Electrical Stimulus Artifact Cancellation and Neural Spike Detection on Large Multi-Electrode Arrays

Gonzalo E. Mena[1,*], Lauren E. Grosberg[3], Paweł Hottowy[4], Alan Litke[5], John Cunningham[1,2], E.J. Chichilnisky[3], and Liam Paninski[1,2].

1 Statistics Department, Columbia University,New York, NY, 10027, USA
2 Grossman Center for the Statistics of Mind and Center for Theoretical Neuroscience, Columbia University.
3 Department of Neurosurgery and Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA 94305, USA
4 Physics and Applied Computer Science, AGH University of Science and Technology, 30-059 Krakow, Poland
5 Santa Cruz Institute for Particle Physics, University of California, Santa Cruz, Santa Cruz, CA 95064, USA

* gem2131@columbia.edu

Abstract

Simultaneous electrical stimulation and recording using multi-electrode arrays can provide a valuable technique for studying circuit connectivity and engineering neural interfaces. However, interpreting these recordings is challenging because the spike sorting process (identifying and segregating action potentials arising from different neurons) is greatly complicated by electrical stimulation artifacts across the array, which can exhibit complex and nonlinear waveforms. Here we develop a scalable algorithm based on a structured Gaussian Process model to estimate and subtract the artifact. The effectiveness of our method is demonstrated in both real and simulated 512-electrode recordings in the peripheral primate retina, with single and two-electrode electrical stimulation. This technology may be helpful in the design of future high-resolution sensory prostheses based on tailored stimulation (e.g., retinal prostheses), and for closed-loop neural stimulation at a much larger scale than currently possible.

1 Introduction

Simultaneous electrical stimulation and recording with multi-electrode arrays (MEAs) serves at least two important purposes for investigating neural circuits and for neural engineering. First, it enables the probing of neural circuits, leading to improved understanding of circuit anatomy and function [1–6]. Second, it can be used to assess and optimize the performance of brain-machine interfaces, such as retinal prostheses [7, 8], by exploring the patterns of stimulation required to achieve particular patterns of neural activity. However, identifying neural activity in the presence of artifacts introduced by electrical stimulation is a major challenge, and automation is required to efficiently analyze recordings from large-scale MEAs. Furthermore, closed-loop experiments require the ability to assess neural responses to stimulation in real time to actively update the stimulus and probe the circuit, so the automated approach for identifying neural activity must be fast [9, 10].

Spike sorting methods [11–13] allow identification of neurons from their spatio-temporal electrical footprints recorded on the MEA. However, these methods fail when used on data corrupted by stimulation artifacts. Although technological advances in stimulation circuitry have enabled recording with significantly reduced

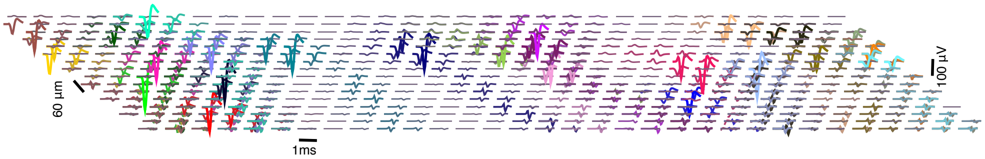

Fig 1. Overlapping electrical images of 24 neurons (different colors) over the MEA, aligned to onset of spiking at $t = 0.5ms$. Each trace represents the time course of voltage at a certain electrode. For each neuron, traces are only shown in the electrodes with a strong enough signal.

artifacts [14–18], identification of neural responses from artifact-corrupted recordings still presents a challenging task — even for human experts — since these artifacts can be much larger than spikes [19], overlap temporally with spikes, and occupy a similar temporal frequency band as spikes.

Although a number of approaches have been previously proposed to tackle this problem [20–23], there are two shortcomings we address here. First, previous approaches are based on restrictive assumptions on the frequency of spikes and their latency distribution (e.g, stimulation-elicited spikes have to occur a certain amount of time after the stimulus). In consequence, it becomes necessary to discard non-negligible portions of the recordings [19, 24], leading to biased results that may miss the regimes where the most interesting neuronal dynamics occur [25, 26]. Second, all of these methods have a local nature, i.e., they are based on electrode-wise estimates of the artifact that don't exploit the shared spatio-temporal information present in MEAs. In general this leads to suboptimal performance. Therefore, a scalable computational infrastructure for spike sorting with stimulation artifacts in large-scale setups is necessary.

