NEUROSCIENCE
RESEARCH ARTICLE

Jeyathevy Sukiban et al. / Neuroscience 414 (2019) 168–185

Evaluation of Spike Sorting Algorithms: Application to Human Subthalamic Nucleus Recordings and Simulations

Jeyathevy Sukiban,[a,b,1] Nicole Voges,[b,*,1] Till A. Dembek,[a] Robin Pauli,[b,d] Veerle Visser-Vandewalle,[e] Michael Denker,[b] Immo Weber,[c] Lars Timmermann[a,c] and Sonja Grün[b,d]

[a]*Department of Neurology, University Hospital Cologne, Germany*

[b]*Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA BRAIN Institute I (INM-10), Jülich Research Centre, Germany*

[c]*Department of Neurology, University Hospital Giessen & Marburg, Marburg, Germany*

[d]*Theoretical Systems Neurobiology, RWTH Aachen University, Germany*

[e]*Department of Stereotactic and Functional Neurosurgery, University Hospital Cologne, Germany*

Abstract—An important prerequisite for the analysis of spike synchrony in extracellular recordings is the extraction of single-unit activity from the multi-unit signal. To identify single units, potential spikes are separated with respect to their potential neuronal origins ('spike sorting'). However, different sorting algorithms yield inconsistent unit assignments, which seriously influences subsequent spike train analyses. We aim to identify the best sorting algorithm for subthalamic nucleus recordings of patients with Parkinson's disease (experimental data ED). Therefore, we apply various prevalent algorithms offered by the 'Plexon Offline Sorter' and evaluate the sorting results. Since this evaluation leaves us unsure about the best algorithm, we apply all methods again to artificial data (AD) with known ground truth. AD consists of pairs of single units with different shape similarity embedded in the background noise of the ED. The sorting evaluation depicts a significant influence of the respective methods on the single unit assignments. We find a high variability in the sortings obtained by different algorithms that increases with single units shape similarity. We also find significant differences in the resulting firing characteristics. We conclude that Valley-Seeking algorithms produce the most accurate result if the exclusion of artifacts as unsorted events is important. If the latter is less important ('clean' data) the K-Means algorithm is a better option. Our results strongly argue for the need of standardized validation procedures based on ground truth data. The recipe suggested here is simple enough to become a standard procedure.

Key words: spike sorting, Parkinson Disease, subthalamic nucleus, spike train, ground truth evaluation, electrophysiology.

INTRODUCTION

The decomposition of extra-cellular multi-unit recordings into single unit activity is a prerequisite for studying neuronal activity patterns (Chibirova et al., 2005; Kühn et al., 2005; Einevoll et al., 2012; Todorova et al., 2014; Yang et al., 2014). The assignment of spikes to individual neurons based on the similarity of their spike shapes is a method commonly referred to as spike sorting (Lewicki, 1998; Quiroga, 2007; Quiroga, 2012). It is composed of three principal steps. First, the spikes are extracted from the band-pass filtered raw signal. Second, salient features of each spike waveform are identified. A common method to automatically extract such features is the principal component analysis Lewicki (1998); Quiroga (2007). Third, a sorting algorithm assigns spikes to putative single neuronal units using the extracted features. Many such spike sorting algorithms are available (Lewicki, 1998; Quiroga, 2007; Quiroga, 2012; Rossant et al., 2016; Chung et al., 2017; Yger et al., 2018) but they typically provide inconsistent results for the same data set (Brown et al., 2004; Wild et al., 2012; Knieling et al., 2016). Such differences in the sorting results affect the subsequent spike train analyses (Brown et al., 2004; Pazienti and Grün, 2006; Todorova et al., 2014). Therefore, a major challenge is to identify an appropriate spike sorting algorithm for a given data set, considering also its impact on the subsequent analysis (Lewicki, 1998; Todorova et al., 2014).

Extracellular recordings from the Subthalamic nucleus (STN, see Table 1 for a list of abbreviations) of patients with

*Corresponding author at: INT, Aix-Marseille University, France.
E-mail address: nicole.voges@gmx.com (Nicole Voges).
[1] Authors contributed equally (co-first authors).

https://doi.org/10.1016/j.neuroscience.2019.07.005

Table 1. List of abbreviations used throughout this study.

Abbreviation	Meaning
ED	experimental data
AD	artificial data
STN	Subthalamic nucleus
PD	Parkinson's Disease
DBS	Deep Brain Stimulation
OFS	Plexon Offline Sorter
TMS	Template Matching
KM(S)	K-Means (scan)
VS(S)	Valley Seeking (scan)
StEM(S)	standard Expectation Maximization (scan)
TDEM(S)	t-distribution Expectation Maximization (scan)
PC	principle component
rpv	refractory period violation
SD	standard deviation
2D	2 dimensional
TP	true positive
TN	true negative
FP	false positive (sorted)
FN	false negative
FPp	false positive (unsorted)
IS	isolation score
Di	internal cluster distance
ISI	inter-spike interval
LV	local coefficient of variation

Parkinson's Disease (PD), obtained during Deep Brain Stimulation (DBS) surgery provide important information about pathological activity patterns (e.g., Reck et al., 2009; Florin et al., 2012; Deffains et al., 2014). The analysis of the corresponding single unit activity contributes to identify and localize pathological patterns (Hutchison et al., 1998).

Besides common spike sorting problems such as bursting activities and overlapping spikes (Lewicki, 1998; Quiroga, 2007), or waveform changes induced by an electrode drift, the separation of single units in human STN data is particularly challenging (Knieling et al., 2016). Microelectrode recordings from brain areas densely packed with neurons, such as the STN (Hamani, 2004), contain spikes from a large number of neurons. The overall recording time is restricted to only a few minutes per recording site since the surgery is exhausting and the patients have to stay awake.[2] The short recording time does not allow to wait for stabilization of tissue and electrode. In contrast, animal studies allow for longer recordings so that it is possible to account for initial stabilization (Raz et al., 2000). Also, simultaneous intra- and extracellular recordings for calibration can be performed in animal studies (Harris et al., 2000) but this is not feasible during DBS surgery. Another advantage in animal studies is the usage of 4-wire close-by electrodes (i.e., tetrodes) or even polytrodes (Rey et al., 2015; Rossant et al., 2016). The resulting recordings generally enable a more accurate spike sorting because one neuron is registered at different wires allowing for triangulation (Harris et al., 2000; Buzsáki, 2004; Lefebvre et al., 2016; Rossant et al., 2016; Yger et al., 2018). In contrast, human DBS recordings are typically performed with up to five single-wire electrodes (McNaughton et al., 1983), typically inserted using a Ben-gun configuration (Gross et al., 2006; Florin et al., 2008; Michmizos and Nikita, 2010; Reck et al., 2012). These electrodes have a maximum diameter of 1 mm with a minimal distance of 2 mm. Thus, the insertion causes a considerable initial tissue movement and the spikes of one neuron are detected on one electrode only.

A few comparative spike sorting studies for human STN recordings have been introduced (Chibirova et al., 2005; Wild et al., 2012; Knieling et al., 2016). Wild et al. (2012) compares three widely used open source sorting toolboxes (WaveClus (Quiroga et al., 2004), KlustaKwik (Harris et al., 2000), and OSort (Rutishauser et al., 2006)) by applying them to artificial data with some STN characteristics. They conclude that WaveClus yields the best results, but does not perform optimally. Knieling et al. (2016) compares a new approach to sort STN spikes to OSort and WaveClus, using the artificial data from (Wild et al., 2012). Chibirova et al. (2005) demonstrate the application of the unsupervised spike sorting algorithm presented in (Aksenova et al., 2003) to both STN recordings from PD patients and artificial data with different noise levels.

Most of the above studies concentrate on open source spike sorting algorithms, whereas many studies recording from the human STN (e.g., Shinomoto et al., 2003; Moran et al., 2008; Schrock et al., 2009; Shimamoto et al., 2013; Yang et al., 2014; Kelley et al., 2018; Lipski et al., 2018) use a commercially available software, the 'Plexon Offline Sorter' OFS. Because of its frequent usage and relevance in the scientific community we focus our studies on the various sorting algorithms offered by the OFS. There are some comparative studies for spike detection and feature extraction (e.g., Wheeler and Heetderks, 1982; Lewicki, 1998; Adamos et al., 2008; Gibson et al., 2008; Yang et al., 2011), but less studies focus on clustering. Here, we concentrate the comparison of the results obtained with the following OFS cluster algorithms: Template Matching (TMS), K-Means (KM), Valley Seeking (VS), standard and t-distribution Expectation Maximization (StEM and TDEM, respectively). Varying the cluster algorithm, we use an identical detection procedure and the first two or three principal components (PCs) as features, since the number of PCs to be used for the sorting is another matter of debate (Hulata et al., 2002). This study does not deal with any parameter or feature selection optimization but aims for reproducible results that are directly comparable to each other. Thus, we apply *unsupervised* clustering without manual, i.e., user-specific, intervention. Still, there is one exception, namely manual TMS, which we use for comparison purposes.

Firstly, we apply all sorting algorithms to the experimental STN data (ED) recorded from PD patients. This enables us to depict the method-dependent differences in the ED sorting results (see Fig. 1) and to subsequently point out their considerable impact on the analysis of spike trains from

[2] The aim of the DBS procedure is not the recording itself but to locate the optimal stimulation site

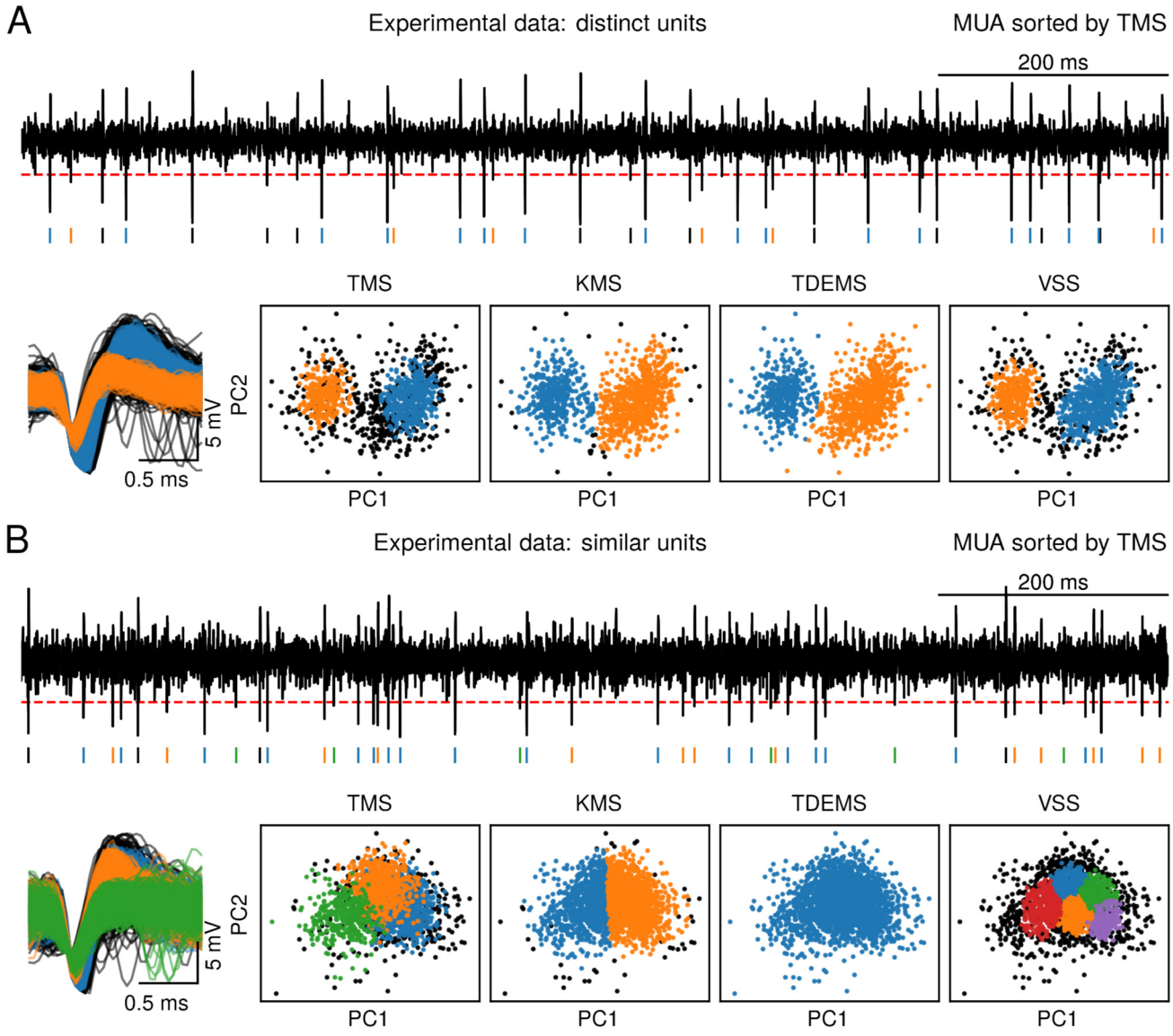

Fig. 1. Exemplary spike sorting of human STN recordings. (A) Two distinct single units and (B) several similar single units were extracted from two STN recordings (top traces with multi-unit activity MUA) by amplitude thresholding (horizontal red line). Colored vertical lines below the continuous traces indicate time stamps of potential spikes. Extracted spike waveforms are shown on the left, the corresponding clustering in 2D feature space on the right: colored dots represent spikes, unsorted events in black. Table 3 lists the Isolations scores and the percentage of refractory period violations for all clusters shown here. The corresponding internal cluster distances (Di) are ranging from 730 to 950, except for TDEMS in panel B with Di = 1100 and VSS in panel B with Di values in between 460 and 580 (not normalized).

real-world recordings. The evaluation of single unit assignments and properties yield significant differences in the spike sorting results and suggests a seemingly best method. For a quantitative comparison, however, ground truth data are necessary, i.e., data with known single unit assignments (e.g., Kretzberg et al., 2009; Wild et al., 2012; Yger et al., 2018). We therefore generate artificial data (AD) with known ground truth that include several features that are close to those of STN recordings. The spike sorting algorithms are then applied to the AD to evaluate their sorting quality. Based on this procedure, we are finally able to identify which methods works best under which circumstances.1–3

MATERIAL AND METHODS

We first briefly describe the ED, followed by a description of the AD generation. Then, we explain the main steps of spike sorting and finally, we detail the comparison and validation of the results of different clustering algorithms.

