analytically by applying the kernels of the network layers in a chain. In our case, the SI corresponds to 281 time bins centered around each classified bin containing a predicted CS event. As our sampling rate was 25 kHz, a CS of 10-ms duration would span 250 time bins. This means that the network was often using information surrounding CS events (281 vs. 250 time bins) to classify CSs. In applications with different sampling rates than ours, the choice of SI can be adjusted to match our strategy of using information surrounding individual CS events.

Training and testing procedures. To prepare the training set, we asked a human expert, experienced in the visual classification of PC spikes, to identify CS events and manually label their start and end points. The expert did this for a total of 159 recorded PCs. For this, the expert used small segments of action potential and LFP recordings, without having access to eye movement data. For each PC, 24 segments, each 250-ms long, were manually labeled. Due to the probabilistic nature of CS occurrence, a recording segment could either contain a CS or not. This resulted in 250-ms-long "labeled input" segments with binary values; 1, between the start and end points of the manually labeled CSs, and 0 elsewhere. For our training set, we only considered those recording segments that contained at least one CS.

For every PC tested for CS detection, we trained a separate network excluding the currently tested PC from the training set. This "leave-one-out" approach allowed us to test how well the network generalized to new data sets, on which it had not been trained, and it also allowed us to have multiple performance tests on our algorithm. Therefore, the training set always comprised the remaining 158 PCs not being currently tested. Depending on which PC was excluded from the set, due to the "leave-one-out" approach, our training set consisted of 2,160–2,192 recording segments that corresponded to a total duration of 540–548 s. Other parameters of network training, such as loss function, learning rate, batch size, and early stopping criterion, were chosen as described in Bellet et al. (2019) for U'n'Eye.

We also performed one more performance test of our and other algorithms, which was concerned with establishing consistency with expert labeling. For seven PCs (out of 159), we asked our human expert to manually label CSs in the entire records and not just a small training subset within each of them. This allowed us to directly compare the labeling of the entire records of these seven PCs by all algorithms that we considered in this study and the human expert. The choice of these seven PCs was based on how well isolated the units were, which allowed the other algorithms to perform at their best capacities in detecting CSs, thus posing a tough competition to our algorithm. Our algorithm in this case was based on training the network on segments from the remaining 158 PCs (other than the currently tested one), as described above.

Postprocessing. We implemented three automated postprocessing steps to enhance the quality of CS detection by our algorithm, for example, to minimize false alarms. First, time shifts between the detected start points of all CSs fired by a particular PC were corrected by realigning them. To this end, we computed the average waveform from the first detection of start times of all detected CSs (Fig. 2*B*, *top*). This average-waveform template was then used as a reference to realign each waveform within a ±2-ms window around CS start so that the cross correlation was maximized (Fig. 2*B*, *bottom*). Second, action potential and LFP waveforms, occurring within 2 ms after CS start, were projected onto a two-dimensional plane (Fig. 2*C*, *left*) using the Uniform manifold approximation and projection (UMAP) dimensionality reduction technique (McInnes et al. 2018). The waveform clusters after dimensionality reduction represented potential candidates for CSs of the recorded PC. Some of these candidates needed to be excluded. For example, if the network in the first step mistakenly classified non-CS events as CSs, then the clustering method would help to refine the classification. This was achieved by using a third postprocessing step (Fig. 2*C*, *right*) to cluster waveforms into suitable CSs and unsuitable ones. For example, among the CS events erroneously detected by the network, there might be SSs that are revealed by a separate cluster in the two-dimensional space (Fig. 2, *C*, *right*, and *D*, black vs. orange and blue). In this third step, groups of waveforms were identified (Fig. 2*D*) using HDBSCAN, a hierarchical clustering algorithm (Campello et al. 2013) that builds a tree to describe the distance between data points. The algorithm minimizes the spanning size of the tree and further reduces the complexity of the tree to end up with a minimum number of leaf nodes, corresponding to the clusters. We used the default parameters for HDBSCAN with the option to find only one cluster. Waveforms were excluded if they belonged to a cluster for which the average predictive probability output from the network remained above 0.5 for less than 3 ms, which was deemed too brief to be an appropriate duration of CS waveform (Fig. 2*E*). Not only non-CS events might have contributed to a distinct cluster separated from the main CS cluster, but true CSs with slightly deviant waveforms (Fig. 2*D*, orange vs. blue) might also have led to separate clusters in the two-dimensional space (Fig. 2*C*, orange vs. blue). For all CS clusters that met the defined threshold criterion on predictive probability (Fig. 2*E*, *cluster 1* and *2*), CS timing and corresponding cluster IDs allowed the user to carefully inspect each cluster and decide whether to include clusters with deviant, yet true, CSs or not.

Optional postverification. After receiving the final output from our algorithm, comprising of CS start and end times, embedding dimensions, as well as cluster IDs, users can optionally add a postverification step to track secondary effects that they may be interested in investigating (e.g., gradual drift in cell position relative to the electrode). To do this in our own analyses, we visually verified the authenticity of the detected CSs. For this, we relied on the shape of the averaged CS waveforms belonging to each cluster as well as the pause induced by the same cluster (putative PC) in SS firing. Similar CSs, albeit grouped under a separate cluster (possibly due to a modified

shape of their waveform), induced pauses of similar duration in SS firing. On the other hand, false positives were excluded based on their inability to match the above criteria. Although not embedded in our automated algorithm, this manual "postverification" step can provide added confidence regarding the performance of our algorithm, and at a minimal cost of time investment. Indeed, all results in this paper (except those in Fig. 10) describe the performance of our algorithm without any manual postverification.

Complex Spike Detection Using Zur and Joshua's PCA-Based Algorithm

To compare the performance of our algorithm to the recently developed PCA-based algorithm (Zur and Joshua 2019), we used the graphical user interface (GUI) provided by the authors at https://github.com/MatiJlab/ComplexSpikeDetection. The code was adjusted to open our data (i.e., .mat files), but we otherwise followed all instructions provided by the authors. To have a fair comparison with our fully automated algorithm, we excluded (from our algorithm) the last "postverification" step based on visual selection of each putative CS.

Quality Metrics

We evaluated the performance of all three algorithms (ours, the PCA-based approach, and the MSD) in detecting CSs using the so-called F1 score (Dice 1945; Sorensen 1948), which compares the consistency of CS labels predicted by the algorithm with the "ground-truth" labels provided by the human expert. The F1 score is the harmonic mean of recall (the ratio of true positive detections and all true CS labels) and precision (the ratio of true positive detections and all CS labels predicted by the algorithm), as given by the following equation:

$$\mathrm{F1} = \frac{2 \times \mathrm{recall} \times \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}}$$

In our case, an F1 score of 1 would suggest that the CSs predicted by our algorithm perfectly matched the "ground-truth" labels provided by the human expert. However, a lower F1 score may suggest that CSs were either erroneously missed or falsely detected. For quality assessment, we also computed the post-CS firing rate of SSs, a signatory feature immune to labels detected by the human expert, which served as a reliable and objective criterion for the identification of a CS. Finally, the resulting CS waveforms were scrutinized by visual inspection.

RESULTS

Our goal in this study was to develop an algorithm for CS detection that matches human-level performance while at the same time minimizing the amount of effort needed in manual labeling and inspection. We achieved this by utilizing a machine learning approach in which a human expert manually labels a very small training data set, which is then used to train a CNN for feature extraction (Fig. 2). We also added additional postprocessing (but still automated) steps that significantly increased the robustness of our algorithm. To establish the utility of our approach for the wider community, we also compared its performance to that of two established methods from the literature. In what follows, we summarize the objective measures of our algorithm's performance. As we show, our algorithm currently outperforms the existing methods in CS detection. Our code and data sets are both available freely for adaptation to individual laboratories' needs and with step-by-step tutorials on use.

Objective Quality Measure Confirms Identity of Complex Spikes

It is well-established that SS firing rate decreases during 10–20 ms after the emission of a CS (Bell and Grimm 1969; Eccles et al. 1966; Latham and Paul 1971; McDevitt et al. 1982; Thach 1967; Fig. 1*F*). This physiological feature, which is independent of the subjective assessment of the human expert, provided us with an important means for objectively measuring the CS labeling quality of our CNN-based algorithm. For 159 PCs, we evaluated SS firing rates before and after the occurrence of CSs detected by our algorithm. As depicted in Fig. 3, CSs identified by our algorithm were followed by a clear and significant decrease in the neurons' SS firing rates (Fig. 3*A*). In the pre-CS period of 3–8 ms, the median SS firing rate of the 159 PCs was 54.9 spikes/s; this dropped to 1.8 spikes/s in the post-CS period of 3–8 ms (Fig. 3*B*, Wilcoxon signed-rank test: $P = 2.1 \times 10^{-35}$). Also, this effect was clearly visible at the level of single neurons (Fig. 3, *B* and *C*), suggesting that the overall suppression of SS firing rate across all PCs (Fig. 3, *A* and *B*) was not merely a consequence of contributions made only by a fraction of the PCs. In the next section, this drop in median firing rate will be compared with that obtained when other algorithms were applied to the same data to demonstrate that our algorithm performs significantly better and with much fewer false positives.

Our Algorithm Outperforms Existing Algorithms

To demonstrate the efficiency of our algorithm, we compared its performance to two other existing approaches. The first approach, the MSD (see MATERIALS AND METHODS), is an online spike sorting application that was based on a template matching algorithm suggested by Wörgötter et al. (1986). Although initially designed to detect fast spiking events such as SSs, the potential of the MSD was quickly realized by several laboratories to detect events with more complex morphological waveforms such as CSs (e.g., Catz et al. 2005). Specifically, the approach with the MSD in terms of CS detection has traditionally been to use this application as an initial "coarse" detector of potential CSs, which was then followed by extensive manual labeling by experts. The second method that we compared our algorithm to was the recent one by Zur and Joshua (2019), which was based on PCA to separate CSs from SSs. This particular algorithm is important to compare with because, like ours, it takes advantage of the deflection in LFP signals occurring at the time of CSs. It is also the latest algorithm available in the literature.

We first investigated the proportion of CSs identified by our algorithm in addition to those found by the other approaches. Across the 159 PCs, our algorithm found a median of 32% additional CSs as compared with the MSD and 5.9% additional CSs as compared with the PCA-based algorithm. In contrast, the additional CSs detected exclusively by the MSD approach but not by our algorithm were significantly less (1.9%; $P = 1.4 \times 10^{-20}$ Wilcoxon signed-rank test; Fig. 4*A*), and the ones found exclusively by the PCA-based algorithm but not by our algorithm were also less (4.3%; $P = 0.097$; Wilcoxon signed-rank test; Fig. 4*D*).