This paper presents a method to identify single-unit spike events in electrical stimulation and recording experiments using large-scale MEAs. We develop a modern, large-scale, principled framework for the analysis of neural voltage recordings that have been corrupted by stimulation artifacts. First, we model this highly structured artifact using a structured Gaussian Process (GP) to represent the observed variability across stimulation amplitudes and in the spatial and temporal dimensions measured on the MEA. Next, we introduce a spike detection algorithm that leverages the structure imposed in the GP to achieve a fast and scalable implementation. Importantly, our algorithm exploits many characteristics that make this problem tractable, allowing it to separate the contributions of artifact and neural activity to data. For example, the artifact is smooth in certain dimensions, with spatial footprints that are different than those of spikes. Also, artifact variability is different than that of spikes: while the artifact does not substantially change if the same stimulus is repeated, responses of neurons in many stimulation regimes are stochastic, enhancing identifiability.

The effectiveness of our method is demonstrated by comparison on simulated data and against human-curated inferred spikes extracted from real data recorded in primate retina. Although some features of our method are context-dependent, we discuss extensions to other scenarios, stressing the generality of our approach.

2 Materials and Methods

In this section we develop a method for identifying neural activity in response to electrical stimulation. We assume access to voltage recordings $Y(e, t, j, i)$ in a MEA with $e = 1, \ldots, E$ electrodes (here, $E = 512$), during $t = 1, \ldots T$ timepoints (e.g., $T = 40$, corresponding to 2 milliseconds for a 20Khz sampling rate) after the presentation of $j = 1, \ldots, J$ different stimuli, each of them being a current pulse of increasing amplitudes a_j (in other words, the a_j are magnification factors applied to an unitary pulse). For each of these stimuli a number n_j of trials or repetitions is available; i indexes trials. Each recorded data segment is modeled as a sum of the true signal of interest s (neural spiking activity on that electrode), plus two types of noise.

The first noise source, A, is the large artifact that results from the electrical stimulation at a given electrode. This artifact has a well defined structure but its exact form in any given stimulus condition is not known *a priori* and must be estimated from the data and separated from occurrences of spikes. Although in typical experimental setups one will be concerned with data coming from many different stimulating electrodes, for clarity we start with the case of just a single stimulating electrode; we will generalize this below.

The second source of noise, ϵ, is additive spherical Gaussian observation noise; that is, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{d'})$, with $d' = T \times E \times \sum_{j=1}^{J} n_j$. This assumption is rather restrictive and we assume it here for computational ease, but refer the reader to the discussion for a more general formulation that takes into account correlated noise.

Additionally, we assume that electrical images (EI) [27, 28] — the spatio-temporal collection of action potential shapes on every electrode e — are available for all the N neurons under study. In detail, each of these EIs are estimates of the voltage deflections produced by a spike over the array in a length T' time window, with onset of the spike aligned to an arbitrary value. They are represented as matrices with dimensions $E \times T'$ and can be obtained in a separate experiment in the absence of electrical stimulation, using standard large-scale spike sorting methods (e.g. [12]). Fig 1 shows examples of many EIs, or templates, obtained during a visual stimulation experiment.

Finally, we assume the observed traces are the linear sum of neural activity, artifact, and other noise sources; that is:

$$Y = A + s + \epsilon. \tag{1}$$

Fig2 illustrates the difficulty of this problem: even if 1) for low-amplitude stimuli the artifact may not heavily corrupt the recorded traces and 2) the availability of several trials can enhance identifiability — as traces with spikes and no spikes naturally cluster into multiple groups — in the general case we will be concerned also with high amplitudes of stimulation. In these regimes, spikes could significantly overlap temporarily with the artifact, occur with high probability and almost deterministically, i.e., with low latency variability. For example, in the rightmost columns of fig 2, spike identification is not straightforward since all the traces look alike, and the shape of a typical trace does not necessarily suggest the presence of neural activity. There, inference of neural activity is only possible given a reasonable estimate of the artifact: for instance, under the assumption that the artifact is a smooth function of the stimulus strength, one can make a good initial guess of the artifact by considering the artifact at a lower stimulation amplitude, where spike identification is relatively easier.

Therefore, a solution of this problem will rely on methods for an appropriate separation of neural activity and artifact, which in turn necessitates the use of sensible models that properly capture the structure of the latter; that is, how it varies along the different relevant dimensions. In the following we develop such a method, and divide its exposition in four parts. We start by describing in 2.1 how to model neural activity, Second, in 2.2 we describe the structure of the stimulation artifacts. Third, in 2.3 we propose a GP model to represent this structure. Finally, in 2.4 we introduce a scalable algorithm that produces an estimate of A and s given recordings Y.