Experimental data

ED were recorded intraoperatively from six awake patients with tremor-dominant idiopathic PD undergoing STN-DBS surgery. The STN was localized anatomically with preoperative imaging and its borders were intraoperatively verified by inspection of multi-unit spiking activity (details are

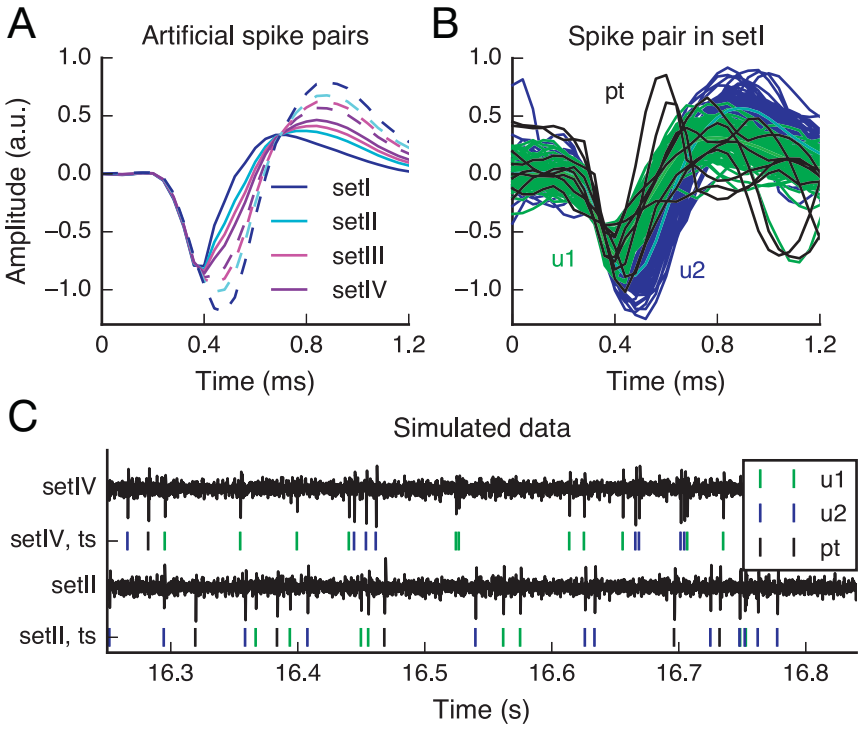

Fig. 2. Artificial data generation. (A) Four pairs (solid and dashed curves) of artificial spikes that are very distinct (setI, blue), distinct (setII, cyan), similar (setIII, magenta), and very similar (setIV, violet). (B) Exemplary setI spike pair (u1,u2) added to the noise (100 spikes each). Thick light blue and green lines indicate the average spike waveforms of u1 and u2, respectively, black lines indicate perturbations (pt). (C) Simulated recording traces (black lines) with the corresponding ground truth spike times (ts). Each trace contains one particular (u1,u2) pair as well as perturbations pt. indicated by green, blue and black markers, respectively).

described below). Up to five combined micro–macro-electrodes recorded single cell activities and LFPs using the INOMED ISIS MER-system 2.4 beta, INOMED Corp., Teningen, Germany. Four of the electrodes were distributed equally on a circle with 2 mm distance from the central electrode using a Ben Gun electrode guide tool. ED were already analyzed in a previous study (Florin et al., 2012) which was approved by the local ethic committee. For more detailed information about the recording setup and recording procedures see Florin et al. (2008); Reck et al. (2012); Gross et al. (2006); Michmizos and Nikita (2010).

The microelectrodes had an impedance of around 1 Mduring each recording session. The signal was amplified by a factor of 20 000, band-pass filtered from 220 to 3000 Hz, using a fixed hardware Bessel filter, and sampled at 25 kHz by a 12 bit A/D converter (± 0.2 V input range). We refrained from a second filtering to avoid further spike shape distortions (Quiroga, 2009). Recording started 6 mm above the previously planned target point. The extracellular multi-unit signals were recorded after moving the electrode closer to its target in 1 mm steps. Before entering the STN, the electrodes traverse the zona incerta characterized by a low background and absent spiking activity. Upon entering the STN background noise suddenly increases and three large amplitude discharge patterns are observed: tonic, irregular, and bursting activity (Rodriguez-Oroz et al., 2001). In patients with tremor, bursting activity may synchronize to tremor frequency. Upon leaving the STN, background activity suddenly drops and spiking activity either significantly reduces or becomes highly regular.

A total of 38 STN recoding traces from six awake PD patients at rest with one to four simultaneous microelectrode trajectories in different recording depths were analyzed. Some example sortings are shown in Fig. 1.

The inclusion criteria for a data trace were a) a minimum length of 20 s, b) no drifts in background activity, c) spiking activity in the STN (based on visual inspection), and d) not exceeding the dynamic range of the A/D converter. The longest stable segment of a given trace was selected for further analysis; the first 2 s of each recording after electrode movement were discarded.

Artificial data generation

A rigorous way to compare spike sorting methods is to test them on data sets with known ground truth, i.e., we know which of the spikes originate from which neuronal units. To this end, we generate AD by first selecting the two most distinct average spike waveforms from one ED trace. To enhance their differences, the larger one is multiplied with a factor of 1.1 so that it exhibits more pronounced peak amplitudes than given in the ED.[3] We call the waveforms w1 and w2. We then linearly combine w1 and w2 to obtain spike pairs (u1,u2) whose similarity can be varied parametrically:

$$u1 = \lambda \cdot w1 + (1-\lambda) \cdot w2 \qquad (1)$$

$$u2 = (1-\lambda) \cdot w1 + \lambda \cdot w2 \qquad \text{with } \lambda \in [0.5, 1]$$

Thus, by varying λ we create data sets with different degrees of similarity of the spike pairs (u1,u2), see Fig. 2A.

For λ = 1 we obtain u1 = w1 and u2 = w2 with u1 and u2 being most different. For λ = 0.5, u1 and u2 are identical. We generate four AD sets, each with one spike pair (u1, u2) obtained for a certain value of λ with λ = 1,0.8,0.7,0.6. The corresponding data sets are called setI (λ = 1, most distinct pair), setII, setIII and setIV (λ = 0.6, most similar pair). The hypothesis behind this choice is that it should be easier to distinguish distinct spikes than similar spikes.

The spike pairs are then added to background noise (Fig. 2B and C). To obtain the noise as realistic as possible, we generate it from the ED, using concatenated recording intervals without any spikes. We reshuffle the phase of the

[3] The smaller one is kept as it is

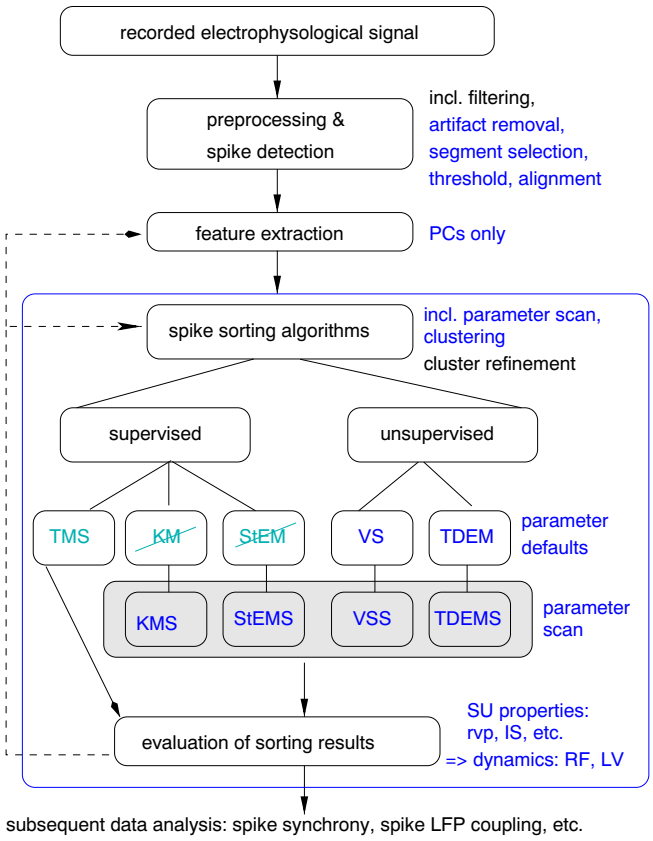

Fig. 3. Spike sorting workflow. Schematic overview of the different steps involved in a general spike sorting workflow: preprocessing, feature extraction, clustering, and cluster evaluation. Blue frame and text indicate the focus of this study. Different clustering methods are separated into originally supervised (cyan) and unsupervised algorithms (blue). Algorithms that are crossed out were only used in combination with a parameter scan (gray background) or with default values to enable unsupervised clustering. Dotted lines indicate possible feedback that can be used for an iterative improvement.

original noise so that the power spectrum is kept constant. The respective pairs of spikes (u1,u2) are added to the noise, each at a rate of 14 Hz as estimated from the ED, assuming a Poisson distribution. Each of the four generated data sets has a length of 40 s, sampled at 25 kHz resulting in approximately 500 spikes per unit. Refractory period violations (rpv) are corrected for by shifting the corresponding spikes of each single unit in time (the second one is shifted forward by 1 to 50 time stamps depending on the closeness of the two spikes) until no more rpv are found.

Inspired by the difficulties that occur during sorting the ED, we additionally include the following challenges. We inject overlapping spikes from different single units by inserting u1 spikes 10 to 22 time stamps (randomly chosen) after some randomly chosen u2 spikes (ca. 2.5% of the total number of spikes per trace as estimated from the ED) and vice versa. We then again correct for rpv. Moreover, we add in total ca. 100 so-called perturbation signals to each trace. These represent artifacts, e.g., noise originating from electrical equipment which may resemble spikes (Horton et al., 2007). Perturbations are given by 8 sinusoidal functions (black lines in Fig. 2B). Each perturbation consists of one cycle of the following frequencies f = 1,0.75,0.5,0.25 with respect to T = 1.52 ms spike length, using a positive amplitude of either the peak amplitude of u2 or half of it. The negative peak amplitude is fixed to the minimum of the spike generated with λ = 0.5. The corresponding insertion times are again Poisson distributed with a firing rate of 3.3 Hz as estimated from the ED. The aim is to investigate how different sorting methods deal with such perturbations. Ideally, perturbation signals should be left as unsorted events.

We generate 10 realizations for each of the four AD data sets. The noise is identical for the four data sets within one realization (setI to setIV) but changes between the 10 realizations. After inserting potential spike events, i.e., u1, u2 and perturbations, into the noise, some threshold crossings vanish while some new crossings without a corresponding AD event emerge, e.g., due to possible overlaps between u1, u2 and perturbations. Therefore, the spike times of the ground truth are obtained as follows: (1) Calculation of the spike detection threshold, i.e., mean signal minus four times the standard deviation (SD) of the complete trace, identical for all sorting methods and identical to the procedure used for the ED. (2) Insertion of u1 spikes into the pure noise and detection of threshold crossings. (3) Repetition of step (2) for u2 and perturbation signals, respectively. When comparing the spike times in ground truth and sorting results we only consider threshold crossings that occurred after insertion and that have a corresponding AD event time stamp. We allow for deviations up to ±0.5 ms. AD were generated using Python2.7.

Spike detection and spike sorting

ED and AD are separated into single units using spike sorting algorithms implemented in the 'Plexon Offline Sorter' OFS (Offline SorterTM, Plexon Inc., Dallas, TX, USA). Fig. 3 shows a general spike sorting workflow composed of four principle steps: preprocessing including spike detection, feature selection, feature based clustering, and finally the evaluation of the resulting single units. The blue frame and text indicate the focus of our comparative approach.