Our algorithm also led to significantly less false positives than both the MSD- and PCA-based algorithms. To demon-

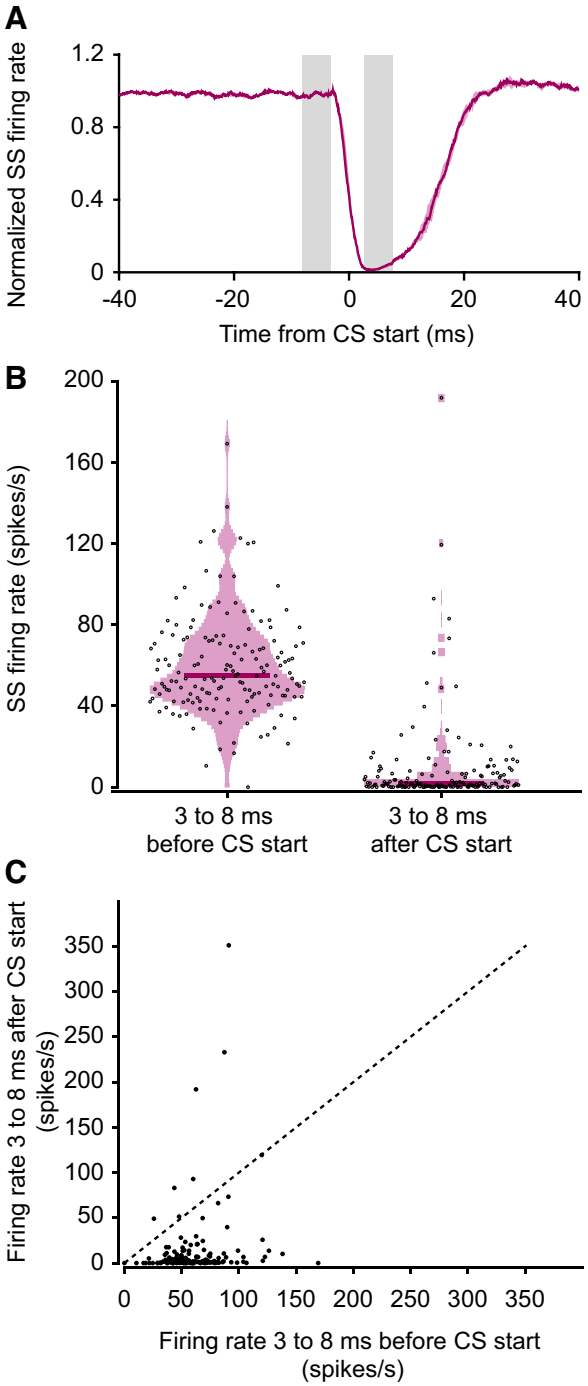


Fig. 3. Decrease of simple spike (SS) rate after complex spikes (CSs). *A*: baseline-normalized median SS firing rate aligned to the start of CSs detected by our algorithm. Data show median ± confidence intervals (obtained by bootstrapping) over 159 Purkinje cells (PCs). Gray shaded regions correspond to the period of 3 to 8 ms before and after CS start that are used for comparing SS firing rates before and after CS start, respectively. *B*: violin plots showing SS firing rate 3 to 8 ms before and after CS start. Each dot represents the average SS firing rate aligned to start time of all CSs corresponding to their respective PCs predicted by our algorithm. Thick lines indicate the median SS firing rate of all PCs. *C*: scatter plot comparing the SS firing rate of each cell (black dots), aligned to CS start, 3 to 8 ms before and after CS start. Dashed back line is the unity line.

strate this, we measured SS firing rate during a post-CS period (3–8 ms after the start times of putative CSs). We did so specifically for CSs that were found exclusively by each algorithm but not the others. A lack of sufficient pause in SS firing provided an objective physiological measure of a falsely identified CS (a true CS should have a pause in SS firing after its occurrence). In Fig. 4*B*, we found that there was a reliable pause in SS firing rate for CSs that were detected by both our algorithm and the MSD (Fig. 4*B*, gray). However, the MSD approach clearly had more false positives than our algorithm because the CSs detected exclusively by the MSD method (and not by our algorithm) had much higher SS firing rates after "putative" CS start times than true detected CSs (Fig. 4*B*, online sorter only data). These exclusively detected CSs were therefore most likely false detections. In contrast, for CS events detected exclusively by our algorithm but not by the MSD, there was still a strong pause in SS firing rate after CS event detection (Fig. 4*B*, our algorithm only data). This means that there were genuine CSs that were missed by the MSD method. The same conclusions could also be reached when we compared our algorithm to the one based on PCA (Fig. 4*E*). Therefore, our algorithm had fewer false positives than both of the other algorithms.

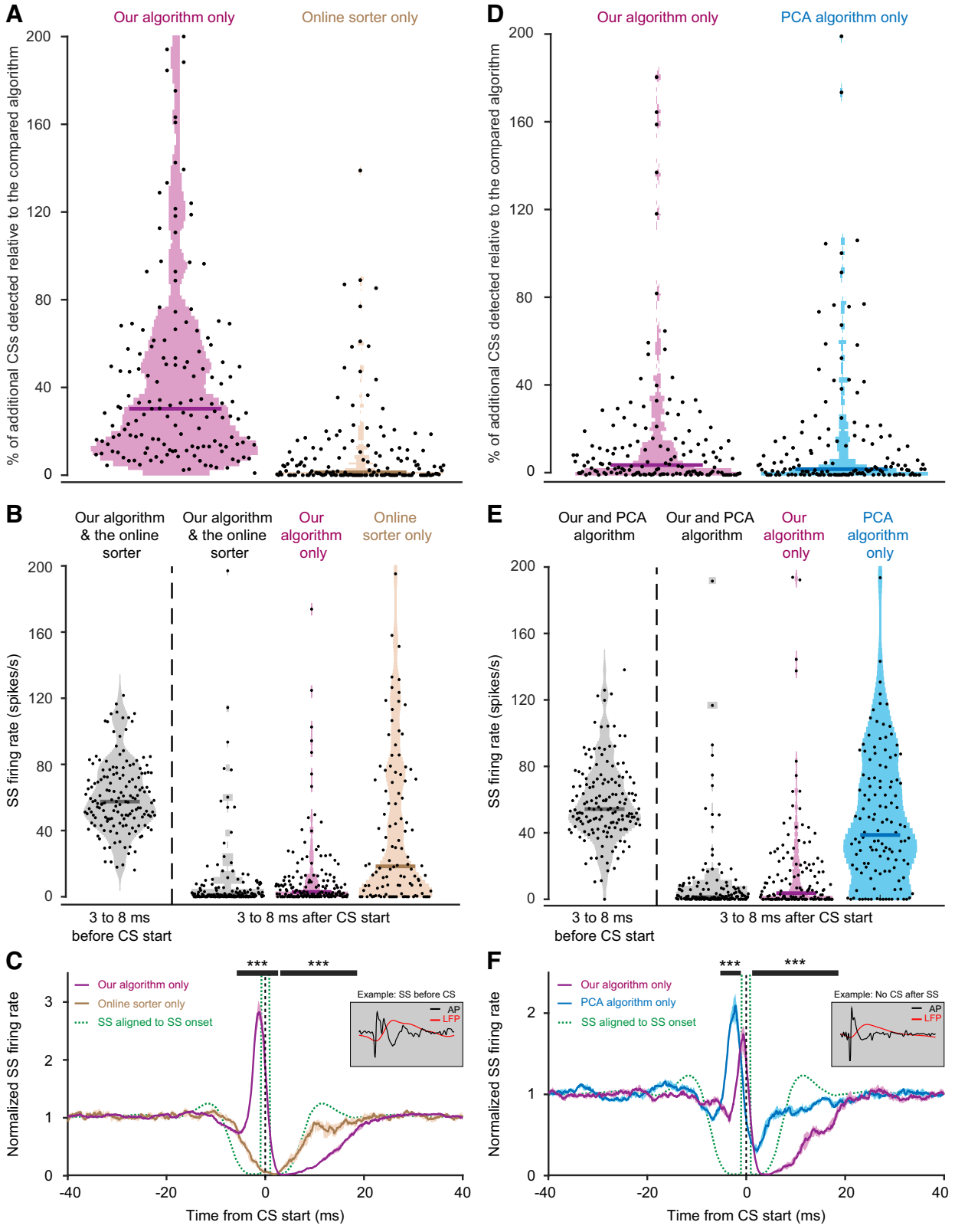
A closer look at the time courses of SS firing rates aligned to the start times of CSs detected exclusively by each algorithm revealed the scenarios that may explain the differential performance across algorithms (and our algorithms' overall superior performance). Specifically, Fig. 4, *C* and *F*, displays the median of normalized SS firing rate aligned to the start times of putative CSs from the pool of 159 PCs. The probability of SS occurrence for CSs detected by the MSD algorithm only (but not ours) resembled a typical autocorrelation function reflecting the refractory period of well-isolated SSs (Fig. 4*C*, dashed green curve). This suggests that the MSD algorithm tended to falsely report a CS at the time of a SS. The MSD approach also often missed CSs when SSs occurred less than 5 ms before their start times (Fig. 4*C*), which is a scenario that is known to modify the shape of CS waveforms (Servais et al. 2004). These waveforms were missed by the MSD algorithm as they differed from the template defined by the experimenter. However, our algorithm did not miss these CSs with modified waveforms (also see Fig. 10). It is the accumulation of SSs occurring just before these CSs detected exclusively by our algorithm (and not by the MSD algorithm) that explains the sharp increase in SS firing rate (Fig. 4*C*) that we observed ~3 ms before the start times of CSs. That is, because this figure shows only the CSs exclusively detected by our algorithm and not the MSD, and because SSs near CS onset modify CS shape, the CSs exclusively detected by our algorithm were ones in which SSs were so close to CS onset that they modified CS waveforms enough for the MSD to completely miss them. Interestingly, the additional CSs only detected by the PCA-based algorithm (and missed by our algorithm) also resulted in a peak in SS firing (Fig. 4*F*). At first glance, one might erroneously argue in favor of the PCA-based algorithm for being more sensitive than ours in detecting CSs. However, a closer look at the much weaker pause in SS firing rate that follows the peak clearly suggests that these additional CSs were actually not real CSs (Fig. 4*F*). They were likely false detections due to potential artifactual LFP modulations around SS events (Fig. 4*F*, *inset*).