2.1 Modeling neural activity

We assume that s is the linear superposition of the activities s^n of the N neurons involved, i.e. $s = \sum_{n=1}^{N} s^n$. Furthermore, each of these activities is expressed in terms

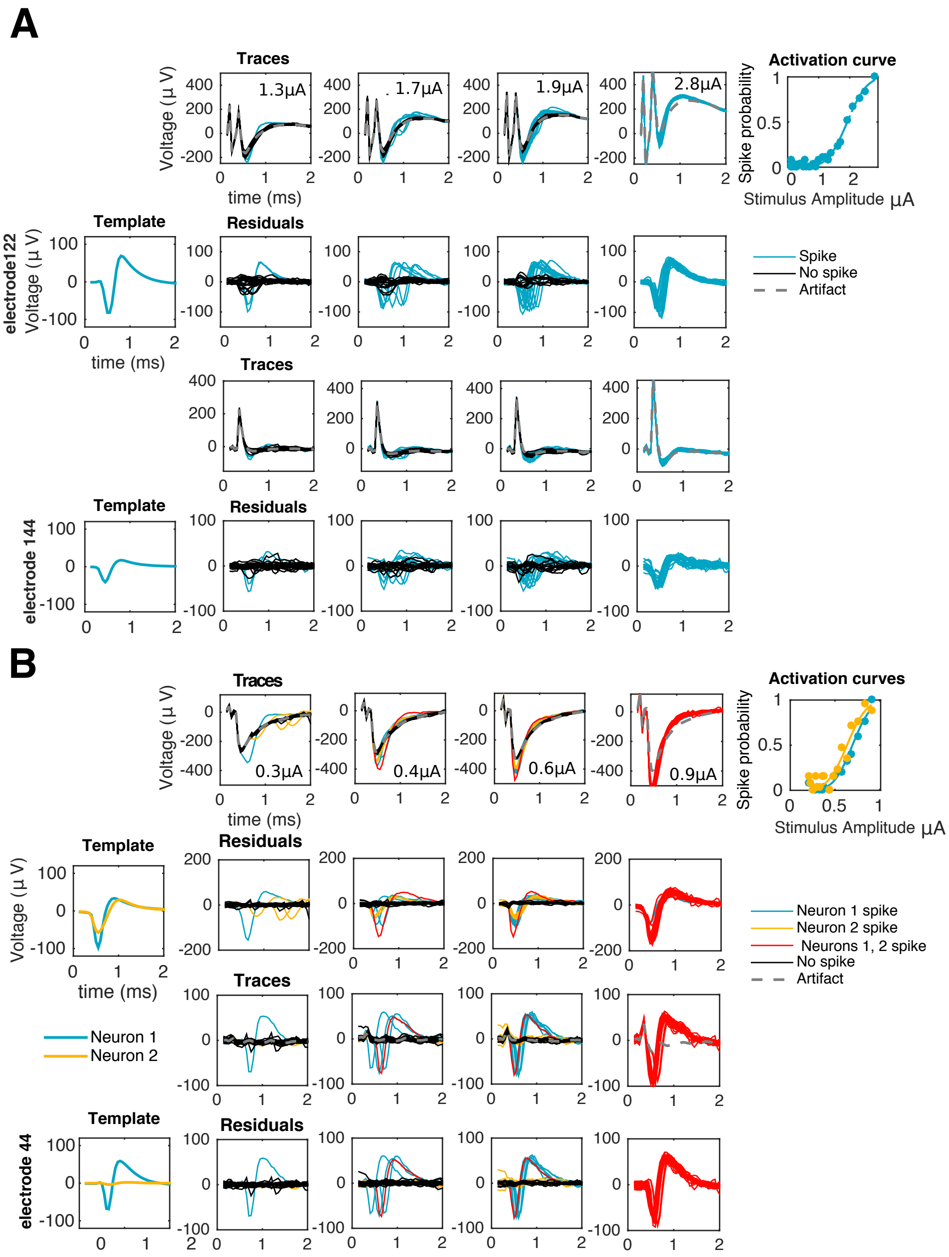

Fig 2. Visual inspection of traces reveals the difficulty of the problem. First column: templates of spiking neurons. Second to fourth columns: responses of one **A**) or two **B**) cells to electrical stimulation at increasing stimulation amplitudes as recorded in the stimulating electrode (first rows) or a neighboring, non-stimulating electrode (third rows). If the stimulation artifact is known (gray traces) it can be subtracted from raw traces to produce a baseline (second and fourth rows) amenable for template matching: traces with spike(s) (colored) match, on each electrode, either a translation of a template (**A** and **B**) or the sum of different translations of two or more templates **B**). As reflected by the activation curves (fifth column) for strong enough stimuli spiking occurs with probability close to one, consistent with the absence of black traces in the rightmost columns.