During a preprocessing step, artifacts (e.g. non-physiological events that may resemble spikes or some of the perturbations in the AD) were identified by visual inspection and removed. The spike detection threshold was set to minus four SD of the background noise[4] (Mrakic-Sposta et al., 2008). After detection, 360 s before and 1160 s after threshold crossing were extracted, resulting in a total spike length of 1520 s (38 time stamps). The spikes are aligned at the point of threshold crossing (cf. Fig. 2B).

Several features of the waveforms such as peak and valley amplitude, peak-valley distance, energy of the signal, and PCs were extracted. Only for the supervised 'manual'

[4] Exceptionally we also used 4.5 SD, depending on the individual signal-to-noise-ratio of the ED spikes.

sorting method TMS (see below), all extracted features were used to visually identify the templates and thus the number of clusters, while the clustering itself uses the complete waveforms. For all other algorithms, clustering is solely based on the first two (2D) or three (3D) principal components. We apply the sorting algorithms TMS, KM, VS, StEM, and TDEM (the methods are described in the following subsections) to both AD and ED.

VS and TDEM automatically determine the number of resulting clusters (unsupervised clustering) but contain method-specific parameters which were set to default values (see Table 2).

In addition, these methods were applied in combination with a parameter scan which optimizes the method-specific parameters (called VSS, TDEMS if used with a scan). During such a scan, a spike sorting algorithm runs repetitively for a wide range of parameter values (varied by step size Δ) to identify and select the run that yields the best sorting quality based on cluster quality metrics, e.g., distances in feature space. TMS, KM, and StEM require user intervention. To enable unsupervised clustering, KM and StEM are only applied in combination with a parameter scan (KMS, StEMS) which automatically computes the appropriate number of clusters. Table 2 and Fig. 3 list the methods that are used with a scan, as well as the corresponding parameter ranges. The idea behind performing a scan and solely using PC features is to obtain user-independent results that can easily be reproduced and compared.

The final step of the spike sorting workflow is the evaluation of the resulting single units, e.g., in terms of separation quality and refractory period violations. Some spike sorting methods (e.g., Harris et al., 2000; Barnett et al., 2016) include a loop that enables cluster refinement, mergence (Yger et al., 2018) or separation in case of ill-defined single units.

In total, 13 different sorting approaches were applied to each data trace: Template Matching with scan (TMS), Valley Seeking with scan (VSS2D and VSS3D), Valley Seeking with default value (VS2D and VS3D), K-Means with scan (KMS2D and KMS3D), standard Expectation Maximization with scan (StDEMS2D and StDEMS3D), t-distribution Expectation Maximization with scan (TDEMSS2D and TDEMSS3D), and t-distribution Expectation Maximization with default value (TDEM2D and TDEM3D), cf. Fig. 3. We apply each unsupervised method using the first two (2D) or three (3D) PCs, enabling a comparison of the corresponding performances. Similarly, we investigate the effect of using a parameter scan compared to using the default value (cf. Table 2).

Table 2. Methods used with a scan and their tunable parameters. The last column indicates the range of tested values, the step size **Δ**, and the default value d if used without scan.

Method	Scanning parameter	Scanning range
KMS	number of single units	1–7 single units
StEMS	number of single units	1–7 single units
TDEM(S)	degree of freedom multiplier DOF	10 to 30, Δ=5, d = 10
VS(S)	Parzen multiplier PM	0.5 to 1.5, Δ=0.2, d = 1

After spike sorting, each threshold crossing event was either labeled as sorted into a cluster or left unsorted if no clear assignment could be made. In the following subsections, we give more details on the spike sorting algorithms used in our study.

Template Matching sorting (TMS)

TMS is a supervised clustering algorithm, the number of clusters has to be predefined by the user. Based on various features the user selects one waveform as template for each cluster. Then, the algorithm calculates the root-mean-square differences D_w for all waveforms w to these templates t: $D_w = \sqrt{1/N \sum_{i=1}^{N} (w_i - t_i)^2}$, where N is the number of time stamps per waveform. TMS identifies the template with minimum difference D_w for each single waveform. If the minimum is smaller than a user defined value for the allowed variability, the particular waveform will be assigned to the cluster defined by this template.

K-Means clustering as Scan (KMS)

The K-Means algorithm requires the user to select a predefined number of clusters and the corresponding cluster centers which are here provided by the scanning algorithm. First, each sample point, i.e., each waveform in PC feature space is assigned to the nearest cluster center, based on Euclidean distances. Second, the cluster centers are recalculated using the center-of-gravity method (Gregory et al., 2007; Dai et al., 2008). Steps one and two are repeated until convergence is reached, i.e., cluster centers are stable. Finally, outliers are removed: Based on mean (μ) and SD of the distances of all sample points from their cluster center, a sample point is removed if it exceeds the outlier threshold, set to μ+2·SD. It is then left as unsorted event.

Valley seeking (VS)

The VS algorithm is based on an iterative non-parametric density estimation (Fukunaga, 1990; Zhang et al., 2007). To subdivide the sample points (i.e., spikes) into exclusive clusters, the algorithm estimates their density in PC feature space using the Parzen approach (Fukunaga, 1990), which estimates the appropriate kernel and its width for the best separation. VS calculates the number of neighbors of each sample point in a fixed local volume and determines the valleys in the density landscape. The critical radius R of the fixed volume is defined as $R = 0.25 \cdot \sigma \cdot PM$, where σ is the SD of the distances of all samples to the overall center point, and PM is the Parzen multiplier, a user-defined parameter. A sample point becomes a seed point of a cluster if its number of neighbors exceeds a threshold. Then, initial clusters are formed by the seed points with the most neighbors. An iterative process classifies still unassigned sample points or leaves them unsorted. We run VS both with the PM default value, and using the scanning algorithm for PM (VSS).

Expectation maximization algorithms (EM)

The standard EM (StEM) algorithm is an iterative, parametric approach that assumes that several Gaussian distributions underlie the distribution of sample points (i.e., spikes). The algorithm requires the user to select the number of Gaussians to be fitted and to define the initial cluster centers (Fukunaga, 1990; Sahani, 1999). To enable unsupervised clustering these are provided by the scanning algorithm. The algorithm starts by running the K-Means algorithm for the first assignment of sample points from which the initial Gaussian parameters are estimated. An iterative process optimizes these parameters until convergence of the Gaussian distributions to stable values. Each iteration consists of an expectation (E)-step that calculates the likelihood for each sample point to belong to each Gaussian, and a maximization (M)-step that maximizes the expected likelihood by optimizing the parameters (Fukunaga, 1990; Sahani, 1999).

The t-distribution EM-algorithm (TDEM) differs from the StEM by fitting wide-tailed t-distributions instead of Gaussians. It has been shown that t-distributions yield a better fit to the underlying statistics of the waveform samples (Shoham et al., 2003). TDEM directly provides unsupervised clustering by starting with a large number of clusters and iteratively optimizing the likelihood function (assignment of samples to clusters) (Shoham et al., 2003). The shape of the t-distribution is determined by the degree of freedom (DOF) multiplier which depends on the sample size and controls the convergence properties (Figueiredo and Jain, 2002; Shoham et al., 2003). We run TDEM both with the DOF default value, and using the automatic scanning algorithm for DOF. In the Plexon implementation of these EM algorithms no events are left unsorted.

Evaluation of spike sorting results

Sorting results are characterized by the number of detected single units and the number of unsorted events. The resulting means and SDs per data set are calculated by averaging over the 38 ED recording traces and the 10 realizations for each AD set, respectively. In the following we detail the evaluation of the corresponding results.

Comparison with ground truth data

To evaluate the accuracy of the clustering algorithms, the resulting single units were compared with a given ground truth. To quantify accordance with and deviations from the ground truth, we calculate the following numbers (cf. Fig. 5D and Fig. 8D):

- **TP** true positives, i.e., correctly assigned spikes: a waveform was given as element of a certain single unit and was sorted into this single unit.
- **FP** false positives (sorted), i.e., wrongly assigned (misclassified) spikes: a waveform was given as element of a certain single unit but was sorted into another single unit.
- **FN** false negatives, i.e., spikes wrongly left unsorted: a waveform was given as element of a certain single unit but was left unsorted.
- **FPp** false positives (unsorted), i.e., wrongly assigned (misclassified) perturbations: a perturbation signal was classified as element of a certain single unit.
- **TN** true negatives, i.e., correctly assigned perturbations: a perturbation signal was left unsorted.

Thus, a 100 % correct classification contains only TP and TN. For each data set sorted by a certain method, we count the corresponding hits (TP, TN) and misses (FP, FPp, FN) and normalize by the number of all events (spikes and perturbations) that are present in both ground truth and the corresponding sorting outcome. For each single unit of the ground truth, we check which unit of the spike sorting outcome contains the most hits and then take this unit as correct. Therefore, we always find TP > FP. Based on these numbers we calculate the following measures. The sensitivity describes how many spikes out of all spike events are correctly assigned: sensitivity = TP/(TP + FP + FN) while the specificity describes how many of the perturbations are correctly left unsorted: specificity = TN/(TN + FPp).

These analyses were performed using MATLAB (Mathworks Inc., Natick, USA). Differences in the general performance of the algorithms were evaluated by comparison to the ground truth values using the Wilcoxon rank sum test. Bonferroni's correction was applied to adjust the significance level for multiple comparisons. To contrast the 2D with the 3D version of a method, we used direct comparisons (Wilcoxon rank sum test without Bonferroni's correction), as well as for the comparison of running a method with a scan versus using the default parameter value.

Quality of spike sorting

We also assess the quality of our sorting results with the following evaluation measures: the percentage of refractory period violations (rpv), the isolation score (IS), and a measure characterizing the internal cluster distance (Di) (Fee et al., 1996b; Joshua et al., 2007; Hill et al., 2011; Einevoll et al., 2012). The amount of rpv indicates the degree of multi-unit contamination in a given single unit. We determine the percentage of inter-spike intervals (ISIs) shorter than 2 ms. The value of 2 ms lies well inside the range of the typically assumed ISI limits of 1.5 ms up to 3 ms (Shinomoto et al., 2003; Moran et al., 2008; Lourens et al., 2013; Kelley et al., 2018). The IS compares the waveforms within one single unit to all other potential spikes in the recording trace based on the normalized and scaled Euclidean distances of their time courses (Joshua et al., 2007). It provides an estimate of how well a single unit is separated from all other potential spikes outside its cluster[5]: IS = 1 means well separated while IS = 0 indicates overlapping clusters. It thus requires the existence of potential spikes outside a given cluster. Since EM methods do not account for unsorted events, we only calculate the IS when there are at least two single units in a given trace. We additionally consider the internal cluster distance Di because it can also be calculated if there is only one unit.

[5] AD were used to calibrate the IS scaling parameter to five.

Table 3. Evaluation measures for Fig. 1. Isolations scores IS and percentage of refractory period violations (rpv) calculated for the clusters shown in Fig. 1. The spike sorting methods TMS, KMS, TDEMS, and VSS yield a variable number of clusters (one to five). If there is only one single unit (TDEMS in panel B) the IS cannot be calculated.

	Unit	1	2	3	4	5
Method	Panel	refractory period violations (%); IS				
TMS	A	0.0; 0.9	0.0; 1.0			
	B	0.2; 0.7	0.2; 0.6	0.3; 0.5		
KMS	A	0.5, 1.0	0.0; 1.0			
	B	2.2; 0.6	0.2; 0.8			
TDEMS	A	0.5; 1.0	0.3; 1.0			
	B	3.7; /				
VSS	A	0.0; 0.9	0.0; 0.9			
	B	0.0; 0.5	0.0; 0.4	0.0; 0.4	0.3; 0.5	0.0; 0.3

This measure uses the first three PCs of each waveform. For each single unit, we calculate the mean waveform and its mean Euclidean distance (in reduced PC space) to all other spikes inside this cluster (normalized by the maximum value of the ED or AD data set). To provide an idea of the meaning of these measures the rvp and IS values of the clusters shown in Fig. 1 are listed in Table 3.

For a consistent scaling behavior of the latter two quality measures we consider 1-Di so that high IS and high 1-Di values indicate well defined clusters.

Firing properties

To investigate the differences in the dynamical properties of the single units we calculate the mean firing rate and the local coefficient of variation LV (Shinomoto et al., 2003) in the ED. The LV characterizes the firing regularity of a single unit:

$$LV = 1/(n-1)\sum_{i=1}^{n-1} 3(T_i - T_{i+1})^2/(T_i + T_{i+1})^2, \quad (2)$$

where T_i is the duration of the ith ISI and n the number of ISIs. LV values enable the following classification (Shinomoto et al., 2003): regular spiking for LV∈[0,0.5], irregular for LV∈]0.5,1], and bursty spiking for LV > 1. For this analysis, only single units with more than 80 spikes and less than 1% rpvs were taken into account to avoid outliers.