To further demonstrate the performance differences among all three algorithms, Fig. 5 shows explicit example "CS" waveforms from five PCs with averaged LFP and action potentials. For each

cell in this figure, we show the average CS waveforms that were detected exclusively by one of the algorithms only but not by the other two. In summary, our algorithm was both more sensitive and less error prone than the online sorting application (MSD) as well as the PCA-based algorithm.

Our CNN Approach Reaches Human Expert-Level Performance

We also evaluated to what extent the predictions from the three approaches agreed with labels from a human expert. To this end, we computed the F1 score (see MATERIALS AND



METHODS) on short recording segments from the same 159 neurons; each currently tested neuron was excluded from the training set. For these segments, we had "ground-truth" labels from the human expert. The F1 score is a measure of consistency in performance between an algorithm and the human expert. As shown in Fig. 6*A*, our algorithm agreed best with the human expert on all CS labels, reflected by an F1 score of 1 or near 1 (Fig. 6*A*). A comparison of F1 scores between our algorithm and the PCA algorithm (Fig. 6*B*, *top*), as well as our algorithm and the online sorter MSD (Fig. 6*B*, *middle*), clearly reveals that for a majority of PCs (52% in the first comparison and 79% in the latter) our algorithm achieved overall higher F1 scores than the other approaches. Comparing the F1 scores of the PCA-based algorithm to the ones achieved by the MSD (Fig. 6*B*, *bottom*), suggests that the former approach also outperformed the latter in a majority (64%) of PCs. In sum, the predictions by our approach were more "human-like" than the ones labeled by the MSD and PCA-based algorithms.

Our algorithm also did not need extensive training sets to achieve good performance in terms of the F1 score. To show this, we plotted the performance of our algorithm as a function of the amount of training data that we used to optimize the CNN's weights (Fig. 6*C*). With a training set of ~35 s, our algorithm already led to better median F1 score performance than both the PCA-based and MSD approaches ($P = 1.2662 \times 10^{-5}$and $P = 7.2048 \times 10^{-10}$ respectively, Wilcoxon signed rank test) (Fig. 6*C*).

We also summarized the performance of our algorithm against the two other algorithms using confusion matrices in Fig. 6*D*. The total number of truly detected CSs (true positives) relative to the human expert was highest in the case of our algorithm (2,053) as compared with the other algorithms (1,908 for PCA and 1,578 for MSD). Similarly, the sum of our algorithm's false positive and false negative rates was the lowest. It should be noted here that, in this context, true positive, false positive, and false negative rates are always reported relative to the human expert labels, unlike our use of these terms in our analysis of SS pauses across different algorithms (Fig. 4).

Finally, for seven PCs, we asked our human expert to fully label the entire recorded data for each neuron, instead of only a subset (MATERIALS AND METHODS). We then compared the CS labels of the three algorithms to the ones placed by the human expert on the entire records of the neurons (spanning a time range of ~8–14 min of neural recording). Overall, the predictions of our algorithm agreed very well with the human labeling (Fig. 7, see "Expert vs. Our algorithm"). A few events were identified as CSs by our algorithm but not by the human expert. However, also the waveforms of these events matched the waveforms of CSs that were labeled by the human expert (Fig. 7, *cells 3*, *5*, and *6*), indicating that the CSs ignored by the expert were indeed genuine CSs (they were probably reflecting mental lapses during manual labeling by the expert). For one of the PCs, the waveforms of additionally detected CSs indicated that our algorithm mistakenly labeled some SSs as CSs (Fig. 7, *cell 7*). These false positive detections, whose average predictive probability remained above the threshold (0.5) for more than 3 ms and were not removed during automatic postprocessing, however, would appear as isolated clusters after dimensionality reduction (Fig. 2*C*). Hence, such false detections could be easily removed post hoc by inspecting the properties of the CSs in the respective isolated cluster. For false positive labels, the average duration of the SS pause (i.e., 15–20 ms) after these events would also be reduced to the average refractory period of SSs in this recording. As compared with the human expert, the PCA-based algorithm resulted in more false positives (Fig. 7, see "Expert vs. PCA algorithm," *cells 1*, *3*, *4*, and *7*) and false negatives (*cells 3* and *5*). The MSD made mistakes mostly because of false negatives (Fig. 7, see "Expert vs. Online sorter").

The comparison with human labels further showed that our algorithm reliably identified the ends of CSs and, considering the knowledge of CS start, provided a quantitative estimate of CS duration. For the recording segments from the 159 PCs, we compared the end times of all CSs that were detected by both our algorithm and the human expert. Correspondingly, average CS durations per cell predicted by our algorithm and the human expert were highly correlated ($\rho = 0.89$, $P = 6.15 \times 10^{-41}$, Spearman correlation; Fig. 8*A*). In light of a possible CS duration code supplementing a CS rate code (Herzfeld et al. 2015, 2018; Junker et al. 2018; Warnaar et al. 2015; Yang and Lisberger 2014), it is important to precisely identify the end times of CSs and to track changes in CS duration in conjunction with behavioral changes even within individual PCs, a particularly tedious task for the expert who has to scrutinize the data. Our algorithm was indeed capable of identifying small variations in CS duration similar to the expert. This is indicated by a strong correlation ($\rho = 0.5$, $P = 2.09 \times 10^{-102}$, Spearman correlation; Fig. 8*B*) of the residuals of human-labeled and algorithm-labeled CS end times of the selected 159 PCs,

Fig. 4. Comparison of complex spike (CS) detection by our algorithm, the principal component analysis (PCA)-based algorithm and the online sorter application Multi Spike Detector (MSD). *A*: violin plots showing the percentage of additional CSs detected exclusively by our algorithm and the online sorter. The percentage of additional CSs detected by an algorithm was calculated using the formula: (CSs detected by algorithm − CSs detected by both/CSs detected by both) × 100, where 100% corresponds to the number of CSs detected by both methods. Our algorithm detected significantly more CSs than the MSD. *B*: violin plots showing simple spike (SS) firing rate aligned to the start of the CSs predicted by both algorithms (gray) or of the events additionally labeled as CSs by either our algorithm (pink) or the online sorter (beige). The decrease in SS firing after CSs, predicted by our algorithm but not by the online sorter, indicates a higher sensitivity of our algorithm. *A* and *B*: each dot represents the average SS firing rate aligned to all CSs for the recording of 1 neuron. Thick lines indicate the median. *C*: pause in the baseline normalized median (±confidence intervals) SS firing rate following a CS. The sharp increase in SS firing rate ~3 ms before CS start (vertical dashed line in black), observed only for CSs detected by our algorithm (pink), and not the MSD (beige), suggests that these SSs occurring shortly before the start of CSs might have altered their waveform. Note how the pause in the SS firing due to CSs detected by the MSD (beige) resembles the SS autocorrelation (green dashed line). Only our algorithm was sensitive enough to detect such CSs with altered waveforms. Black bars on *top* show (with ***) intervals with a significant difference between the 2 traces (random permutations cluster corrected for multiple comparisons). AP, action potential. *C*, *inset*: an example of such a waveform. *D*, *E*, and *F*: Same as in *A*, *B*, and *C*, except now the comparison is made to the PCA-based algorithm (cyan). The sharp increase in SS activity just before the start of CSs detected by the PCA-based algorithm may suggest that their algorithm was also sensitive enough to capture changes in CS waveform. However, unlike the pause in SS firing induced by CSs detected by our algorithm, the pause observed in their case was much weaker. This suggests that the additional events detected by the PCA algorithm were not real CSs but rather other events like SSs paralleled by deflections in the local field potential (LFP) signal. *F*, *inset*: an example of such false detection.