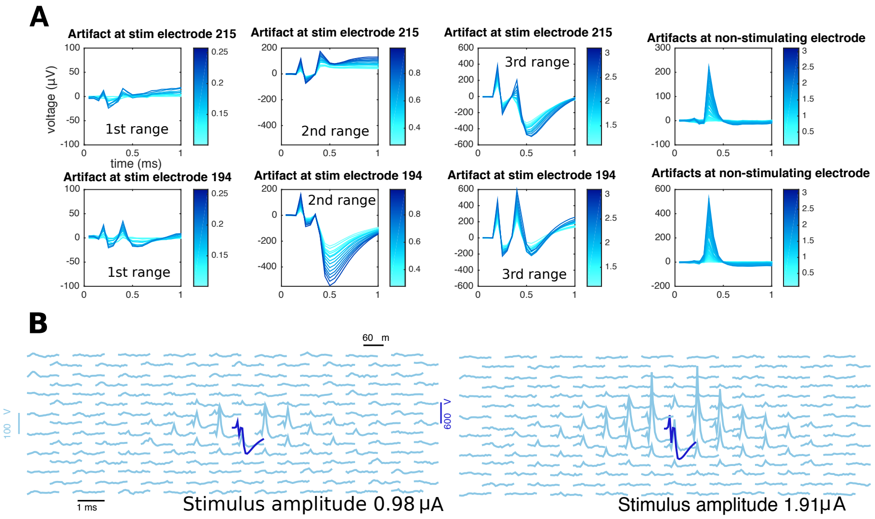

Fig 3. Properties of the electrical stimulation artifact revealed by TTX experiments. **A)** local, electrode-wise properties of the stimulation artifacts. Overall, magnitude of the artifact increases with stimulation strength (different shades of blue). However, unlike non-stimulating electrodes, where artifacts have a typical shape of a bump around 0.5 ms (fourth column), the case of the stimulating electrode is more complex: besides the apparent increase in artifact strength, the shape itself is not a simple function of stimulating electrode (first and second rows). Also, for a given stimulating electrode the shape of the artifact is a complex function of the stimulation strength, changing smoothly only within certain stimulation ranges: here, responses to the entire stimulation range are divided into three ranges (first, second, and third column) and although traces within each range look alike, traces from different ranges cannot be guessed from other ranges. **B)** stimulation artifacts in a neighborhood of the stimulating electrode, at two different stimulus strengths (left and right). Each trace represents the time course of voltage at a certain electrode. Notice that stimulating electrode (blue) and non-stimulating electrodes (light blue) are plotted in different scales.

of the binary vectors b^n that indicate spike occurrence and timing: specifically, if $s^n_{j,i}$ is the neural activity of neuron n at trial i of the j-th stimulation amplitude, we write $s^n_{j,i} = M^n b^n_{i,j}$, where M^n is a matrix that contains on each row a copy of the EI of neuron n (vectorizing over different electrodes) aligned to spiking occurring at different times. (M^n is defined for notational convenience only here; we never need to actually construct these matrices.) Notice that this binary representation immediately entails that: 1) on each trial each neuron fires at most once (this is the case in reality, as the recording window is comparable with the refractory period) and 2) that spikes can only occur over a discrete set of times (a strict subset of the entire recording window), which here corresponds to all the time samples between 0.25 ms and 1.5 ms. We refer the reader to [29] for details on how to relax this assumption that here we have taken for the sake of simplicity.

2.2 Stimulation Artifacts

Electrical stimulation experiments where neural responses are inhibited (e.g., using the neurotoxin TTX) provide qualitative insights about the structure of the stimulation artifact $A(e, t, j, i)$ (figure 3); that is, how it varies as a function of all the relevant covariates: space (represented by electrode, e), time t, amplitude of stimulus a_j, and stimulus repetition i. Repeating the same stimulation leads to the same artifact, up to small random fluctuations, and so by averaging several trials these fluctuations can be reduced, and we can conceive the artifact as a stack of movies $A(e,t,j)$, one for each amplitude of stimulation a_j.

We treat the stimulating and non-stimulating electrodes separately because of their observed different qualitative properties.

2.2.1 Stimulating electrode

Modeling the artifact in the stimulating electrode requires special care because it is this electrode that typically will capture the strongest neural signal in attempts to directly activate a soma (e.g. figure 3). The artifact is more complex in the stimulating electrode [16] and has the following properties in this preparation: 1) its magnitude is much greater than of the non-stimulating electrodes; 2) its effect persist at least 2 ms after the onset of the stimulus; and 3) it is a piece-wise continuous function of the stimulus strength. Discontinuities occur at a pre-defined set of stimulus amplitudes, the “breakpoints" (known beforehand), resulting from gain settings in the stimulation hardware that must change in order to apply stimuli of different magnitude ranges [16]. Notice that these discontinuities are a rather technical and context-dependent feature that may not necessarily apply to all stimulation systems, unlike the rest of the properties described here. Artifact waveforms resulting from stimulus amplitudes within each of the ranges defined by the breakpoints change smoothly (see figure 3A).

2.2.2 Non-stimulating electrodes

The artifact here is much more regular and of lower magnitude, and has the following properties (see figure 3): 1) its magnitude peaks around $.4ms$ following the stimulus onset, and then rapidly stabilizes; 2) the artifact magnitude typically decays with distance from the stimulating electrode; 3) the magnitude of the artifact increases with increasing stimulus strength.