RESULTS

We first present the evaluation of the results obtained by applying the 13 sorting algorithms to the STN recordings (ED), followed by an investigation of the impact of the sortings on the dynamical properties of the resulting single units. The ED sorting evaluation leaves us in doubt about the best method. Therefore, we then evaluate the results of applying the identical methods to the AD which allows for an objective ground truth comparison. This procedure enables us to finally identify the best sorting methods.

Spike sorting of experimental data

We aim at sorting the ED under the additional constraint of identifying an *unsupervised* sorting algorithm that enables a fast, reliable and reproducible extraction of single units. Our criteria for a successful sorting are: (1) all true spikes are detected and (2) artifacts and strongly distorted spikes are not extracted but left unsorted. For a quantitative evaluation, we first sort the data using TMS, a manual sorting method that puts the user in complete control. According to our criteria it was performed with precise visual inspection aiming for clearly separated single units that are free of artifacts and wrongly classified spikes. During this preprocessing procedure, we did not observe a firing rate dependent spike shape modulation in our STN traces. Hence, we neither performed a quantitative test to check for spike amplitude modulations, nor did we single out bursting activity for specific analyses. In search for an automatic sorting algorithm, we apply the following unsupervised methods offered by Plexon: VS and TDEM (both applied with parameter scan and default value, respectively, in 2D and 3D, respectively), as well as KMS and StEMS (both only applied with scan, in 2D and 3D, respectively). Some exemplary results are shown in Fig. 1. Here, we assume that the TMS sorting represents the ground truth, because TMS is a widely used method (Raz et al., 2000; Levy et al., 2002; Rutishauser et al., 2006; Steigerwald et al., 2008), subjectively often perceived as the best sorting.

For a quantitative sorting analysis we evaluate the number of detected single units (Fig. 4A), the percentage of unsorted events (Fig. 4B), and the percentage of refractory period violations rpv (Fig. 4C). The number of detected single units is highly variable, depending on the sorting method. TMS and KMS3D detect on average two single units, TDEM(S) detect on average significantly less units whereas VSS2D and VSS3D yield significantly more units than TMS. EM methods do not account for unsorted events, they do not leave any spike unsorted. KMS methods yield the least unsorted events, followed by VS(S)2D while VS(S)3D and TMS show the highest percentage of unsorted events. TMS, all VS methods and KMS2D and KMS3D result in less than 1.5% rpv. Methods that do not account for unsorted events yield more potential spikes per single unit which results in a higher probability of rpv occurrences. In the literature, the percentage of tolerated rpv ranges from 0.5% up to 2.5% (Moran et al., 2008; Lourens et al., 2013; Yang et al., 2014). We consider a single unit to be clean if it has less than 1% rpvs.

We now compare the assignments of the different sortings to the ground truth given by TMS using the terminology introduced in Methods: TP, TN, FP, FPp and FN rates (Fig. 5D). Since TMS aims at ‘clean’ single units it results in a high TN rate of 39% and thus only 61% TP, see Fig. 5A. All other methods leave less events unsorted, resulting in lower TN (black) and accordingly higher FPp (dark gray) rates. They show a similar amount of misclassified spikes (FP, light blue) but clearly differ in terms of their FN (light gray), FPp and TN rates. EM methods, e.g., yield no TN but only FPp, since they do not allow for unsorted

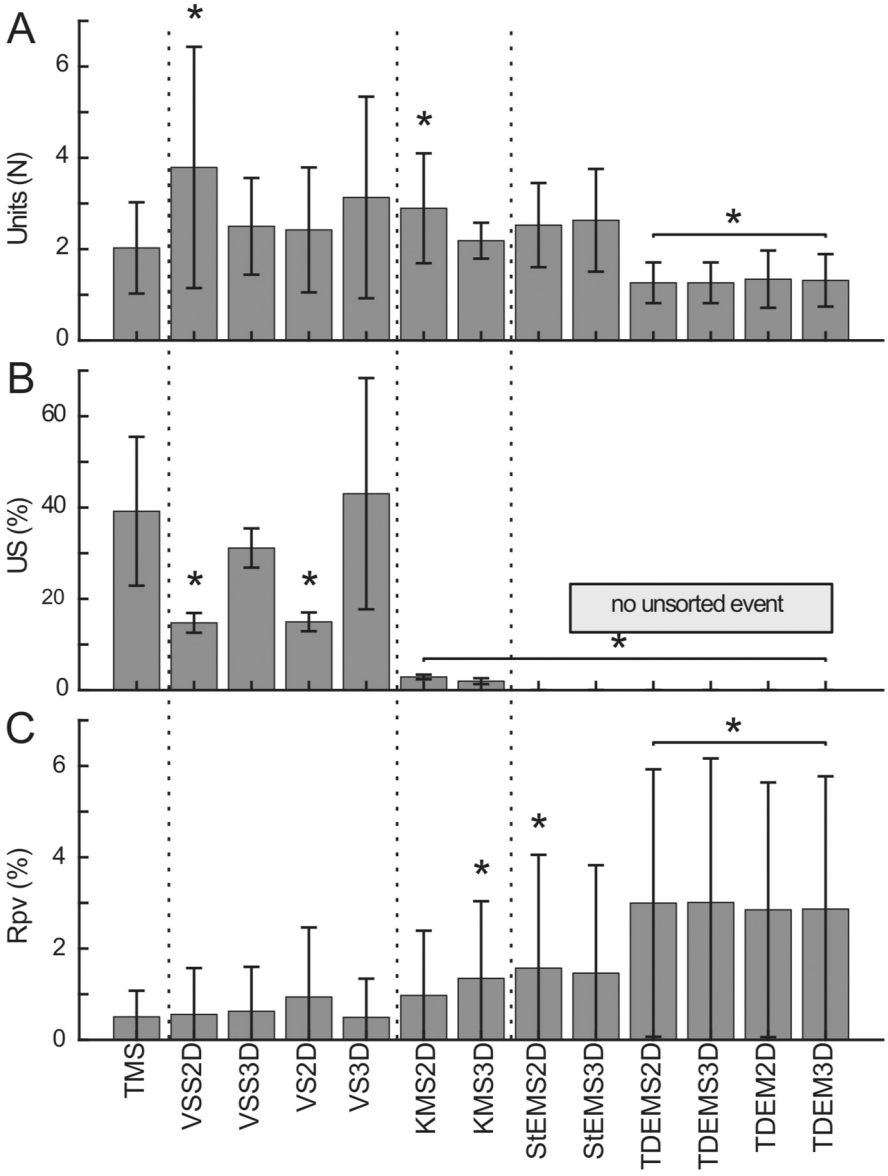

Fig. 4. Evaluation of ED sorting results. (A, B, C) Bar plots of the average number of detected single units (A), the percentage of unsorted events US (B) and the percentage of refractory period violations rpv (C), in dependence of the sorting method. Shown are mean ± SD of the 38 recording traces. Stars indicate a significant difference (P <.0042 after Bonferroni correction) to the TMS results.

events, hence the high amount of rvp (cf. Fig. 4C). Compared to TMS (the assumed ground truth), TP rates are reduced for all other methods, the most for VS methods due to the high FN rates. KMS yield relatively high TP rates but very low TN rates because only a very few events are left unsorted (cf. Fig. 4B).

The sensitivity (i.e., percentage of TP relative to the total number of true spikes in TMS) and specificity (percentage of TN relative to the number of unsorted events in TMS) measures in Fig. 5B1 and Fig. 5B2 summarize these results. We aim at both a high sensitivity (i.e., correctly classified spikes) and a high specificity (events correctly left unsorted). In total, the sensitivity varies between 44% and 80% (Fig. 5B1), and the specificity between 0% and 64% (Fig. 5B2). All EM methods result in a high sensitivity, but zero specificity, because they do not account for unsorted events. KMS methods also yield a high sensitivity, but a low specificity. VS methods result in a high specificity combined with a rather low sensitivity. Combining these two measures, VSS3D seems to provide the best result, since it shows the highest sensitivity of all VS methods.

Fig. 5C1 and Fig. 5C2 asses the sorting quality from another perspective, independently of the assumed ground truth:

The isolation score (IS) and the internal cluster distance (Di) indicate how well the resulting single units are clustered (cf. *Fig* 1). For well separated clusters without artifact contamination, IS and 1-Di are close to one. The large vertical spread indicates a large variability for all methods, mostly due to the high variability in the number of single units detected by each method, cf. Fig. 4A. TMS yields a rather low IS although the low percentage of rpv indicates a successful sorting (cf. Fig. 4C). The high IS values for TDEM methods do not indicate well defined clusters due to a high amount of rpv (cf. Fig. 4C). They are simply a consequence of the fact that the IS can only be calculated when there is more than one single unit which is not often the case, cf. Fig. 4A. The Di measure considers all single units and indeed indicates poorly defined clusters. KMS methods yield relatively high IS and 1-Di values and a reasonable amount of rpv. VS methods show relatively high 1-Di values but comparably low IS scores. Together with the high FN rate (Fig. 5A) this indicates that many spikes are left unsorted.

We assessed the time consumption of all methods that account for unsorted events by applying them to five randomly chosen STN channels of 31.5 *s* ± 9.2 *s* length. The method with the highest time consumption is clearly manual TMS which needs 269.8 *s* ± 76.2 *s*, including the time to asses the number of clusters, to define the templates, visual inspection of results in feature space, etc. VS2D and VS3D need 0.5 *s* ± 0.1 *s* and 0.6 *s* ± 0.2 *s*, respectively, VSS2D and VSS3D need 2.5 *s* ± 0.6 *s* and 2.7 *s* ± 0.3 *s*, respectively, while KMS2D and KMS3D need 2.0 *s* ± 0.6 *s* and 2.3 *s* ± 0.5 *s*, respectively.

In summary, we find that VSS3D agrees best with the TMS results, suggesting that VSS3D is the best sorting method. However, a detailed comparison of the assignment of individual spikes indicates considerable

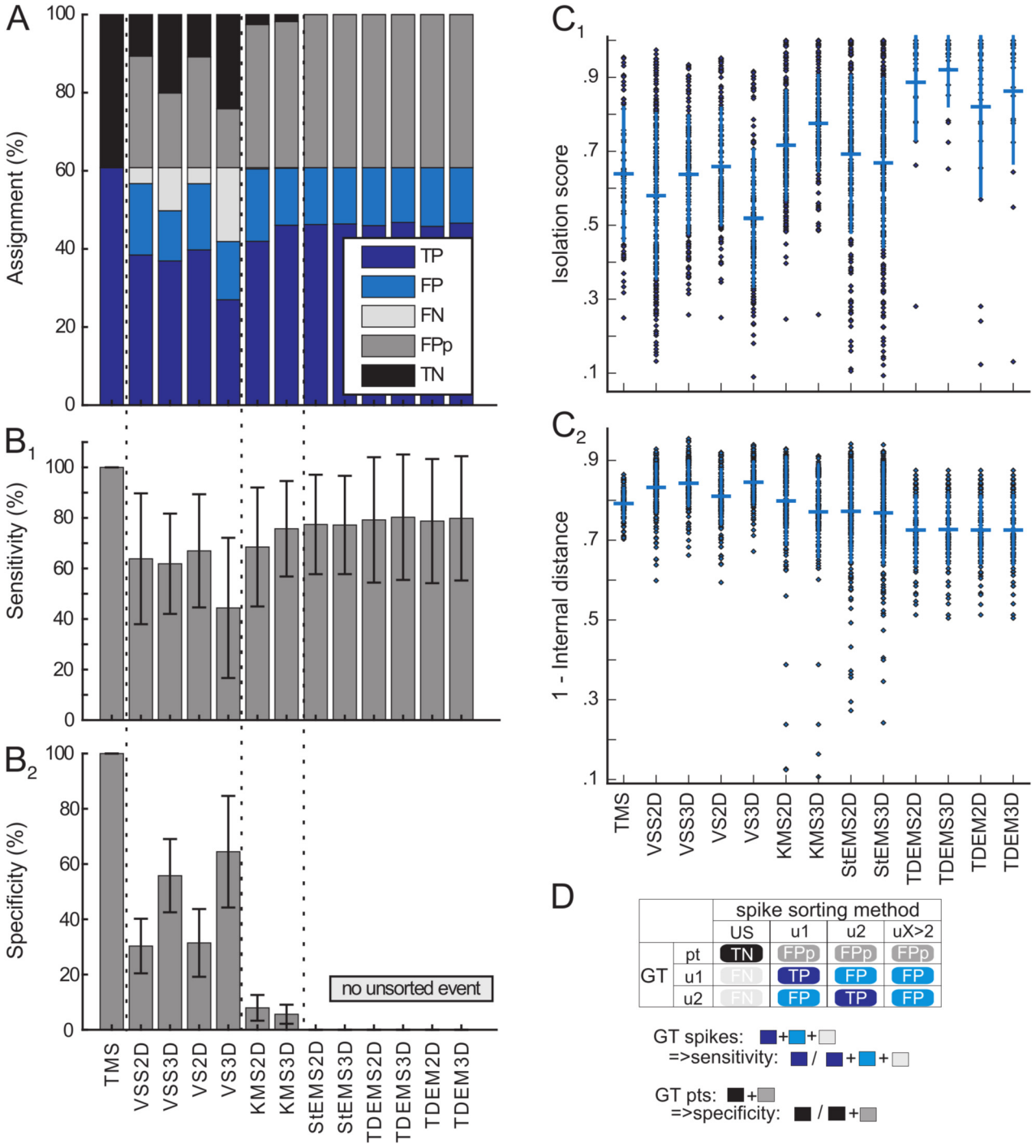

Fig. 5. Evaluation of ED sorting results with ground truth (TMS) comparison. (A) Stacked bar plot showing the percentage of correct and wrong assignments of ED spikes in dependence of the sorting methods using TMS as ground truth reference: TP indicate correctly classified spikes, FP misclassified spikes, FN spikes left unsorted, FPp indicate unsorted events of TMS that are classified as spikes, and TN indicates unsorted events of TMS that are also left unsorted by the other methods. (B1, B2) Sensitivity and specificity measure in dependence of the sorting methods (mean ± SD of all recording traces). (C1, C2) Cluster quality measures IS and 1-Di (Di values normalized to their maximum) applied to ED: each dot represents the value obtained for one single unit in the 38 recording traces. Horizontal lines indicate the average over all single units, vertical lines indicate the corresponding SD. (D) Summary of TP, TN, FP and FN notations and definition of sensitivity and specificity.