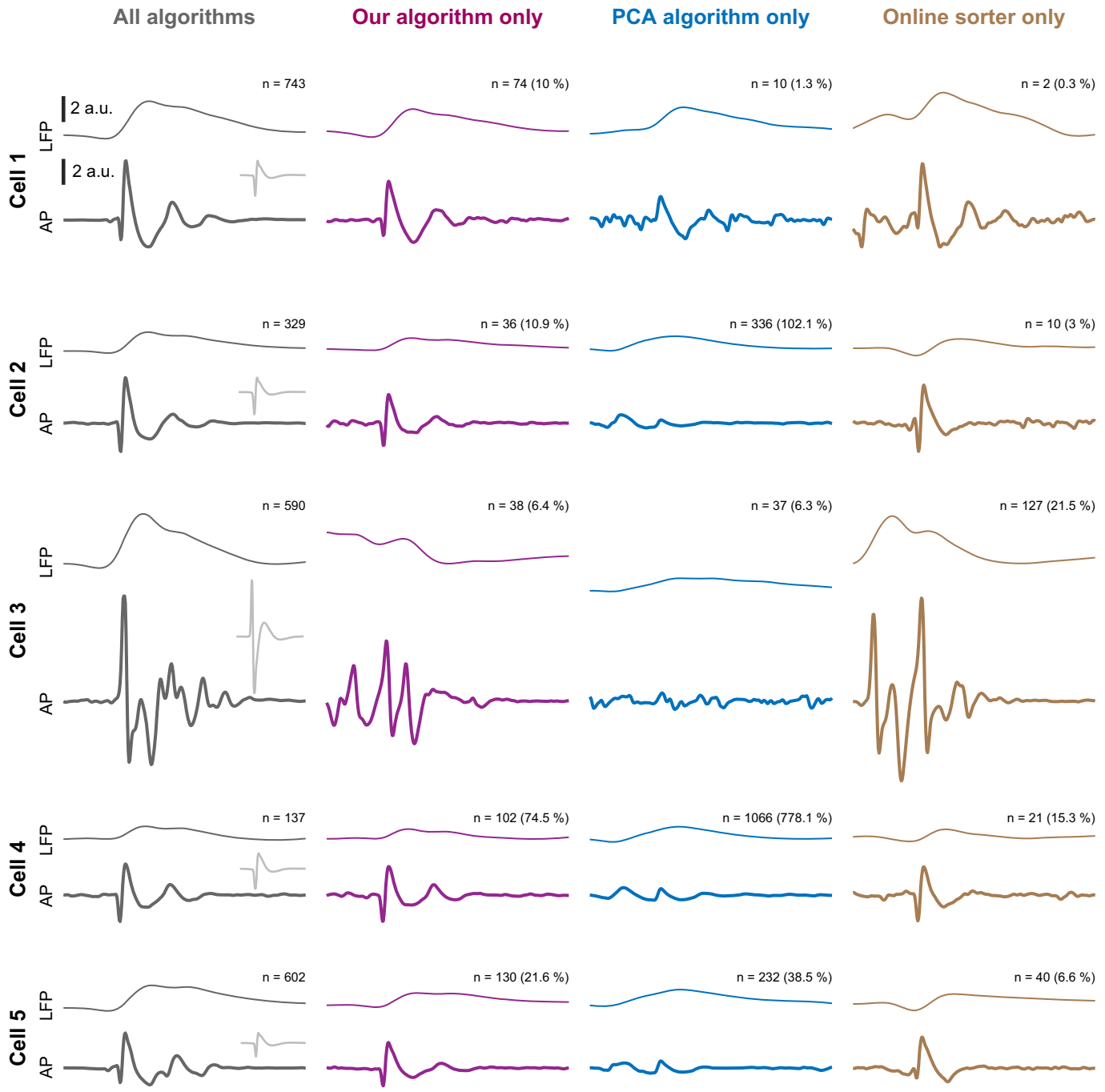


Fig. 5. Waveforms of events labeled as complex spikes (CSs) by our algorithm, the principal component analysis (PCA)-based algorithm, and the online sorting application Multi Spike Detector (MSD). Examples from 7 neurons showing the average waveform in the local field potential (LFP) and action potentials of CSs detected exclusively by all 3 methods: our algorithm only (pink), the PCA-based algorithm only (cyan), and the online sorter only (beige). CSs detected commonly by all 3 algorithms are shown in gray. Averaged simple spike waveforms (light gray) of each Purkinje cell (PC), scaled down by 50% relative to the CSs, are shown in *column 1*, *insets*.

obtained by subtracting the mean CS duration of the respective PC (Fig. 8*C*). As shown in Fig. 8*C*, the estimate of CS end times provided by our algorithm and the human expert differed only very slightly.

Practical Considerations for Using Our Algorithm

Our CNN-based algorithm uses the LFP and action potential signals simultaneously as input signals that pass through a series of steps to deliver CS start and end times as the final output. These steps have been summarized in Fig. 9. In short, the workflow of our algorithm can be divided into three main stages. The first stage, "Network training" (Fig. 9*A*), requires segments of manually labeled inputs as well as action potential and LFP signals. These act as the training set. The second stage, "CS detection" (Fig. 9*B*), is fully automated (including automated postprocessing) and utilizes the network weights learned during training to detect CS events among recordings of new sets of PCs. For this stage, the action potential and LFP

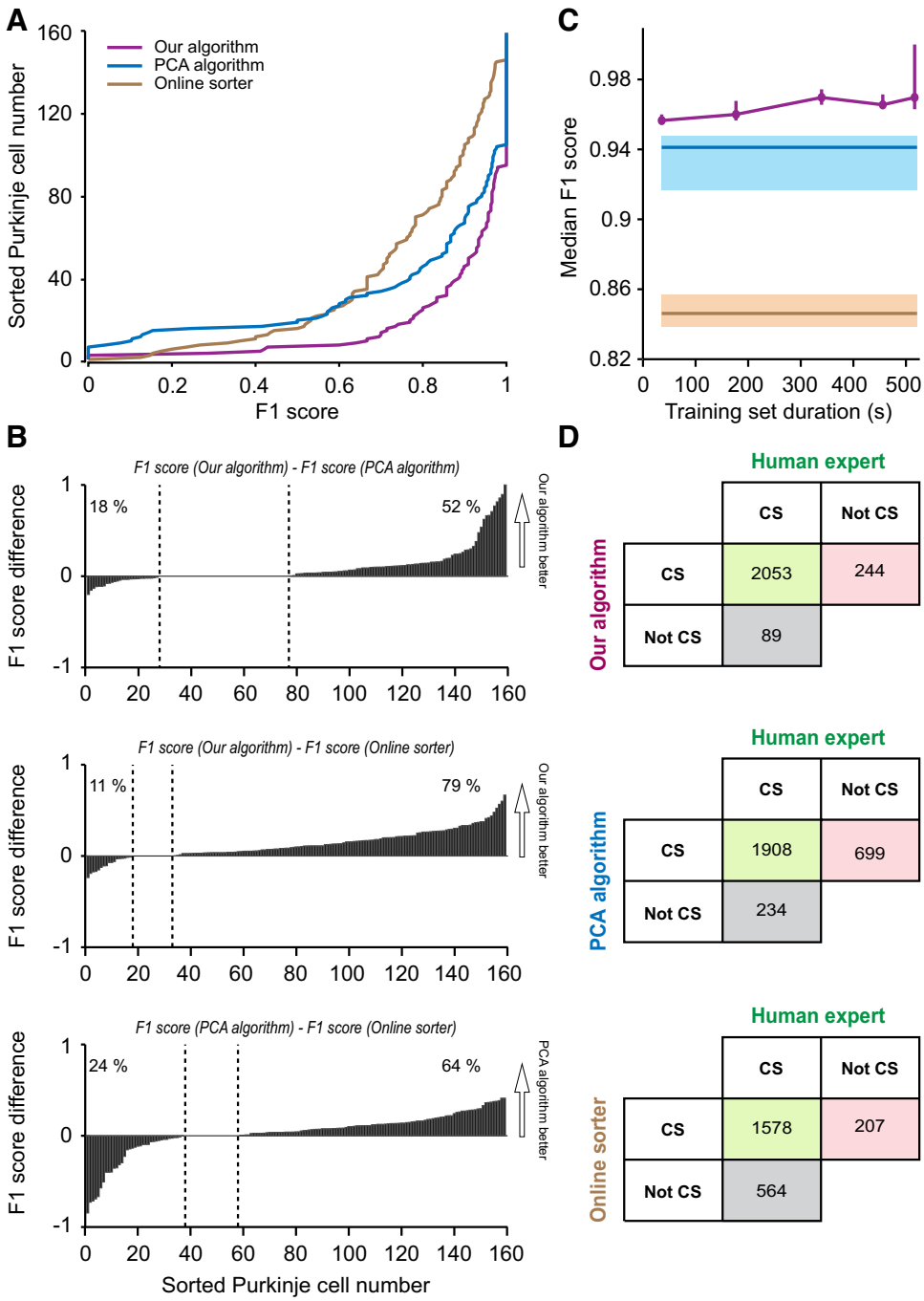


Fig. 6. Classification agreement of our algorithm, the principal component analysis (PCA)-based algorithm, and the online sorter application (MSD) with a human expert. *A*: distribution of F1 scores of our algorithm (pink), the PCA-based algorithm (cyan), and the online sorter (beige) computed by comparing complex spike (CS) labels with the human expert. Data from 159 neurons. *B*: difference between F1 scores obtained by our algorithm and the PCA-based algorithm (*top*), our algorithm and the online sorter (*middle*), and the PCA-based algorithm and the online sorter (*bottom*). *C*: F1 score of our algorithm as a function of the total duration of the training set (pink). Filled circles indicate the median, and error bars represent the 95% confidence interval of the median obtained by bootstrapping. As a reference, the F1 score achieved by the online sorter (beige) and by the PCA-based algorithm (cyan) are also displayed. Thick lines indicate the median and the shaded area represents 95% confidence interval of the median obtained by bootstrapping. *D*: confusion matrixes summarizing the CS detection performance of all 3 algorithms relative to the human expert. Green boxes represent the correctly detected CSs (true positives), gray boxes represent the falsely missed CSs (false negatives), and the red boxes represent the falsely detected CSs (false positives).

signals from an entire recording can be passed to the algorithm in one shot without the need for segmentation, which is automatically implemented by the algorithm.

One of the key requirements for correct CS classification is the quality of the recorded PC signal, which may naturally depend on several factors. For example, subtle drifts between electrode tip and the cell body during a recording session can lead to sudden or gradual changes in the signal-to-noise ratio of the PC signal and potentially change the morphology of the CS waveform. Also, several SSs firing in close proximity to each other might lead to complex waveforms that may erroneously be detected as CS events. Furthermore, there is also a possibility of CS waveforms being modified by the presence of preceding SSs (Servais et al. 2004; Zang et al. 2018). For our algorithm to be more resilient to such influences, it utilizes the three automatic postprocessing steps at the output of the CNN (see *Postprocessing* in MATERIALS AND METHODS). These postprocessing steps allow easy optimization of our algorithm's output.

The very final stage, "Postverification" (Fig. 9*C*), allows users to scrutinize every PC one last time to confirm whether the events marked by the algorithm were real CSs or just false positives.

We now summarize how to handle the few cases in which special care may be warranted when using our algorithm. Specifically, while in most cases our algorithm performed accurately in labeling true events, there were some cases in which distinct clusters of events were detected. During post-