Based on these observations, we develop a general framework for artifact modeling based on GPs.

2.3 A structured GP model for stimulation artifacts

From the above discussion we conclude that the artifact is highly non-linear (on each coordinate), non-stationary (i.e., the variability depends on the value of each coordinate), but structured. The GP framework [30] provides powerful and computationally scalable methods for modeling non-linear functions given noisy measurements, and leads to a straightforward implementation of all the usual operations that are relevant for our purposes (e.g., interpolation and/or extrapolation across time or different electrodes) in terms of some tractable conditional Gaussian distributions.

To better understand the rationale guiding the choice of GPs, consider first a simple Bayesian regression model for the artifact as a noisy linear combination of M basis functions $\Phi_i(e,t,j)$ (e.g polynomials); that is, $A(e,t,j)=\sum_{i=1}^{M} w_i\Phi_i(e,t,j)+\epsilon$, with a regularizing prior $p(w)$ on the weights. If $p(w)$ and ϵ are modeled as Gaussian, and if we consider the collection of $A(e,t,j)$ values (over all electrodes e, timesteps t, and stimulus amplitude indices j) as one large vector A, then this translates into an assumption that the vector A is drawn from a high-dimensional Gaussian distribution. The prior mean μ and covariance K of A can easily be computed in terms of Φ and $p(w)$. Importantly, this simple model provides us with tools to estimate the posterior distribution of A given partial noisy observations (for example, we could estimate the posterior of A at a certain electrode if we are given its values on the rest of the array).

Since A in this model is a stochastic process (indexed by e, t, and j) with a Gaussian distribution, we say that A is modeled as a Gaussian process, and write $A \sim \mathcal{GP}(\mu, K)$.

The main problem with the approach sketched above is that one has to solve some challenging model selection problems: what basis functions Φ_i should we choose, how large should M be, what parameters should we use for the prior $p(w)$, and so on. We can avoid these issues by instead directly specifying the covariance K and mean μ (instead of specifying K and μ indirectly, through $p(w)$, Φ, etc.).

The parameter μ informs us about the mean behavior of the samples from the GP (here, the average values of the artifact). Briefly, we estimate $\hat{\mu}$ by taking the mean of the recordings at the lowest stimulation amplitude and then subtract off that value from all the traces, so that μ can be assumed to be zero in the following. We refer the reader to the supporting information for details, and stress that all the figures shown in the main text are made after applying this mean-subtraction pre-processing operation.

Next we need to specify K. This “kernel" can be thought of as a square matrix of size $\dim(A) \times \dim(A)$, where $\dim(A)$ is as large as $T \times E \times J \sim 10^6$ in our context. This number is large enough so all elementary operations (e.g. kernel inversion) are prohibitively slow unless further structure is imposed on K — indeed, we need to avoid even storing K in memory, and estimating such a high-dimensional object is impossible without some kind of strong regularization. Thus, instead of specifying every single entry of K we need to exploit a simpler, lower-dimensional model that is flexible enough to enforce the qualitative structure on A that we described in the preceding section.

Specifically, we impose a separable Kronecker product structure on K, leading to tractable and scalable inferences [31, 32]. This Kronecker product is defined for any two matrices as $(A \otimes B)_{((i_1,i_2),(j_1,j_2))} = A_{(i_1,j_1)} B_{(i_2,j_2)}$. The key point is that this Kronecker structure allows us to break the huge matrix K into smaller, more tractable pieces whose properties can be easily specified and matched to the observed data. The result is a much lower-dimensional representation of K that serves to strongly regularize our estimate of this very high-dimensional object.

We state separate Kronecker decompositions for the non-stimulating and stimulating electrodes. For the non-stimulating electrode we assume the following decomposition:

$$K = \rho K_t \otimes K_e \otimes K_s + \phi^2 I_{\dim(A)}, \tag{2}$$

where K_t, K_e and K_s are the kernels that account for variations in the time, space, and stimulus magnitude dimensions of the data, respectively. One way to think about the Kronecker product $K_t \otimes K_e \otimes K_s$ is as follows: start with an array $z(t, e, s)$ filled with independent standard normal random variables, then apply independent linear filters in each direction t, e, and s to z so that the marginal covariances in each direction correspond to K_t, K_e, and K_s, respectively. The dimensionless quantity ρ is used to control the overall magnitude of variability and the scaled identity matrix $\phi^2 I_{\dim(A)}$ is included to capture the fact that what is finally observed is a noise-corrupted version of the actual artifact. Notice that we distinguish between this noise variance ϕ^2 and the observation noise variance σ^2, associated with the error term ϵ of Eq 1.