differences: The FPp, FP, and FN rates for VSS3D sum up to approximately 40%. Other issues are the low IS score for both TMS and VSS3D, the higher TP rate for KMS compared to VSS3D, as well as the fact that we might lose a lot of true spikes when using TMS or VSS3D due to 39% unsorted events. Moreover, all VS and KMS methods detect more single units compared to TMS. In the end, we are left with the suspicion that the subjective TMS sorting and thus VSS3D might, after all, not be the best methods to sort our data. We therefore apply all methods again to

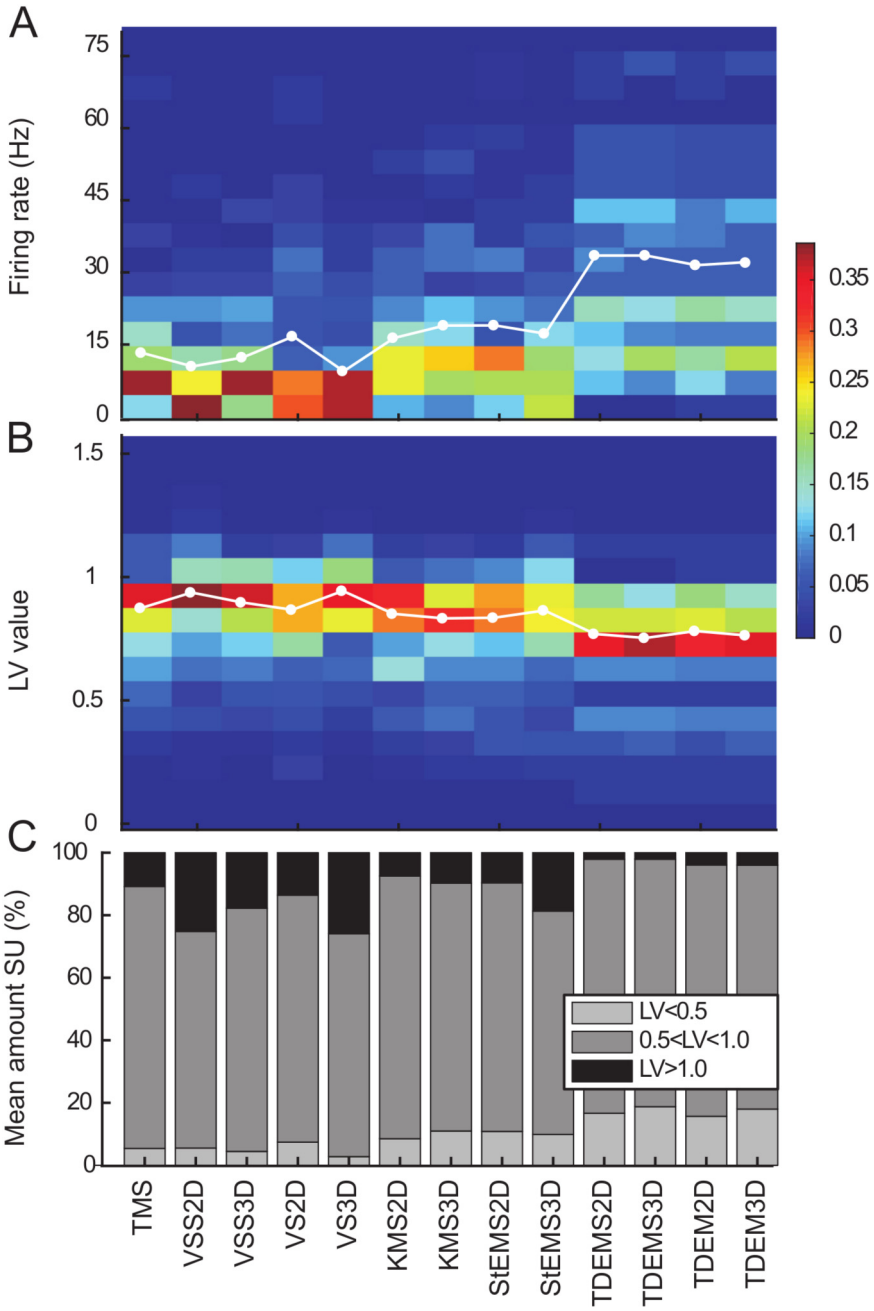

Fig. 6. Variability in the ED firing statistics. Differences in firing patterns characterized by (A) firing rate and (B) local coefficient of variation LV. The color code indicates the number of detected single units SU normalized to the total number of detected single units per method (binned and averaged over all traces). Mean firing rate and mean LV (averaged over all single units) are indicated by white dots. (C) Amount of single units with regular (LV < 0.5), irregular ($0.5 \leq LV \leq 1.0$) and bursty (LV > 1.0) firing patterns.

AD which provides an objectively given ground truth to compare with.

Impact of spike sorting methods on single unit firing properties

To characterize the differences in the firing patterns of all detected single units that result from using different sorting algorithms, we calculate the firing rate and the local coefficient of variation for each single unit. Fig. 6A and Fig. 6B show the corresponding distributions obtained by binning and averaging across all STN recording traces. Each entry is averaged over all single units identified in the ED. Fig. 6A shows clear discrepancies in the firing rate distributions. This is a consequence of the distinct number of single units obtained from the different methods, as well as of the amount of unsorted events (cf. Fig. 4). Single units obtained with TMS and VS show lower frequencies (maximally up to 30 Hz), while KMS and StEM yield single units with up to 40 Hz. Single units obtained with TDEM methods have the highest firing rates (60 Hz) since these methods detect mostly one single unit and leave no events unsorted.

Another characteristic property of spiking activity is the LV which quantifies the regularity in neuronal firing (Fig. 6B). We again observe method-dependent deviations, based on the variable number of single units: the lower the number of detected single units, the more regular are the subsequent spike trains. When classifying the single units according to their LV value in regular (LV < 0.5), irregular ($0.5 \leq LV \leq 1.0$), and bursty (LV > 1) firing neurons (Shinomoto et al., 2003; Steigerwald et al., 2008; Lourens et al., 2013) we find clear differences (6C): TDEM methods yield more regular and less bursty single units (less than 5%) whereas VS2D and VS3D result in less regular and more bursty single units (up to 25%).

Spike sorting of artificial data

The ED results left us undecided concerning the best sorting method. In need of an objective ground truth we now evaluate the results of sorting artificially generated data. For the AD we know the correct spike and perturbation assignments, the latter representing artifacts to be left unsorted. Please note that there are a few overlapping spikes from distinct single units. If their shapes are seriously distorted (large overlap) they are most likely classified as extra unsorted events.

Fig. 7 presents the first part of the sorting results obtained for AD sets with varying spike pair similarity. For each set we again evaluate the resulting number of single units (Fig. 7A), the percentage of unsorted events (Fig. 7B) and rpv (Fig. 7C). The ground truth is shown on the very left of the panels. For setI and setII (distinct spike pairs), the

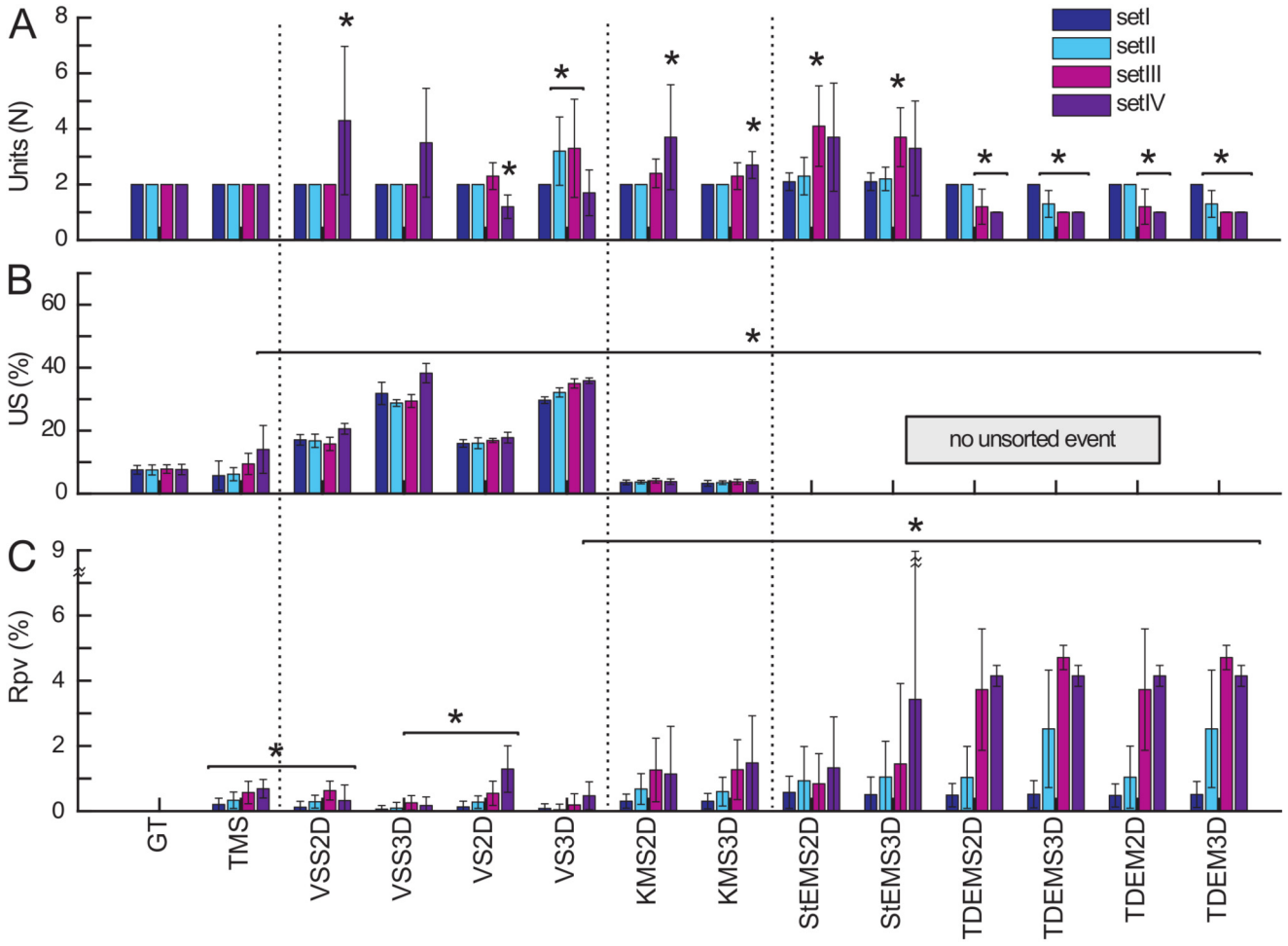

Fig. 7. Evaluation of AD sorting results. Bar plots of the sorting results in dependence of the sorting methods (color coded spike pair similarity in setI to setIV): (A) average number of detected single units SU, (B) percentage of events left unsorted, and (C) percentage of refractory period violations (rpv). Shown are mean ± SD of the 10 realizations, stars indicate a significant difference compared to the ground truth values (P <.039 after Bonferroni correction).

number of resulting single units is mostly quite similar and close to the ground truth. For setIII and setIV (similar spike pairs) all TDEM algorithms detect significantly fewer units, whereas StEMS methods find significantly more units, similar to KMS and VSS (Fig. 7A). These observations are similar to the corresponding ED results (cf. Fig. 4A).