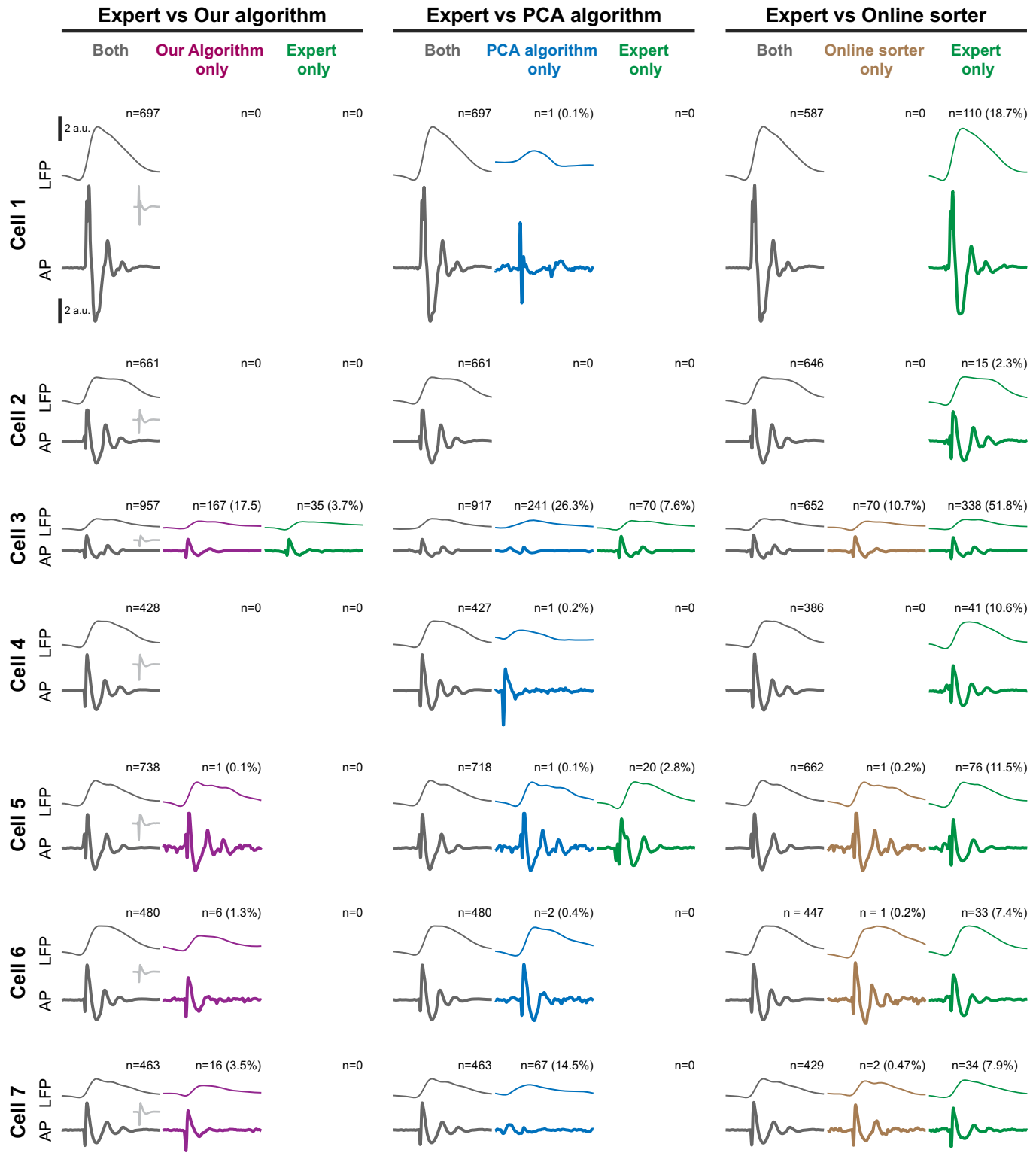


Fig. 7. Comparison of waveforms of events labeled as complex spikes (CSs) by all 3 algorithms with the human expert. Examples from 7 neurons showing the average CS waveforms in the local field potential (LFP) and action potentials (APs). PCA, principal component analysis. Expert vs. Our algorithm: CSs detected by the expert and our algorithm (*left*), our algorithm only (*middle*), and the expert only (*right*). Expert vs. PCA algorithm: CSs detected by the expert and the PCA algorithm (*left*), PCA algorithm only (*middle*), and expert only (*right*). Expert vs. Online sorter: CSs detected by the expert and online sorter (*left*), online sorter only (*middle*), and expert only (*right*). *Insets*, *1st column*: the averaged simple spike waveforms (light gray) of the corresponding Purkinje cells (PCs), scaled down by 50% relative to the CSs.

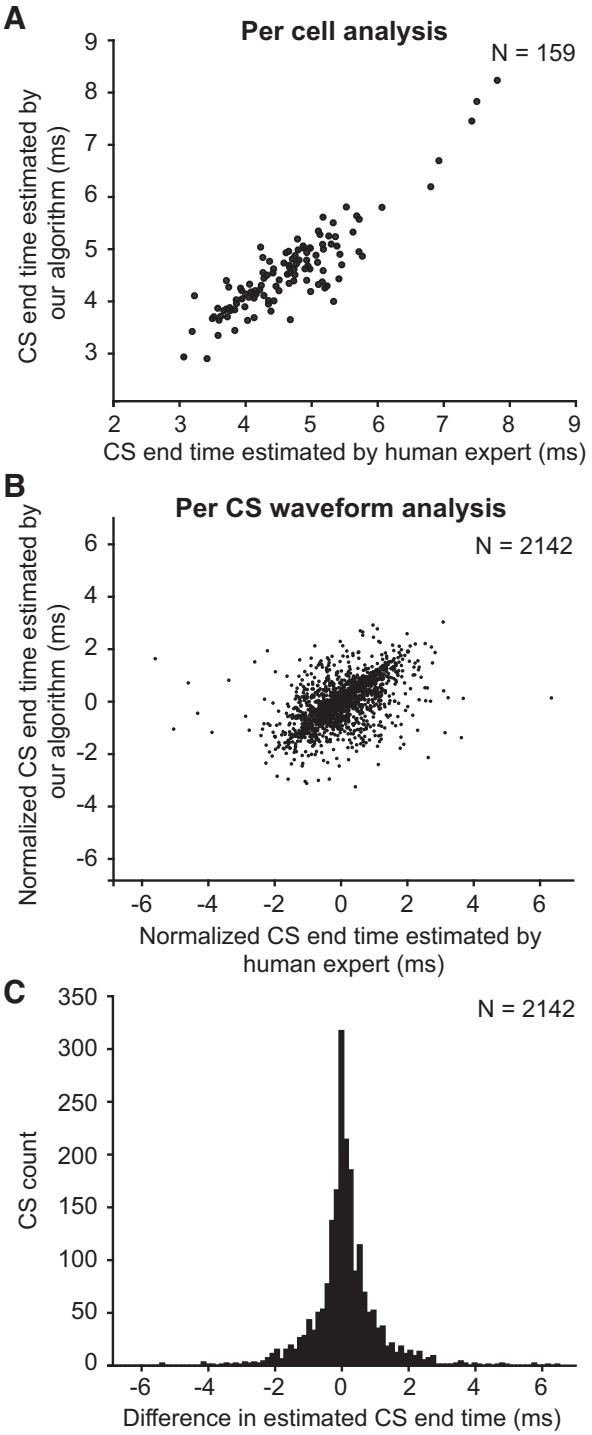


Fig. 8. Comparison of complex spike (CS) end times estimated by our algorithm and by the human expert. *A*: correlation of CS end times estimated by our algorithm and the human expert. Each dot shows the average end time of all CSs from 1 neuron. *B*: correlation of all CS end times pooled across the 159 neurons. The end time of each CS was normalized by subtracting the average end time of the respective neuron. *C*: distribution of difference in CS end times labeled by our algorithm and by the human expert. Data show all CSs detected by both our algorithm and the human expert in short recording segments from 159 neurons.

verification, a closer look at the waveforms of those distinct clusters with separate IDs revealed that these events were true CSs having their waveforms somehow modified. One such possibility is depicted in Fig. 10, *A* and *B*, where the amplitude of CS waveform is reduced gradually over time, most probably due to subtle position shifts between the tip of the recording electrode and the targeted neuron. These modified waveforms, seen as separate clusters (Fig. 10*A*), were in fact separated by time as seen in the raster plot of SSs aligned to CS start time (Fig. 10*B*): the CSs with *cluster ID 1* (red) appeared early during the recording session, while the ones with *cluster ID 3* (cyan) appeared later. Plotting the mean CS waveforms of *cluster 1* (early) and *cluster 3* (late) on top of each other (Fig. 10*B*, *top*) clearly shows a reduction in amplitude of LFP and action potential signals over time.

Also, it is likely that there can be interactions between SS occurrence and CS waveform appearance. Specifically, and as mentioned earlier, a study on PCs in nonanesthetized mice has demonstrated that the shape of the CS waveform can be altered by preceding SSs (Servais et al. 2004). Furthermore, recently conducted experiments on climbing fiber responses in PCs have revealed that the potassium currents, by means of voltage gating in a branch-specific manner, can regulate the climbing fiber driven calcium ion influx leading to changes in CS waveform amplitude (Zang et al. 2018). This may explain why the additional CSs detected by our algorithm might have potentially deceived other algorithms. An example of CS waveforms being modified by the presence of preceding SS is shown in Fig. 10, *C* and *D*, yellow trace. The genuine nature of the additional CSs detected by our algorithm in all cases was confirmed with the help of another prominent physiological marker: the pause in spontaneous firing activity of SSs 10–20 ms right after the occurrence of a CS. Although we observed this pause in the vast majority (92%) of PCs, there was only a small subset of PCs where the suppression of SS activity during the post-CS period was either very weak or missing (Fig. 3*C*). This may allow us to question the credibility of this physiological marker in confirming the presence of a CS. However, it is very unlikely since the lack of this SS pause may simply be an artifact of the poor signal-to-noise ratio of the recorded PCs that potentially led to falsely detected SS events by the online sorter MSD. In this study, we focused on SSs that were detected using the MSD. However, in principle, other CNN based approaches designed specifically for detection of fast spiking events (Rácz et al. 2020) could also be paired with our method.

DISCUSSION

This study proposes a largely automated approach to CS detection as a sensitive and reliable alternative to tedious and experience-dependent manual labeling. After training with surprisingly little data, our algorithm outperformed a widely used spike sorter as well as the latest PCA-based algorithm designed exclusively for CS detection. Moreover, our algorithm also easily caught up with the performance of an experienced human expert. Searching manually for rare events like CSs, amidst a sea of high-frequency SS signals, not only requires several weeks of tedious effort but, as demonstrated by research on visual search (Evans et al. 2011; Wolfe et al. 2005), is also error prone, even among experts. Our network renders CS detection not just feasible but, also, more objective and systematic. Steps describing the general workflow of our algorithm are summarized in Fig. 9.