Likewise, for the stimulating electrode we consider the kernel:

$$K' = \sum_{i=1}^{r} \rho^r K_t^r \otimes K_s^r + \phi^2 I_{T \times J}. \tag{3}$$

Here, the sum goes over the stimulation ranges defined by consecutive breakpoints; and for each of those ranges, the kernel K_s^r has non-zero entries only for the stimulation values within the r-th range between breakpoints. In this way, we ensure artifact information is not shared for stimulus amplitudes across breakpoints. Finally, ρ^r and ϕ'^2 play a similar role as in Eq (2).

Now that this structured kernel has been stated it remains to specify parametric families for the elementary kernels $K_t, K_e, K_s, K_t^r, K_s^r$. We construct these from the Matérn family, using extra parameters to account for the behaviors described in 2.2.

2.3.1 A non-stationary family of kernels

We consider the Matérn(3/2) kernel, the continuous version of an autoregressive process of order 2. Its (stationary) covariance is given by

$$K_\lambda(x_1, x_2) = K_\lambda(\delta = |x_1 - x_2|) = \left(1 + \sqrt{3}\delta\lambda\right) \exp\left(-\sqrt{3}\delta\lambda\right). \qquad (4)$$

The parameter $\lambda > 0$ represents the (inverse) length-scale and determines how fast correlations decay with distance. We use this kernel as a device for representing smoothness; that is, the property that information is shared across a certain dimension (e.g. time). This property is key to induce reasonable extrapolation and filtering estimators, as required by our method (see 2.4). Naturally, given our rationale for choosing this kernel, similar results should be expected if the Matérn(3/2) was replaced by a similar, stationary smoothing kernel.

We induce non-stationarities by considering the family of unnormalized gamma densities $d_{\alpha,\beta}(\cdot)$:

$$d_{\alpha,\beta}(x) = \exp(-x\beta)x^\alpha. \qquad (5)$$

By an appropriate choice of the pair $(\alpha, \beta) > 0$ we aim to expressively represent non-stationary 'bumps' in variability. The functions $d_{\alpha,\beta}(\cdot)$ are then used to create a family of non-stationary kernels through the process $Z_{\alpha,\beta} \equiv Z_{\alpha,\beta}(x) = d_{\alpha,\beta}(x)Y(x)$ where $Y \sim GP(0, K_\lambda)$. Thus Y here is a smooth stationary process and d serves to modulate the amplitude of Y. $Z_{\alpha,\beta}$ is a *bona fide* GP [33] with the following covariance matrix ($D_{\alpha,\beta}$ is a diagonal matrix with entries $d_{\alpha,\beta}(\cdot)$):

$$K(\lambda, \alpha, \beta) = D_{\alpha,\beta} K_\lambda D_{\alpha,\beta}. \qquad (6)$$

For the non-stimulating electrodes, we choose all three kernels K_t, K_e, K_s as $K(\lambda, \alpha, \beta)$ in Eq (6), with separate parameters λ, α, β for each. For the time kernels we use time and t as the relevant covariate (δ in Eq (4) and x in Eq (5)). The case of the spatial kernel is more involved: although we want to impose spatial smoothness, we also need to express the non-stationarities that depend on the distance between any electrode and the stimulating electrode. We do so by making δ represent the distance between recording electrodes, and x represent the distance between stimulating and recording electrodes. Finally, for the stimulus kernel we take stimulus strength a_j as the covariate but we only model smoothness through the Matérn kernel and not localization (i.e. $\alpha, \beta = 0$).

Finally, for the stimulating electrode we use the same method for constructing the kernels K_t^r, K_s^r on each range between breakpoints.

2.4 Algorithm

Now we introduce an algorithm for the joint estimation of A and s, based on the GP model for A. Roughly, the algorithm is divided in two stages: first, the hyperparameter that govern the structure of A have to be found. This is described in 2.4.1. Second, given the inferred hyperparameters we perform the actual inference of A, s given these hyperparameters. This is described in 2.4.2 and 2.4.3. We base our approach on posterior inference for $p(A, s|Y, \theta, \sigma^2) \propto p(Y|s, A, \sigma^2)p(A|\theta)$, where the first factor in the right hand side is the likelihood of the observed data Y given s, A, and the noise variance σ^2, and the second stands for the noise-free artifact prior; $A \sim GP(0, K^\theta)$. A summary of all the involved operations is shown in pseudo-code in algorithm 1.

```
Algorithm 1 Spike detection and Artifact cancellation with electrical stimulation
Input: Traces $Y = (Y_j)_{j=1,\ldots,J}$, in response to $J$ stimuli.
Output: Estimates of artifact $\hat{A}$ and neural activity $\hat{s}^n$ for each neuron.
        EIs of $N$ neurons (e.g. obtained in a visual stimulation experiment).
Initialization
Estimate $\phi^2$ (artifact noise) and $\theta$.        ▷ Hyperparameter estimation, Eq (7)
Also, estimate $\sigma^2$ (neural noise) from traces
Artifact/neural activity inference via coordinate ascent and extrapolation
for $j = 1, \ldots J$ do
    Estimate $A_j^0$ from $A_{[j-1]}$ ($A_1^0 \equiv 0$).        ▷ Extrapolation, Eq (11)
    while some $\hat{s}_{j,i}^n$ change from one iteration to the next do        ▷ Coordinate ascent
        • Estimate $\hat{s}_{j,i}^n$ (for each $i, n$) greedily        ▷ Matching pursuit, Eq (9)
          until no spike addition increases the likelihood.
        • Estimate $\hat{A}_j$ from residuals $Y_j - \sum_{n=1}^N s_j^n$        ▷ Artifact filtering, Eq (10).
    end while
end for
```