EM algorithms do not account for unsorted events and KMS methods leave only a very few events unsorted while VS methods yield many more unsorted events than present in the ground truth (15% to 40% compared to 10%, see Fig. 7B). The percentage of unsorted events resulting from TMS is mostly close to the ground truth, only setIV yields more than 10% unsorted events due to a nearly impossible distinction between perturbations and spikes. The major difference to the ED results is the small amount of unsorted events in the ground truth: Here, unsorted events represent artifacts whereas the large amount of unsorted events in the ED are mostly spikes that were left unsorted because no clear assignment could be made.

Most methods induce a significant percentage of rpv (Fig. 7C). The percentage of rpv is larger the more similar the embedded spike pairs are, as it is more difficult to separate similar spikes which induces sorting errors. As observed for the ED, we find that methods that do not account for unsorted events result in a high percentage of rpv, e.g., TDEM methods with more than 3% rpv for setIII and setIV. In contrast, VS methods yield mostly less than 1% rpv.

Fig. 8A1 to A4 show the second part of the AD results: the TP, TN, FP, FPp and FN assignments made for the four sets. The 100% correct ground truth assignment consists of two parts: 90% TP, i.e., correctly classified spikes and 10% TN, i.e., perturbations that were correctly left unsorted.

Concerning the spike events in setI, most methods perform quite well, yielding a TP rate close to 90%. Only VS (S) methods leave 8% (2D) to 15% (3D) of spikes unsorted which results in a comparably high FN and low TP rate. However, as expected from the ED results, VS(S) methods also correctly leave most perturbations unsorted (TN close to 10%). KMS, StEMS, and TDEMS yield generally high FPp and low FN rates: many perturbations are wrongly classified as spikes and only a few or no spikes are left unsorted. FN, FPp and TN rates change only slightly with increasing spike pair similarity (setI to setIV) since perturbations are identical in all sets. The number of misclassified spikes, however, clearly increases with increasing spike

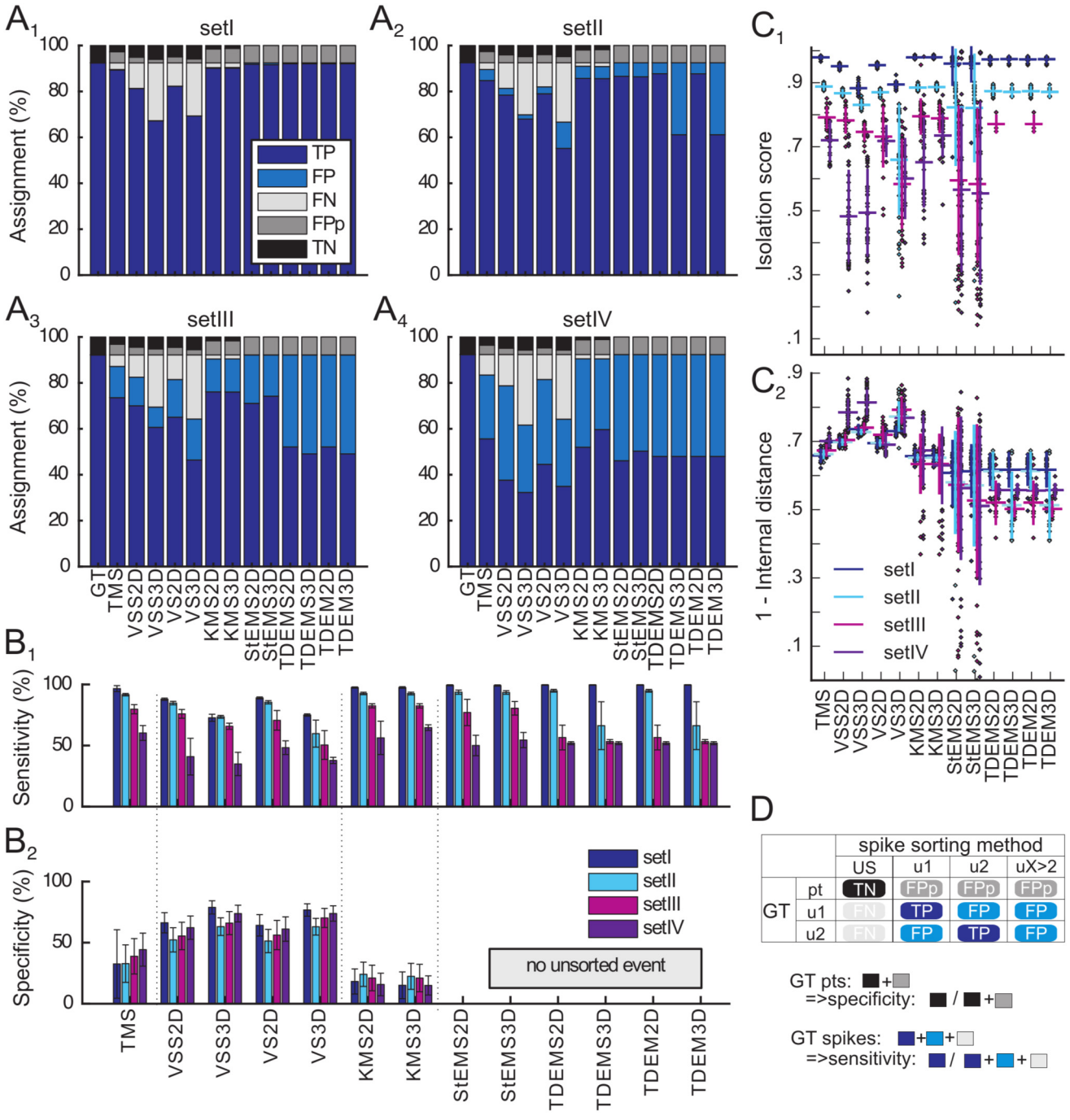

Fig. 8. Evaluation of AD sorting results with ground truth comparison. (A1-A4) Stacked bar plots showing the average percentage of correct and wrong assignments of spikes and perturbations in dependence of the sorting methods: TP indicates correctly classified spikes, FP misclassified spikes, FN spikes wrongly left unsorted, FPp indicates perturbations wrongly classified as spikes, and TN perturbations correctly left unsorted (cf. D). Stars indicate a significant difference to the ground truth values (P <.039 after Bonferroni correction). (B1, B2) Sensitivity and specificity measures in dependence of the sorting methods (mean ± SD of the 10 realizations). (C1, C2) Cluster quality measures IS and 1-Di (Di values normalized to their maximum) in dependence of the sorting methods: each dot represents the value obtained for one single unit. Horizontal lines indicate the average over all single units in 10 realizations, vertical lines indicate the corresponding SD. The color code in B1, B2, C1, and C2 represents the spike pair similarity. (D) Summary of TP, TN, FP and FN notations and definition of sensitivity and specificity.

pair similarity: In setII, TDEM(S)3D already show 30% FP due to collapsing the spikes from two single units into one single unit while all other methods yield 2% to 6% FP (Fig. 8A2). For setIII, all TDEM methods yield approximately 45% FP while most other methods result in fewer FP (5% to 15%) and correspondingly higher TP rates (60% to 80%). Only VS3D yields less than 50% TP due to the relatively high percentage of 20% FN.

Fig. 8C1 and Fig. 8C2 show that the cluster quality measures IS and Di mostly reflect the results obtained by the

above ground truth comparison: the more similar the spike pairs, the lower the TP rate and the average IS. This agreement holds only partially for the Di. If the number of identified single units is large, the resulting clusters are small and naturally have a small internal distance, e.g., the large 1-Di values for VSS in setIV (Fig. 8C2, cf. Fig. 7A). Thus, IS and Di have to be considered in relation to the number of single units. VS methods show relatively high 1-Di values but low TP rates, an effect of the high FN rates which bear less influence on the Di measure (Joshua et al., 2007). For the TDEM(S) methods applied to setII, however, the Di results match the TP rates better than the IS results.

We expected that using more PCs yields better results. However, the significant (P<.05) differences in the percentage of unsorted events and TP between VS(S)2D and VS(S)3D indicate the opposite: the 2D results are closer to the ground truth. Still, VS(S)3D yield significantly less rpv compared to VS(S)2D, but this is simply the consequence of leaving many events unsorted. Similarly, some of the TDEM(S)2D results (number of units and TP rate for setII) are significantly closer to the ground truth than the TDEM(S)3D results. Therefore we conclude, that VS and TDEM work better in 2D as compared to 3D feature space.

The differences between the results obtained with and without automatic scan are inconsistent and only pertain to VS methods. For example, VSS3D versus VS3D yields mostly significantly (P<.05) different values for the number of single units, unsorted events, and TP where the results obtained with scan are closer to the ground truth for the number of single units and unsorted events but without scan, the TP rates are closer to the ground truth. Thus, we see no advantage in applying an automatic parameter scan.

Fig. 8B1 and Fig. 8B2 summarize our findings. The sensitivity (normalized TP rate) clearly decreases with increasing spike similarity, independently of the sorting method. The more similar a spike pair is, the harder is the task to distinguish the spikes and to sort them into different units. For setIV, all sensitivity values are close to 50% indicating that the sorting task is so difficult that the success rates are bound to be close to chance level. However, there is no clear dependency of the specificity (normalized TN rate) on the task difficulty. Since EM methods do not account for unsorted events, their specificity is zero. As expected from the ED results, VS methods show a high specificity but their sensitivity is rather low. In contrast, KMS and TMS show again, as observed for the ED, a low specificity while their sensitivity is relatively high. For the AD, we conclude that KMS and VS(S)2D yield the best compromise between high sensitivity, i.e., many correctly classified spikes, and high specificity, i.e., many identified perturbations. Hence, the doubts about VSS3D being the best sorting method for the ED are justified. Fig. 9 shows the so-called success rate, i.e., the sum of TP and TN rates (filled circles) for AD sets II and III with 90% spikes and 10% perturbations. It shows that TMS and KMS are the most successful methods, followed by VS2D. The open circles are an estimate obtained via re-normalization with changed proportions for spikes (50%) and perturbations (50%).

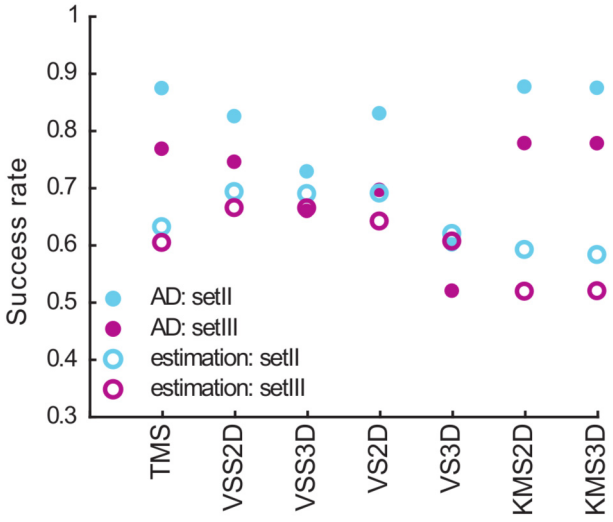

Fig. 9. AD sorting results in dependence of the amount of perturbations. Shown are the sum of TP and TN rate of the AD setII and setIII for the sorting methods that account for unsorted events (ground truth is at one). Filled circles indicate the proportion given in the AD, i.e., 90% spikes and 10% perturbations. Open circles indicate an estimate calculated by re-normalization assuming 50% spikes and 50% perturbations. The color code represents the spike pair similarity.

In this case VS methods show a higher success rate than KMS and TMS. Thus, the best method to sort the data depends on the amount of perturbations (i.e., artifacts) which are to be left unsorted.

DISCUSSION

The classification of multi-unit activity into single units is an important prerequisite for many types of data analysis, e.g., neuronal correlations, spike-LFP phase coupling, or tuning properties of single cells. The sorting evaluation procedure described in this study is generally applicable. We provide a comparative analysis that depicts and characterizes the differences in the results of a selected set of sorting algorithms, applied to ED and AD with known ground truth, respectively. Comparing the results of the ED to the four AD sets we find that the task difficulty in the ED is most similar to setIII of the AD, i.e., a hard task due to similar spike shapes.

We evaluate sorting methods provided by the 'Plexon Offline Sorter', a frequently used software package (Shinomoto et al., 2003; Moran et al., 2008; Schrock et al., 2009; Shimamoto et al., 2013; Yang et al., 2014; Kelley et al., 2018; Lipski et al., 2018). Aiming for an objective comparison without any user intervention, we focus on algorithms that either run with a given default parameter value or in combination with a parameter scan, always using PC as features. We additionally use supervised TMS in order to contrast our results with this widely-used (Raz et al., 2000; Levy et al., 2002; Rutishauser et al., 2006; Steigerwald et al., 2008) method. The user intervention in such supervised methods is time consuming, it inherently includes a human bias (Wood et al., 2004) and it typically requires a parameter and/or a feature selection optimization (Wild et al., 2012).