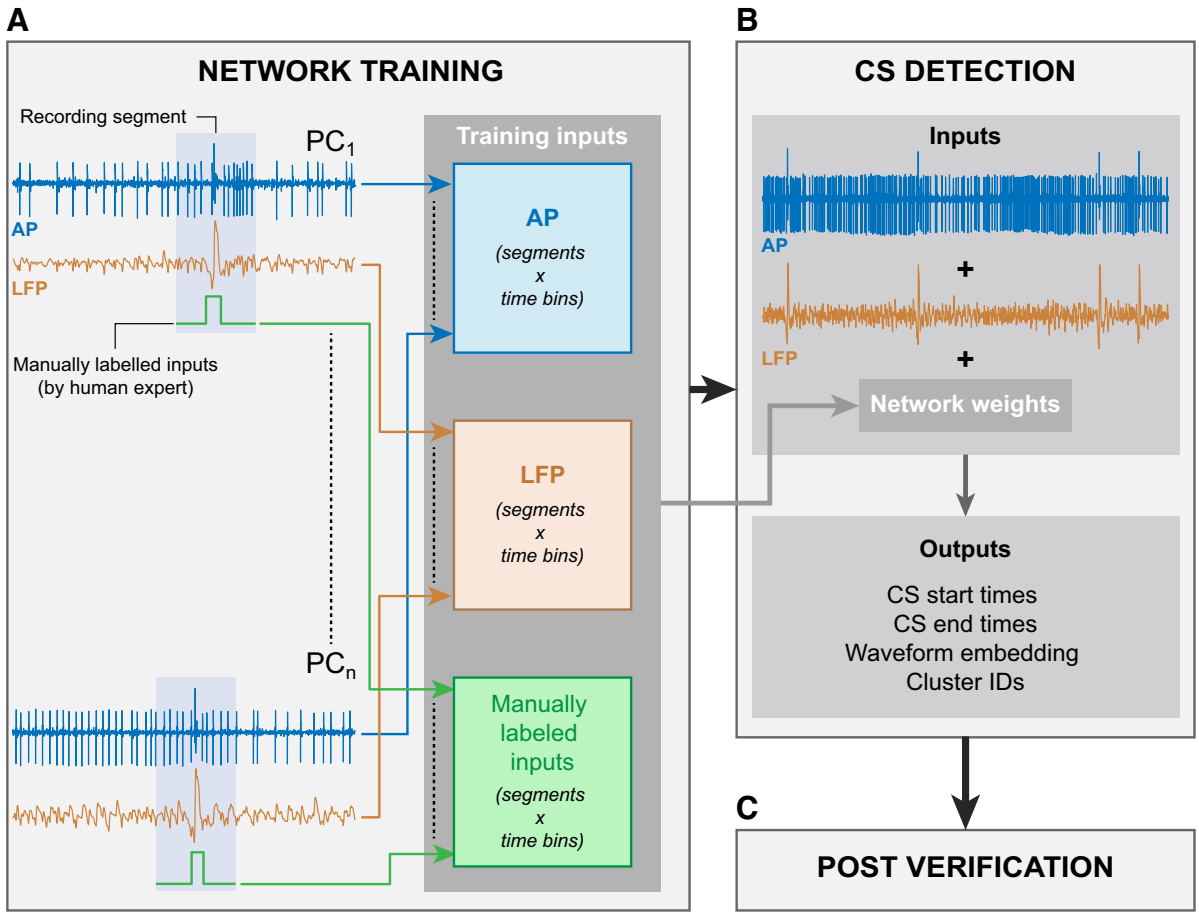


Fig. 9. Workflow for using our algorithm. *A*: the experimenter selects small segments of signal containing at least 1 complex spike (CS) each. Each segment is fed into the neural network in the form of 3 matrixes containing the action potentials (APs), the local field potentials (LFPs), and the labels separately. After training, the network outputs a set of weights. *B*: the weights are used for evaluating new signals. The output of the algorithm delivers information about CS start and end times, their IDs, and waveform shape. *C*: this information can be used by the analysts to manually verify the CS labels during postverification. PC, Purkinje cell.

Challenges Associated with Complex Spike Detection

When looking at the raw trace of a well-isolated PC neuron, like the one in Fig. 1*A*, one might argue that the problem of CS detection is rather trivial; a simple voltage-threshold-based detection could easily solve this problem. However, no matter how well isolated a PC may be, there may be fluctuations of the raw signal being recorded, which can occur at different time scales (whether fast or due to gradual drifts in the position of the neuron relative to the electrode). These fluctuations necessarily modify the start and end times of detected events using simple voltage thresholds. Therefore, even when clean signals can sometimes allow simple detection with thresholds for some applications, the relevance of CSs in the field of cerebellum research extends beyond "mere detection." Precise characterization of CSs and their overall durations, as well as the characterization of their morphology, may matter a great deal for function (Herzfeld et al. 2015, 2018; Junker et al. 2018; Warnaar et al. 2015; Yang and Lisberger 2014).

The major challenge that any approach for detecting CSs meets is the polymorphic complexity and rarity of these neural events (Warnaar et al. 2015). Experienced human experts may in principle reach a high level of agreement by using visual search to identify CS events. However, this approach is very tedious and therefore inevitably associated with fluctuations of attention, which jeopardizes the analyst's performance (Wolfe et al. 2005). The tediousness of the manual detection approach is increased even further if attempts are made to pinpoint the times of CS start and end or to identify distinct features of the CS morphology such as its spikelet architecture (Warnaar et al. 2015). Therefore, conventional spike sorters based on template matching (Catz et al. 2005; Dash et al. 2010; Herzfeld et al. 2015, 2018; Junker et al. 2018) or even simpler voltage-threshold crossings can be useful to facilitate visual inspection. However, the need to double check detected CS events will forestall gains in investments of time and effort only minimally.

Our Algorithm Is More Sensitive and Performs Better than Other Existing Approaches

Although initially designed to detect fast spiking events like SSs, the use of MSD was extended by cerebellum researchers to detect CSs (Catz et al. 2005). However, the challenges associated with CS detection made it difficult for the MSD to be used as a tool for CS detection and also limited its use in assisting visual inspection. To highlight these challenges, we compared the performance of our algorithm to that of the MSD, and clearly, our algorithm was better.

In a very recent publication, a PCA-based algorithm (Zur and Joshua 2019), designed exclusively for detecting CS

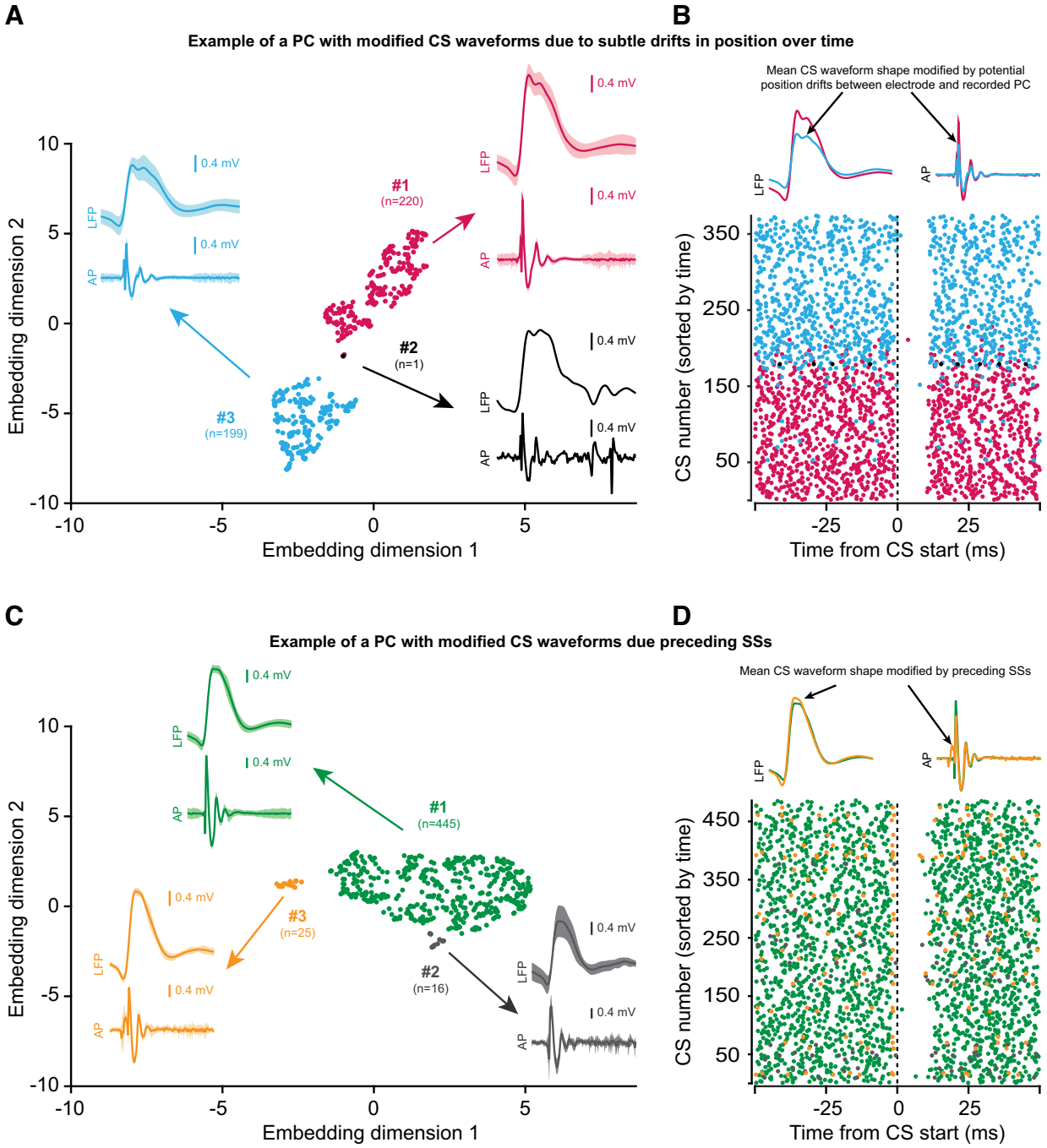


Fig. 10. Postverification of complex spikes (CSs) with modified waveforms. *A* and *B*: example of a Purkinje cell (PC) where the shape of the CS waveform was modified potentially due to subtle drifts between the position of the recording electrode tip and the targeted neuron over time. *A*: CSs detected by our algorithm based on their waveform shape appear as distinct clusters (*1–3*) with separate IDs in a 2-dimensional feature space. *Insets*: averaged CS waveforms ± SD in local field potential (LFP) and action potential for each cluster. *B*: raster plot showing simple spike (SS) activity aligned to all CSs detected by our algorithm sorted by time. Note how the clusters separated in feature space appear early (red) and late (cyan) during the recording session when sorted by time. The LFP and action potential signals of these late waveforms were clearly smaller in amplitude as compared with the earlier waveforms. Nevertheless, the pause induced in SS firing by both clusters was the same, suggesting that the CSs belonging to these clusters were the same. *C* and *D*: example of a PC where the shape of the CS waveform was modified by the presence of preceding SSs. Note how the presence of SSs just before the CSs, seen as large SDs around the averaged action potential waveform (*C*, *insets*), modified the average shape of the CS waveform as compared with that of the main (green) cluster (*D*, *top*). Also, as seen in the raster plot, the pause induced by these distinct CS clusters in SS firing was the same.

events by utilizing the LFP and action potential frequency bands, demonstrated improved sorting of CSs. Since our CNN-based algorithm was also trained on both frequency bands, we adapted our PC data as per the PCA-based algorithm's requirements to directly compare its performance with ours. The virtue of the PCA-based approach notwithstanding, a comparison of F1 scores reveals that it is clearly outperformed by our network-based method. Unlike the PCA-based approach that relies on an arbitrary threshold initially set by the user and may not capture all CSs in situations where oscillations may occur

in the LFP signal, our CNN-based algorithm uses the UMAP dimensionality reduction technique, thus making it more resistant to such changes. Moreover, our algorithm also extracts a number of key parameters on CS timing and morphology in a regularized and systematic manner that are not provided by the PCA-based approach, which are, as stated earlier, potentially important for understanding the functional role of CSs.