2.4.1 Initialization: hyperparameter estimation

From Eqs (2,3, 4) and (6) the GP model for the artifact is completely specified by the hyperparameters $\theta = (\rho, \alpha, \lambda, \beta)$ and ϕ^2. The standard approach for estimating θ is to optimize the marginal likelihood of the observed data Y [30]. However, in this setting computing this marginal likelihood entails summing over all possible spiking patterns s while simultaneously integrating over the high-dimensional vector A; exactly computing this large joint sum and integral is computationally intractable. Instead we introduce a simpler approximation that is computationally relatively cheap and quite effective in practice. We simply optimize the Gaussian likelihood of $\tilde{A}$,

$$\max_{\theta} \log p(\tilde{A}|\theta, \phi^2) = \min_{\theta} \frac{1}{2} \tilde{A}^t \left(K^{(\theta,\phi^2)} \right)^{-1} \tilde{A} + \frac{1}{2} \log \left| K^{(\theta,\phi^2)} \right|, \tag{7}$$

where $\tilde{A}$ is a computationally cheap proxy for the true A. Here, $K^{(\theta,\phi^2)} = K^{\theta} + \phi^2 I_d$ with $K^{\theta} = \rho K_t \otimes K_e \otimes K_s$ for the non-stimulating electrode or $K^{\theta} = \rho K_t \otimes K_e \otimes K_s$ for the stimulating electrode. Due to the Kronecker structure of these matrices, once $\tilde{A}$ is obtained the terms in Eq(7) can be computed quite tractably, with computational complexity $O(d^3)$, with $d = \max\{E, T, J\}$ ($\max\{T, J\}$ in the stimulating-electrode case), instead of $O(\dim(A)^3)$, with $\dim(A) = E \cdot T \cdot J$, in the case of a general non-structured K. Thus the Kronecker assumption here leads to computational efficiency gains of several orders of magnitude. See e.g. [32] for a detailed exposition of efficient algorithmic implementations of all the operations that involve the Kronecker product that we have adopted here; some potential further accelerations are mentioned in the discussion section below.

Now we need to define $\tilde{A}$. The stimulating electrode case is a bit more straightforward here: since the artifact A is much bigger than the effect of spiking activity s on this electrode, the effect of s on data Y recorded at the stimulating electrode can be neglected, and we have found that setting $\tilde{A}$ to the mean or median of Y across trials and then solving Eq(7) leads to reasonable hyperparameter settings. We estimate distinct kernels K_t^r, K_s^r for each stimulating electrode (since from Fig3A we see that there is a good deal of heterogeneity across electrodes), and each of the ranges between breakpoints. Fig 4B shows an example of some kernels estimated following this approach.