Highly variable sorting results for different methods

In agreement with Brown et al. (2004); Wild et al. (2012); Knieling et al. (2016) we show that the results obtained by using different sorting methods differ significantly, in both ED and AD. There are deviations in the number of detected single units, in the percentage of unsorted events and rpv, as well as differences in the cluster quality measures. The IS is typically used to select well isolated single units, e.g., by rejecting clusters with IS <0.7 (Joshua et al., 2007; Lourens et al., 2013; Deffains et al., 2014). The percentage of tolerated rpv is typically assumed to be 0.5% up to 2.5% (Moran et al., 2008; Lourens et al., 2013; Yang et al., 2014) for refractory periods assumed to be 1 ms up to 4 ms (Bar-Gad et al., 2001a; Moran et al., 2008; Eden et al., 2012; Lourens et al., 2013; Shimamoto et al., 2013; Kelley et al., 2018). Given this wide range of values, Knieling et al. (2016) suggests that 1 ms could be the absolute refractory period while larger values pertain to the relative refractory period. The application of these measures to the AD verifies an increased occurrence of well isolated single units if the corresponding waveforms are distinct and thus easy to separate.

Most extracellular recordings (in particular STN data, see Introduction) contain perturbations, e.g., movement or speech artifacts, and distorted spikes. Therefore, it is a clear disadvantage of EM methods that they do not leave any event unsorted. Even though such perturbations are often removed during a preprocessing procedure (cf. Methods), a considerable percentage is typically not identified. For example, approximately 8 of 10% perturbations survived the preprocessing of our AD. Consequently, the resulting single units of all EM methods are contaminated, resulting in high FPp rates and a high amount of rvp. Among the EM methods, StEM yields the highest sensitivity and the fewest rpv. Hill et all. (Hill et al., 2007; Hill et al., 2011) discuss that the assumption of Gaussian distributions in StEM may be inappropriate for spike clusters due to spike shapes varying with time. The latter can be caused by bursting activity which is a prominent feature in STN recordings (Hutchison et al., 1998; Beurrier et al., 1999; Chibirova et al., 2005; Steigerwald et al., 2008). Still, StEM works comparably well for our data, possibly due to the short recoding time and relatively constant spike shapes.

VS algorithms yield the most specific sorting results, they leave nearly all perturbations unsorted. Yet, all VS methods also leave a considerable amount of spikes unsorted which decreases their sensitivity. For the AD, only VS(S)2D methods provide a good compromise between specificity and sensitivity. A previous study (Kretzberg et al., 2009) details that TDEM performs better than VS in clustering artificial data adapted to resemble extracellular recordings from a turtle's retina. Our AD, however, explicitly contains perturbations. In such a complex case, as typical for STN data (Lewicki, 1998), the non-parametric approach taken in the VS(S)2D might provide an advantage because the valleys separating the single units do not have to obey a specific parametric form (Fukunaga, 1990; Hill et al., 2011).

KMS is the most sensitive algorithm, only a few spikes are left unsorted and the amount of misclassifications is acceptable. Yet, it detects only a very few perturbation and thus has a low specificity. There is no significant difference between KMS2D and KMS3D. At first sight, one expects that more information (i.e., 3D) yields a better performance but VS and TDEM perform better in 2D feature space than in 3D. The additional dimension may capture the variability in the background noise (Lewicki, 1998; Bishop, 2006) and thus lead to misclassifications.

Another important point for selecting an appropriate sorting method is the type of analysis that the user aims to perform with the resulting single units. Missed spikes (FN), for example, reduce the significance of spike synchrony stronger than misclassified spikes (FP) (Pazienti and Grün, 2006). Thus, for the analysis of neuronal correlation in STN recordings (Weinberger et al., 2006; Moran et al., 2008) KMS is a better choice than VS. Another example are tuning curves, i.e., the distributions of neuronal firing rates with respect to a movement (Georgopoulos et al., 1982) or stimulus (Hubel and Wiesel, 1959) direction. In this case, misclassified events (spikes, perturbations) can induce incorrect multimodal distributions while missed spikes lead to an underestimation of the true firing rates (Hill et al., 2011).

Dynamical properties of single units and their relation to spike sorting

Typically, the average firing rates measured in the STN of Parkinson patients are reported to range from 25 Hz up to 50 Hz (Benazzouz et al., 2002; Steigerwald et al., 2008; Remple et al., 2011; Lourens et al., 2013; Deffains et al., 2014). We observe rates ranging from 14 Hz up to 39 Hz, purely depending on the sorting algorithm. Thus, we find lower rates than reported in the literature which can have several reasons: the specific disease type (tremor dominant versus akinetic-rigid), behavioral tasks (passive or active or no limb movements) during the recording (Rodriguez-Oroz et al., 2001), disease duration (Remple et al., 2011), as well as the exact recording place (Deffains et al., 2014). The method-dependent dispersion of average firing rate values observed here is 25 Hz which is identical to the rate dispersion reported in the literature. Similarly, the amount of regular, irregular, and bursty single units strongly depends on the sorting method. These three firing patterns have been demonstrated in the physiological state of human STNs. The most common one is irregular firing in up to 52% of the cells, followed by 36% bursting activity and regular firing observed in 12% of the cells in essential tremor patients (Steigerwald et al., 2008). In the parkinsonian state, the percentage of bursting cells strongly increases to up to 70% (Bergman et al., 1994; Steigerwald et al., 2008). Thus, bursting single units are a characteristic feature of STN recordings in PD patients (Beurrier et al., 1999; Levy et al., 2001; Chibirova et al., 2005; Lourens et al., 2013) and are reported to vary from 5% to 25% (Chibirova et al., 2005) or 15% to 34% (Lourens et al., 2013), depending on the exact recording site. We find a similar amount of variability,

namely 7% to 25% bursting single units, solely ascribed to the sorting method.

Recordings with intervals of high multi-unit firing are a particular challenge in spike sorting (Lewicki, 1998; Einevoll et al., 2012). The most difficult case is several bursting single units registered on one electrode. Such recordings typically comprise overlapping spikes which are hard to cluster due to distorted spike shapes (Lewicki, 1998; Harris et al., 2000; Bar-Gad et al., 2001b). Another issue of bursting activity is a possible spike amplitude reduction over time (Fee et al., 1996b; Quirk and Wilson, 1999; Harris et al., 2000). The latter can be consequence of sustained firing, but spike shape modulations may also occur during electrode drifts (Fee et al., 1996b; Quirk and Wilson, 1999; Knieling et al., 2016). In such cases, the spike amplitude should not be used as major feature for the clustering (Harris et al., 2000), nor should one expect circular but rather elongated clusters in PC feature space (Fee et al., 1996a; Fee et al., 1996b). We do not directly consider these issues since a visual inspection of the STN data did not show spike shape modulations. This is most likely related to the relatively low average firing rates in our data, as well as to our short recording times. Our AD, however, includes overlapping spikes which are most likely classified as perturbation signals. Moreover, our AD contains spike pairs with a controlled difference in amplitudes, representing the similarity categories. Their effect on the sorting results is roughly identical for all methods. Only for the assumption of Gaussian distributions (TDEM versus StEM methods) we observe a tendency to collapse similar spikes into one single cluster. Such tendencies presumably contribute to the diversity in the dynamical properties of STN cells reported above, especially in the case of high firing rates.

We restrict ourselves to the *unsupervised* methods offered by the OFS and we do not evaluate the sorting performance in different firing regimes. In terms of artificial data, an analysis focusing on the latter is principally straightforward, but beyond the scope of this study. Modern methods are often particularly suited for dense recoding arrays (Rossant et al., 2016; Chung et al., 2017; Yger et al., 2018) which, among many other advantages, enable a better handling of overlapping spikes. Many of these and other methods involve parameter and/or feature optimization which is often automatized (Fee et al., 1996b; Quiroga et al., 2004; Wild et al., 2012; Chung et al., 2017), but may also require user intervention, for example during cluster curation (Rossant et al., 2016; Yger et al., 2018). Typically, spike sorting is performed in the time and/or frequency domain (Lewicki, 1998; Einevoll et al., 2012; Quiroga, 2012; Wild et al., 2012). Here, the issue related to bursting cells can be accounted for by using the ISI distribution (Delescluse and Pouzat, 2006) or by assuming non-Gaussian distributions (Fee et al., 1996a). Another possibility is presented in Aksenova et al. (2003); Chibirova et al. (2005); Caro-Martin et al. (2018), namely template matching (TMPS) or K-means clustering with feature optimization (K-TOPS) in phase space, i.e. using spike derivative-based features. Chibirova et al. (2005) demonstrate a successful application of their unsupervised TMPS to STN recordings from PD patients.

In summary, different spike sorting approaches yield highly variable results. In order to recommend a sorting method we distinguish between two cases: 'clean' and 'noisy' data. With 'clean' we mean that a first visual inspection of the data indicates that there are only a few artifacts and distorted spike shapes – or the given perturbations can easily be identified and removed otherwise. With 'noisy data' we mean frequent perturbations that are difficult to identify and to remove. If the data is relatively clean we recommend to use the KMS method since it offers the highest success rate (Fig. 9) due to a high sensitivity (Fig. 8B1) and relatively well isolated clusters. If the data is particularly noisy and if missed spikes are less relevant for the subsequent analysis, VS(S)2D is probably a better choice. It combines a high specificity with an intermediate sensitivity (Fig. 8B1,B2) so that its success rate is higher in case of many perturbations (Fig. 9) and yields very few rpv.

The procedure described here could generally serve as a pre-analysis step to select the appropriate sorting method for a specific data set: One first generates an AD set with known ground truth which is adapted to the experimental recordings. If necessary, bursts of spikes and spike shape modulations can be included. The sorting algorithms in question are then applied to the AD and the results are evaluated in relation to the ground truth. Finally, one selects the method with the best results and applies it to the experimental recordings. It is elementary enough to be generally applicable but yields results specific to the given data. Our results clearly show the importance of a careful spike sorting method selection.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft Grant 1753/3-1 Klinische Forschergruppe (KFO219, TP12), Deutsche Forschungsgemeinschaft Grant GR 1753/4-2 & DE 2175/2-1 Priority Program (SPP 1665), the Helmholtz Association through the Helmholtz Portfolio Theme Supercomputing and Modeling for the Human Brain (SMHB), and by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 720270 & 785907 (Human Brain Project SGA1 & SGA2). We thank Paul Chorely and Alexa Riehle for technical help and fruitful discussions. Tragically, Paul Chorely died before we were able to finish the manuscript.

REFERENCES

Adamos DA, Kosmidis EK, Theophilidis G. (2008) Performance evaluation of PCA-based spike sorting algorithms. COMPUT METH PROG BIO 91:232-244.

Aksenova TI, Chivirova OK, Dryga O, Tetko IV, Benabid AL, Villa AEP. (2003) An unsupervised automatic method for sorting neuronal spike waveforms in awake and freely moving animals. Methods 30:178-187.

Bar-Gad I, Ritov Y, Bergman H. (2001) The neuronal refractory period causes a short-term peak in the autocorrelation function. J Neurosci Meth 104:155-163.

Bar-Gad I, Ritov Y, Vaadia E, Bergman H. (2001) Failure in identification of overlapping spikes from multiple neuron activity causes artificial correlations. J Neurosci Meth 107:1-13.