Not only was our algorithm more sensitive in detecting CSs, but it also rejected more false CSs, as compared with the other algorithms. This can best be seen in the example of Fig. 2*C*. In Fig. 2*C*, the *cluster 1* waveforms, despite sharing a similar shape of the initial spike component with the genuine CSs in *cluster 3*, appeared as a clearly separated group in our dimensionality reduced space. These erroneous waveforms were therefore safely rejected. On the other hand, waveforms belonging to *cluster 2*, neighboring the main *cluster 3*, were still accepted due to close resemblance of their features to the genuine ones.

It is well established (Eccles and Szentágothai 1967) that individual adult PCs, unlike PCs during early stages of development, usually receive input from only one climbing fiber. Only in rare cases also adult PCs have been found to be innervated by more than one climbing fiber (Nishiyama and Linden 2004). Consequently, it is usually very unlikely to find a second CS with completely different properties in addition to the first CS in records of individual PCs. However, we found a subset of nine (out of 159) PCs for which the CNN delineated a completely separate, large cluster of CSs in addition to the main cluster, suggesting input from more than one climbing fiber. The ability of our algorithm to identify such PCs is yet another demonstration of the high sensitivity and selectivity of our approach.

To test whether our algorithm could really take over the burden of labeling CSs manually, we made a one to one comparison of the performance of the CNN and the human expert on records of seven PCs for which all CSs had been labeled manually. Indeed, our algorithm's performance matched the human-level expertise in detecting CSs in all PCs, except for one in which additional CSs were detected by our algorithm (Fig. 7, *cell 7*). The location of these CSs in a distinct cluster in a two-dimensional feature space allowed the experimenter to easily evaluate the validity of the identification of the waveform as a potential CS and, in this case, to conclude that it was spurious. A similar comparison of the PCA-based algorithm's performance to that of the human expert yielded spurious detections in more cases, suggesting that the performance of our algorithm was closest to the performance of the human expert.

Our Algorithm Detects Start and End Points of CSs with Human-Level Performance

The prevailing idea of CSs serving as the "teaching-signal" for postsynaptic PCs (Albus 1971; Marr 1969), for which the occurrence of each CS event might be the only source of relevant information (Gellman et al. 1985; Rushmer et al. 1976), has been challenged by studies that demonstrated that the duration of action potential bursts fired by olivary neurons may vary and that this may be reflected by changes in the duration and the spikelet architecture of CSs (Bazzigaluppi et al. 2012; De Gruijl et al. 2012; Llinás and Yarom 1981; Maruta et al. 2007; Mathy et al. 2009; Rasmussen et al. 2013; Ruigrok and Voogd 1995; Zang et al. 2018). These observations have suggested that not only the occurrence of a CS but also its duration may be relevant for motor learning. Addressing this possibility requires experimenters to invest even more time to manually label the start and end times of CS waveforms in addition to just detecting the events themselves. Not surprisingly, given the amount of time and effort involved, only a handful of attempts have been made to test this idea (Herzfeld et al. 2015, 2018; Junker et al. 2018; Yang and Lisberger 2014) with inconsistent results. To achieve consensus, larger data sets collected under more diverse conditions would have to be explored, a necessity that researchers have been reluctant to meet because of the hassles of the manual timing analysis. Since our CNN-based approach is able to effortlessly follow the performance of the human expert in detecting the start and end of the CS waveforms, by applying the expert's "mental rules" learned during training, quantifying task-related changes in the architecture of CSs collected at different times in an experiment will become much more feasible in the future.

Deep Learning as a Research Tool

More broadly, deep learning allows modeling nonlinear relationships between input and output for which no analytical solutions may exist. It is exactly this property of deep learning that explains why this machine learning approach has recently emerged as a potentially powerful research tool, which can tremendously reduce the workload of scientists (Bellet et al. 2019; Cireşan et al. 2012; Havaei et al. 2017; Oztel et al. 2017). In light of recent developments, in which deep learning has been successfully utilized to not only design stimuli with controlled higher order statistics (Gatys et al. 2015), but also to model nonlinear relationships in neural data (Ecker et al. 2018), it is not hard to imagine that the full potential of deep learning will significantly boost the pace of neuroscientific research in the coming years. Certainly, in the case of cerebellar neurophysiology, we believe that our use of deep learning to detect the rare, but relevant, CS events will allow much renewed investigation of the contentious functional role of these highly peculiar spikes in motor control and beyond.

Conclusion

So far, all analysis involving CSs has been based on extremely laborious, manual, or semiautomated methods. This enormously slows down the pace of developments in the field. Our deep learning approach can reverse this reality. For example, for a database like ours (159 PCs), our approach requires the human expert to invest only 2–3 h of CS labeling for training purposes and another 3–4 h to later verify the results. Given that it takes a comparable time to manually label all CSs found in recordings of just one PC, this investment in time is negligible compared with the alternative of manually labeling all recorded PCs. Moreover, our automated algorithm performs this task on par with human experts, and it renders more systematic valuable information about the timing and morphology of CS waveforms. The algorithm has been made available for use via an open source implementation at https://github.com/jobellet/detect_CS with provisions for retraining the network to new users' own measurements. The data from all 159 PCs are available for download at https://figshare.com/articles/

Extracellular_recording_of_cerebellar_Purkinje_cells_and_labels_of_complex_spikes_from_expert/11872227. We strongly believe that the gains in time and reliability that our tool offers may substantially facilitate the quest for the functional role of the still largely mysterious CSs.

GRANTS

Supported by Deutsche Forschungsgemeinschaft (DFG) Research Unit 1847 "The Physiology of Distributed Computing Underlying Higher Brain Functions in Non-Human Primates" Projects FOR 1847-A3 (TH 425/13-2) and FOR 1847-A6 (HA 6749/2-1). Z.M.H. was also supported by DFG Collaborative Research Centre 1233 "Robust Vision" (Project No. 276693517).

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

A.M., J.B., Z.M.H., and P.T. conceived and designed research; A.M. performed experiments; A.M., J.B., M.E.B., J.I., Z.M.H., and P.T. interpreted results of experiments; A.M., J.B., J.I., Z.M.H., and P.T. prepared figures; A.M., J.B., M.E.B., Z.M.H., and P.T. drafted manuscript; A.M., J.B., M.E.B., Z.M.H., and P.T. edited and revised manuscript; A.M., J.B., J.I., Z.M.H., and P.T. approved final version of manuscript; J.B., M.E.B., J.I., and A.M. analyzed data.

ENDNOTE

At the request of the authors, readers are herein alerted to the fact that additional materials related to this manuscript may be found at the institutional website of one of the authors, which at the time of publication they indicate is for the electrophysiological signals and labels (No Behavior): https://figshare.com/articles/Extracellular_recording_of_cerebellar_Purkinje_cells_and_qjlabels_of_complex_spikes_from_expert/11872227; for the training set: https://figshare.com/articles/CS_detect_training_set/11891283; and for every network weights: https://figshare.com/articles/detect_CS_network_weights/11788833. These materials are not a part of this manuscript, and have not undergone peer review by the American Physiological Society (APS). APS and the journal editors take no responsibility for these materials, for the website address, or for any links to or from it.

REFERENCES

Albus JS. A theory of cerebellar function. *Math Biosci* 10: 25–61, 1971. doi:10.1016/0025-5564(71)90051-4.

Bazzigaluppi P, De Gruijl JR, van der Giessen RS, Khosrovani S, De Zeeuw CI, de Jeu MT. Olivary subthreshold oscillations and burst activity revisited. *Front Neural Circuits* 6: 91, 2012. doi:10.3389/fncir.2012.00091.

Bechert K, Koenig E. A search coil system with automatic field stabilization, calibration, and geometric processing for eye movement recording in humans. *Neuroophthalmology* 16: 163–170, 1996. doi:10.3109/01658109609009677.

Bell CC, Grimm RJ. Discharge properties of Purkinje cells recorded on single and double microelectrodes. *J Neurophysiol* 32: 1044–1055, 1969. doi:10.1152/jn.1969.32.6.1044.

Bellet ME, Bellet J, Nienborg H, Hafed ZM, Berens P. Human-level saccade detection performance using deep neural networks. *J Neurophysiol* 121: 646–661, 2019. doi:10.1152/jn.00601.2018.

Campello RJ, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. New York: Springer, 2013, p. 160–172.

Catz N, Dicke PW, Thier P. Cerebellar complex spike firing is suitable to induce as well as to stabilize motor learning. *Curr Biol* 15: 2179–2189, 2005. doi:10.1016/j.cub.2005.11.037.

Cireşan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification (Preprint). *arXiv* 12022745, 2012.

Dash S, Catz N, Dicke PW, Thier P. Specific vermal complex spike responses build up during the course of smooth-pursuit adaptation, paralleling the decrease of performance error. *Exp Brain Res* 205: 41–55, 2010. doi:10.1007/s00221-010-2331-2.

Davie JT, Clark BA, Häusser M. The origin of the complex spike in cerebellar Purkinje cells. *J Neurosci* 28: 7599–7609, 2008. doi:10.1523/JNEUROSCI.0559-08.2008.

De Gruijl JR, Bazzigaluppi P, de Jeu MT, De Zeeuw CI. Climbing fiber burst size and olivary sub-threshold oscillations in a network setting. *PLOS Comput Biol* 8: e1002814, 2012. doi:10.1371/journal.pcbi.1002814.

Dice LR. Measures of the amount of ecologic association between species. *Ecology* 26: 297–302, 1945. doi:10.2307/1932409.

Eccles JC, Llinás R, Sasaki K. The action of antidromic impulses on the cerebellar Purkinje cells. *J Physiol* 182: 316–345, 1966. doi:10.1113/jphysiol.1966.sp007826.

Eccles JC, Szentágothai J. *The Cerebellum as a Neuronal Machine*. Oxford, UK: Springer-Verlag, 1967.

Ecker AS, Sinz FH, Froudarakis E, Fahey PG, Cadena SA, Walker EY, Cobos E, Reimer J, Tolias AS, Bethge M. A rotation-equivariant convolutional neural network model of primary visual cortex (Preprint). *arXiv* 180910504, 2018.

Evans KK, Cohen MA, Tambouret R, Horowitz T, Kreindel E, Wolfe JM. Does visual expertise improve visual recognition memory? *Atten Percept Psychophys* 73: 30–35, 2011. doi:10.3758/s13414-010-0022-5.

Fujita Y. Activity of dendrites of single Purkinje cells and its relationship to so-called inactivation response in rabbit cerebellum. *J Neurophysiol* 31: 131–141, 1968. doi:10.1152/jn.1968.31.2.131.

Gatys LA, Ecker AS, Bethge M. A neural algorithm of artistic style (Preprint). *arXiv* 150806576, 2015.

Gellman R, Gibson AR, Houk JC. Inferior olivary neurons in the awake cat: detection of contact and passive body displacement. *J Neurophysiol* 54: 40–60, 1985. doi:10.1152/jn.1985.54.1.40.

Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, Pal C, Jodoin P-M, Larochelle H. Brain tumor segmentation with deep neural networks. *Med Image Anal* 35: 18–31, 2017. doi:10.1016/j.media.2016.05.004.

Herzfeld DJ, Kojima Y, Soetedjo R, Shadmehr R. Encoding of action by the Purkinje cells of the cerebellum. *Nature* 526: 439–442, 2015. doi:10.1038/nature15693.

Herzfeld DJ, Kojima Y, Soetedjo R, Shadmehr R. Encoding of error and learning to correct that error by the Purkinje cells of the cerebellum. *Nat Neurosci* 21: 736–743, 2018. doi:10.1038/s41593-018-0136-y.

Ito M. Neural design of the cerebellar motor control system. *Brain Res* 40: 81–84, 1972. doi:10.1016/0006-8993(72)90110-2.

Judge SJ, Richmond BJ, Chu FC. Implantation of magnetic search coils for measurement of eye position: an improved method. *Vision Res* 20: 535–538, 1980. doi:10.1016/0042-6989(80)90128-5.

Junker M, Endres D, Sun ZP, Dicke PW, Giese M, Thier P. Learning from the past: a reverberation of past errors in the cerebellar climbing fiber signal. *PLoS Biol* 16: e2004344, 2018. doi:10.1371/journal.pbio.2004344.

Kitazawa S, Kimura T, Yin P-B. Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature* 392: 494–497, 1998. doi:10.1038/33141.

Kostadinov D, Beau M, Blanco-Pozo M, Häusser M. Predictive and reactive reward signals conveyed by climbing fiber inputs to cerebellar Purkinje cells. *Nat Neurosci* 22: 950–962, 2019. [Erratum in *Nat Neurosci* 23: 468, 2020. 10.1038/s41593-020-0594-x.] doi:10.1038/s41593-019-0381-8.

Latham A, Paul DH. Spontaneous activity of cerebellar Purkinje cells and their responses to impulses in climbing fibres. *J Physiol* 213: 135–156, 1971. doi:10.1113/jphysiol.1971.sp009373.

Leznik E, Llinás R. Role of gap junctions in synchronized neuronal oscillations in the inferior olive. *J Neurophysiol* 94: 2447–2456, 2005. doi:10.1152/jn.00353.2005.

Llinás R. Eighteenth Bowditch lecture. Motor aspects of cerebellar control. *Physiologist* 17: 19–46, 1974.

Llinás R, Sugimori M. Electrophysiological properties of in vitro Purkinje cell dendrites in mammalian cerebellar slices. *J Physiol* 305: 197–213, 1980. doi:10.1113/jphysiol.1980.sp013358.

Llinás R, Yarom Y. Electrophysiology of mammalian inferior olivary neurones in vitro. Different types of voltage-dependent ionic conductances. *J Physiol* 315: 549–567, 1981. doi:10.1113/jphysiol.1981.sp013763.

Marr D. A theory of cerebellar cortex. *J Physiol* 202: 437–470, 1969. doi:10.1113/jphysiol.1969.sp008820.

Maruta J, Hensbroek RA, Simpson JI. Intraburst and interburst signaling by climbing fibers. *J Neurosci* 27: 11263–11270, 2007. doi:10.1523/JNEUROSCI.2559-07.2007.

Mathy A, Ho SS, Davie JT, Duguid IC, Clark BA, Häusser M. Encoding of oscillations by axonal bursts in inferior olive neurons. *Neuron* 62: 388–399, 2009. doi:10.1016/j.neuron.2009.03.023.

McDevitt CJ, Ebner TJ, Bloedel JR. The changes in Purkinje cell simple spike activity following spontaneous climbing fiber inputs. *Brain Res* 237: 484–491, 1982. doi:10.1016/0006-8993(82)90460-7.

McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction (Preprint). *arXiv* 180203426, 2018.

Medina JF, Lisberger SG. Links from complex spikes to local plasticity and motor learning in the cerebellum of awake-behaving monkeys. *Nat Neurosci* 11: 1185–1192, 2008. [Erratum in *Nat Neurosci* 12: 808, 2009.] doi:10.1038/nn.2197.

Nishiyama H, Linden DJ. Differential maturation of climbing fiber innervation in cerebellar vermis. *J Neurosci* 24: 3926–3932, 2004. doi:10.1523/JNEUROSCI.5610-03.2004.

Ohmae S, Medina JF. Climbing fibers encode a temporal-difference prediction error during cerebellar learning in mice. *Nat Neurosci* 18: 1798–1803, 2015. doi:10.1038/nn.4167.

Oscarsson O. Functional organization of olivary projection to the cerebellar anterior lobe. In: *The Inferior Olivary Nucleus: Anatomy and Physiology*, edited by Courville J, de Montigny C, Lamarre Y. New York: Raven Press, 1980, p. 279–290.

Oztel I, Yolcu G, Ersoy I, White T, Bunyak F. Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway, NJ: IEEE, 2017, p. 1195–1200.

Prsa M, Dicke PW, Thier P. The absence of eye muscle fatigue indicates that the nervous system compensates for non-motor disturbances of oculomotor function. *J Neurosci* 30: 15834–15842, 2010. doi:10.1523/JNEUROSCI.3901-10.2010.

Rácz M, Liber C, Németh E, Fiáth R, Rokai J, Harmati I, Ulbert I, Márton G. Spike detection and sorting with deep learning. *J Neural Eng* 17: 016038, 2020. doi:10.1088/1741-2552/ab4896.

Rasmussen A, Jirenhed D-A, Zucca R, Johansson F, Svensson P, Hesslow G. Number of spikes in climbing fibers determines the direction of cerebellar learning. *J Neurosci* 33: 13436–13440, 2013. doi:10.1523/JNEUROSCI.1527-13.2013.

Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. New York: Springer, 2015, p. 234–241.

Ruigrok TJ, Voogd J. Cerebellar influence on olivary excitability in the cat. *Eur J Neurosci* 7: 679–693, 1995. doi:10.1111/j.1460-9568.1995.tb00672.x.

Rushmer DS, Roberts WJ, Augter GK. Climbing fiber responses of cerebellar Purkinje cells to passive movement of the cat forepaw. *Brain Res* 106: 1–20, 1976. doi:10.1016/0006-8993(76)90069-X.

Servais L, Bearzatto B, Hourez R, Dan B, Schiffmann SN, Cheron G. Effect of simple spike firing mode on complex spike firing rate and waveform in cerebellar Purkinje cells in non-anesthetized mice. *Neurosci Lett* 367: 171–176, 2004. doi:10.1016/j.neulet.2004.05.109.

Sorensen TA. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol Skar* 5: 1–34, 1948.

Streng ML, Popa LS, Ebner TJ. Climbing fibers control Purkinje cell representations of behavior. *J Neurosci* 37: 1997–2009, 2017. doi:10.1523/JNEUROSCI.3163-16.2017.

Stuart G, Häusser M. Initiation and spread of sodium action potentials in cerebellar Purkinje cells. *Neuron* 13: 703–712, 1994. doi:10.1016/0896-6273(94)90037-X.

Thach WT Jr. Somatosensory receptive fields of single units in cat cerebellar cortex. *J Neurophysiol* 30: 675–696, 1967. doi:10.1152/jn.1967.30.4.675.

Thach WT. Discharge of Purkinje and cerebellar nuclear neurons during rapidly alternating arm movements in the monkey. *J Neurophysiol* 31: 785–797, 1968. doi:10.1152/jn.1968.31.5.785.

Warnaar P, Couto J, Negrello M, Junker M, Smilgin A, Ignashchenkova A, Giugliano M, Thier P, De Schutter E. Duration of Purkinje cell complex spikes increases with their firing frequency. *Front Cell Neurosci* 9: 122, 2015. doi:10.3389/fncel.2015.00122.

Wolfe JM, Horowitz TS, Kenner NM. Cognitive psychology: rare items often missed in visual searches. *Nature* 435: 439–440, 2005. doi:10.1038/435439a.

Wörgötter F, Daunicht WJ, Eckmiller R. An on-line spike form discriminator for extracellular recordings based on an analog correlation technique. *J Neurosci Methods* 17: 141–151, 1986. doi:10.1016/0165-0270(86)90067-1.

Yang Y, Lisberger SG. Purkinje-cell plasticity and cerebellar motor learning are graded by complex-spike duration. *Nature* 510: 529–532, 2014. doi:10.1038/nature13282.

Zang Y, Dieudonné S, De Schutter E. Voltage- and branch-specific climbing fiber responses in Purkinje cells. *Cell Reports* 24: 1536–1549, 2018. doi:10.1016/j.celrep.2018.07.011.

Zur G, Joshua M. Using extracellular low frequency signals to improve the spike sorting of cerebellar complex spikes. *J Neurosci Methods* 328: 108423, 2019. doi:10.1016/j.jneumeth.2019.108423.