Barnett AH, Magland JF, Greengard LF. (2016) Validation of neural spike sorting algorithms without ground-truth information. J Neurosci Meth 264:65-77.
Benazzouz A, Breit S, Koudsie A, Pollak P, Krack P, Benabid AL. (2002) Intraoperative microrecordings of the subthalamic nucleus in Parkinson's disease. Mov Disord 17:S145-S149.
Bergman H, Wichmann T, Karmon B, DeLong MR. (1994) The primate subthalamic nucleus. II Neuronal activity in the MPTP model of parkinsonism J Neurophysiol 72:507-520.
Beurrier C, Congar P, Bioulac B, Hammond C. (1999) Subthalamic nucleus neurons switch from single-spike activity to burst-firing mode. J Neurosci 19:599-609.
Bishop CM. (2006) Pattern recognition and machine learning, 2006. Springer-Verlag New York.
Brown EN, Kass RE, Mitra PP. (2004) Multiple neural spike train data analysis: state-of-the-art and future challenges. Nat Neurosci 7:456-461.
Buzsáki G. (2004) Large-scale recording of neuronal ensembles. Nat Neurosci 7:446-451.
Caro-Martin CR, Delgado-GarcÃa JM, Gruart A, Sanchez-Campusano R. (2018) Spike sorting based on shape, phase and distribution features, and K-TOPS clustering with validity and error indices. Sci Rep 8:17796.
Chibirova OK, Aksenova TI, Benabid AL, Chabardes S, Larouche S, Rouat J, Villa AEP. (2005) Unsupervised spike sorting of extracellular electrophysiological recording in subthalamic nucleus of parkinsonian patients. BioSystems 79:159-171.
Chung JE, Magland JF, Barnett AH, Tolosa VM, Tooker AC, Lee KY, Shah KG, Felix SH, Frank LM, Greengard LF. (2017) A fully automated approach to spike sorting. Neuron 95:1381-1394.
Dai J, Liu X, Yi Y, Zhang H, Wang J, Zhang S, Zheng X. (2008) Experimental study on neuronal spike sorting methods. Proceedings of the 2008 2nd international conference on future generation communication and networking 2. p. 230-233.
Deffains M, Holland P, Moshel S, Ramirez de Noriega F, Bergman H, Israel Z. (2014) Higher neuronal discharge rate in the motor area of the subthalamic nucleus of parkinsonian patients. J Neurophysiol 112:1409-1420.
Delescluse M, Pouzat C. (2006) Efficient spike-sorting of multi-state neurons using inter-spike intervals information. J Neurosci Meth 150:16-29.
Eden UT, Gale JT, Amirnovin R, Eskandar EN. (2012) Characterizing the spiking dynamics of subthalamic nucleus neurons in Parkinson's disease using generalized linear models. Front Integr Neurosci 6:28.
Einevoll GT, Franke F, Hagen E, Pouzat C, Harris KD. (2012) Towards reliable spike-train recordings from thousands of neurons with multielectrodes. Curr Opin Neurol 22:11-17.
Fee MS, Mitra PP, Kleinfeld D. (1996) Automatic sorting of multiple unit neuronal signals in the presence of anisotropic and non-Gaussian variability. J Neurosci Meth 69:175-188.
Fee MS, Mitra PP, Kleinfeld D. (1996) Variability of extracellular spike waveforms of cortical neurons. J Neurophysiol 76:3823-3833.
Figueiredo MAT, Jain AK. (2002) Unsupervised learning of finite mixture models. IEEE T PATTERN ANAL 24:381-396.
Florin E, Reck C, Burghaus L, Lehrke R, Gross J, Sturm V, Fink GR, Timmermann L. (2008) Ten hertz thalamus stimulation increases tremor activity in the subthalamic nucleus in a patient with Parkinson's disease. Clin Neurophysiol 119:2098-2103.
Florin E, Himmel M, Reck C, Maarouf M, Schnitzler A, Sturm V, Fink GR, Timmermann L. (2012) Subtype-specific statistical causalities in parkinsonian tremor. Neuroscience 210:353-362.
Fukunaga, K., 1990. Introduction to statistical Pattern recognition (2Nd Ed.). Academic press professional, Inc., San Diego, CA, USA.
Georgopoulos A, Kalaska J, Caminiti R, Massey J. (1982) On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. J Neurosci 2(11):1527-1537.
Gibson S, Judy JW, Markovic D. (2008) Comparison of spike-sorting algorithms for future hardware implementation. ENG MED BIOL SOC ANN 2008:5015-5020.
Gregory A. Wilkin, G.A., Huang, X., 2007. K-Means Clustering Algorithms: Implementation and Comparison. Proc. 2nd IMSCCS , 133–136.
Gross RE, Krack P, Rodriguez-Oroz MC, Rezai AR, Benabid AL. (2006) Electrophysiological mapping for the implantation of deep brain stimulators for Parkinson's disease and tremor. Movement Disord 21.
Hamani C. (2004) The subthalamic nucleus in the context of movement disorders. Brain 127:4-20.
Harris KD, Henze DA, Csicsvari J, Hirase H, Buzsáki G. (2000) Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. J Neurophysiol 84:401-414.
Hill D, Kleinfeld D, Mehta SB. (2007) Spike Sorting. In: Mitra PP, & Bokil H, editors. In observed brain dynamics, Vol. 1. Oxford Press. chapter Chapter. p. 1-17.
Hill DN, Mehta SB, Kleinfeld D. (2011) Quality metrics to accompany spike sorting of extracellular signals. J Neurosci 31:8699-8705.
Horton PM, Nicol AU, Kendrick KM, Feng JF. (2007) Spike sorting based upon machine learning algorithms (SOMA). J Neurosci Meth 160:52-68.
Hubel DH, Wiesel TN. (1959) Receptive fields of single neurones in the cat's striate cortex. J Physiol 148:574-591.
Hulata E, Segev R, Ben-Jacob E. (2002) A method for spike sorting and detection based on wavelet packets and Shannon's mutual information. J Neurosci Meth 117:1-12.
Hutchison WD, Allan RJ, Opitz H, Levy R, Dostrovsky JO, Lang AE, Lozano AM. (1998) Neurophysiological identification of the subthalamic nucleus in surgery for Parkinson's disease. Ann Neurol 44:622-628.
Joshua M, Elias S, Levine O, Bergman H. (2007) Quantifying the isolation quality of extracellularly recorded action potentials. J Neurosci Meth 163:267-282.
Kelley R, Flouty O, Emmons EB, Kim Y, Kingyon J, Wessel JR, Oya H, Greenlee JD, Narayanan NS. (2018) A human prefrontal-subthalamic circuit for cognitive control. Brain 141(1):205-216.
Knieling S, Sridharan KS, Belardinelli P, Naros G, Weiss D, Mormann F, Gharabaghi A. (2016) An Unsupervised Online Spike-Sorting Framework Int J Neural Syst :1550042.
Kretzberg J, Coors T, Furche J. (2009) Comparison of valley seeking and T-distributed EM algorithm for spike sorting. BMC Neurosci 10:P47.
Kühn AA, Trottenberg T, Kivi A, Kupsch A, Schneider GH, Brown P. (2005) The relationship between local field potential and neuronal discharge in the subthalamic nucleus of patients with Parkinson's disease. Exp Neurol 194:212-220.
Lefebvre B, Yger P, Marre O. (2016) Recent progress in multi-electrode spike sorting methods. J of Physiol-Paris 110:327-335.
Levy R, Dostrovsky JO, Lang AE, Sime E, Hutchison WD, Lozano AM. (2001) Effects of apomorphine on subthalamic nucleus and globus pallidus internus neurons in patients with Parkinson's disease. J Neurophysiol 86:249-260.
Levy R, Hutchison WD, Lozano AM, Dostrovsky JO. (2002) Synchronized neuronal discharge in the basal ganglia of parkinsonian patients is limited to oscillatory activity. J Neurosci 22:2855-2861.
Lewicki MS. (1998) A review of methods for spike sorting: the detection and classification of neural action potentials. Network 9:R53-R78.
Lipski WJ, Alhourani A, Pirnia T, Jones PW, Dastolfo-Hromack C, Helou LB, Crammond DJ, Shaiman S, Dickey MW, Hold LL, Turner RS, Fiez JA, Richardson RM. (2018) Subthalamic nucleus neurons differentially encode early and late aspects of speech production. J Neurosci 38(24):5620-5631.
Lourens MAJ, Meijer HGE, Contarino MF, van den Munckhof P, Schuurman PR, van Gils SA, Bour LJ. (2013) Functional neuronal activity and connectivity within the subthalamic nucleus in Parkinson's disease. Clin Neurophysiol 124:967-981.
McNaughton BL, O Keefe J, Barnes CA. (1983) The stereotrode: a new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records. J Neurosci Methods 8:391-397.
Michmizos, K.P., Nikita, K.S., 2010. Can we infer subthalamic nucleus spike trains from intranuclear local field potentials?, in: ENG MED BIOL SOC ANN, pp. 5476–5479.
Moran A, Bergman H, Israel Z, Bar-Gad I. (2008) Subthalamic nucleus functional organization revealed by parkinsonian neuronal oscillations and synchrony. Brain 131:3395-3409.

Mrakic-Sposta S, Marceglia S, Egidi M, Carrabba G, Rampini P, Locatelli M, Foffani G, Accolla E, Cogiamanian F, Tamma F, Barbieri S, Priori A. (2008) Extracellular spike microrecordings from the subthalamic area in Parkinson's disease. J Clin Neurosci 15:559-567.
Pazienti A, Grün S. (2006) Robustness of the significance of spike synchrony with respect to sorting errors. J Comput Neurosci 21:329-342.
Quirk MC, Wilson MA. (1999) Interaction between spike waveform classification and temporal sequence detection. J Neurosci Meth 94:41-52.
Quiroga RQ. (2007) Spike sorting. Scholarpedia 2:3583.
Quiroga RQ. (2009) What is the real shape of extracellular spikes? J Neurosci Meth 177:194-198.
Quiroga RQ. (2012) Spike sorting. Curr Biol 22:R45-R46.
Quiroga RQ, Nadasdy Z, Ben-Shaul Y. (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput 16:1661-1687.
Raz A, Vaadia E, Bergman H. (2000) Firing patterns and correlations of spontaneous discharge of pallidal neurons in the normal and the tremulous 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine vervet model of parkinsonism. J Neurosci 20:8559-8571.
Reck C, Florin E, Wojtecki L, Krause H, Groiss S, Voges J, Maarouf M, Sturm V, Schnitzler A, Timmermann L. (2009) Characterisation of tremor-associated local field potentials in the subthalamic nucleus in Parkinson's disease. Eur J Neurosci 29:599-612.
Reck C, Maarouf M, Wojtecki L, Groiss SJ, Florin E, Sturm V, Fink GR, Schnitzler A, Timmermann L. (2012) Clinical outcome of subthalamic stimulation in Parkinson's disease is improved by intraoperative multiple trajectories microelectrode recording. J Neurol Surg 73:377-386.
Remple MS, Brandenham CH, Kao CC, Charles PD, Neimat JS, Konrad PE. (2011) Subthalamic nucleus neuronal firing rate increases with parkinson's desease progession. Movement Disord 26(9):1657-1662.
Rey HG, Pedreira C, Quiroga RQ. (2015) Past, present and future of spike sorting techniques. Brain Res Bull 119:106-117.
Rodriguez-Oroz MC, Rodriguez M, Guridi J, Mewes K, Chockkman V, Vitek J, DeLong MR, Obeso JA. (2001) The subthalamic nucleus in parkinson's disease: somatotopic organization and physiological characteristics. Brain 124:1777-1790.
Rossant C, Kadir SN, Goodman DFM, Schulman J, Hunter MLD, Saleem AB, Grosmark A, Belluscio M, Denfield GH, Ecker AS, Tolias AS, Solomon S, Buzsaki G, Carandini M, Harris KD. (2016) Spike sorting for large, dense electrode arrays. Nat Neurosci 19:634-641.
Rutishauser U, Schuman EM, Mamelak AN. (2006) Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. J Neurosci Meth 154:204-224.
Sahani M. (1999) Latent variable models for neural data analysis. California Institute of Technology Pasadena, CA, USA: Doctoral Dissetation, 1999.
Schrock LE, Ostrem JL, Turner RS, Shimamoto SA, Starr PA. (2009) The subthalamic nucleus in primary dystonia: single-unit discharge characteristics. J Neurophysiol 102(6).
Shimamoto SA, Ryapolova-Webb ES, Ostrem JL, Galifianakis NB, Miller KJ, Starr PA. (2013) Subthalamic nucleus neurons are synchronized to primary motor cortex local field potentials in Parkinson's disease. J Neurosci 33:7220-7233.
Shinomoto S, Shima K, Tanji J. (2003) Differences in spiking patterns among cortical neurons. Neural Comput 15:2823-2842.
Shoham S, Fellows MR, Normann RA. (2003) Robust, automatic spike sorting using mixtures of multivariate t-distributions. J Neurosci Meth 127:111-122.
Steigerwald F, Pötter M, Herzog J, Pinsker M, Kopper F, Mehdorn H, Deuschl G, Volkmann J. (2008) Neuronal activity of the human subthalamic nucleus in the parkinsonian and nonparkinsonian state. J Neurophysiol 100:2515-2524.
Todorova S, Sadtler P, Batista A, Chase S, Ventura V. (2014) To sort or not to sort: the impact of spike-sorting on neural decoding performance. J Neural Eng 11.
Weinberger M, Mahant N, Hutchison WD, Lozano AM, Moro E, Hodaie M, Lang AE, Dostrovsky JO. (2006) Beta oscillatory activity in the subthalamic nucleus and its relation to dopaminergic response in Parkinson's disease. J Neurophysiol 96:3248-3256.
Wheeler BC, Heetderks WJ. (1982) A comparison of techniques for classification of multiple neural signals. IEEE T BIO-MED ENG 29:752-759.
Wild J, Prekopcsak Z, Sieger T, Novak D, Jech R. (2012) Performance comparison of extracellular spike sorting algorithms for single-channel recordings. J Neurosci Meth 203:369-376.
Wood F, Black MJ, Vargas-Irwin C, Fellows M, Donoghue JP. (2004) On the variability of manual spike sorting. IEEE T BIO-MED ENG 51:912-918.
Yang AI, Vanegas N, Lungu C, Zaghloul KA. (2014) Beta-coupled high-frequency activity and beta-locked neuronal spiking in the subthalamic nucleus of Parkinson's disease. J Neurosci 34:12816-12827.
Yang C, Olson B, Si J. (2011) A multiscale correlation of wavelet coefficients approach to spike detection. Neural Comput 23:215-250.
Yger P, Spampinato GLB, Esposito E, Lefebvre B, Deny S, Gardella C, Stimberg M, Jetter F, Zeck G, Picaud S, Duebel J, Marre O. (2018) A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo. eLife 7e34518.
Zhang C, Zhang X, Zhang MQ, Li Y. (2007) Neighbor number, valley seeking and clustering. Pattern Recogn Lett 28:173-180.

(Received 21 December 2018, Accepted 1 July 2019)
(Available online 9 July 2019)