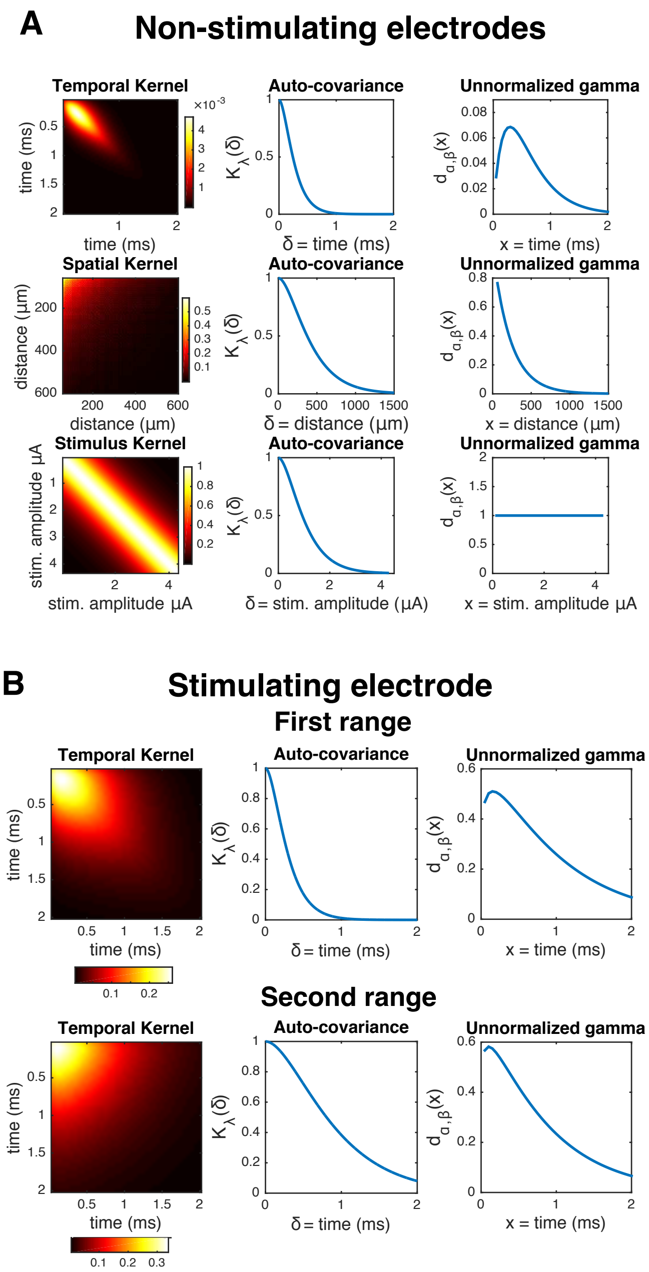

Fig 4. Examples of learned GP kernels. A) *Left*: inferred kernels K_t, K_e, K_s in the top, center and bottom rows, respectively. *Center*: corresponding unnormalized 'gamma-like' envelopes $d_{\alpha,\beta}$ (Eq 5). *Right*: corresponding stationary auto-covariances from the Matérn(3/2) kernels (Eq 4). The inferred quantities are in agreement with what is observed in Fig 3B: first, the shape of temporal term $d_{\alpha,\beta}$ reflects that the artifact starts small, then the variance amplitude peaks at $\sim .5$ ms, and then decreases rapidly. Likewise, the corresponding spatial $d_{\alpha,\beta}$ indicates that the artifact variability induced by the stimulation is negligible for electrodes greater than 700 microns away from the stimulating electrode. **B)** Same as **A**), but for the stimulating electrode. Only temporal kernels are shown, for two inter-breakpoint ranges (first and second rows, respectively).

For non-stimulating electrodes, the artifact A is more comparable in size to the spiking contributions s, and this simple average-over-trials approach was much less successful. On the other hand, for non-stimulating electrodes the artifact shape is much more reproducible across electrodes, so some averaging over electrodes should be effective. We found that a sensible estimate can be obtained by assuming that the effect of the artifact is a function of the position relative to the stimulating electrode. Under that assumption we can estimate the artifact by translating, for each of the stimulating electrodes, all the recorded traces as if they had occurred in response to stimulation at the center electrode, and then taking a big average for each electrode. In other words, we estimate

$$\tilde{A}(e,t,j) = \frac{1}{E}\sum_{e_s=1}^{E}\frac{1}{n_j}\sum_{i=1}^{n_j} Y^{e_s}(\bar{e},t,j), \tag{8}$$

where Y^{s_e} are the traces in response to stimultion on electrode s_e and $\bar{e}$ is the index of electrode e after a translation of electrodes so that s_e is the center electrode. This centered estimate leads to stable values of θ, since combining information across many stimulating electrodes serves to average-out stimulating-electrode-specific neural activity and other outliers.

Some implementation details are worth mentioning. First, we do not combine information of all the E stimulating electrodes, but rather take a large-enough random sample to ensure the stability of the estimate. We found that using ~ 15 electrodes is sufficient. Second, as the effect of the artifact is very localized in space, we do not utilize all the electrodes, but consider only the ones that are close enough to the center (here, the 25% closest). This leads to computational speed-ups without sacrificing estimate quality; indeed, using the entire array may lead to sub-optimal performance, since distant electrodes essentially contribute noise to this calculation. Third, we do not estimate ϕ^2 by jointly maximizing Eq (7) with respect to (θ, ϕ). Instead, to avoid numerical instabilities we estimate ϕ^2 directly as the background noise of the fictitious artifact. This can be easily done before solving the optimization problem, by considering the portions of A with the lowest artifact magnitude, e.g. the last few time steps at the lowest amplitude of stimulation at electrodes distant from the stimulating electrode. Fig 4A shows an example of kernels K_t, K_e, and K_s estimated following this approach.

2.4.2 Coordinate Ascent

Once the hyperparameters θ are known we focus on the posterior inference for A, s given θ and observed data Y. The non-convexity of the set over which the binary vectors b^n are defined makes this problem difficult: many local optima exist in practice and, as a result, for global optimization there may not be a better alternative than to look at a huge number of possible cases. We circumvent this cumbersome global optimization by taking a greedy approach, with two main characteristics: first, joint optimization over A and s is addressed with alternating ascent (over A with s held fixed, and then over s with A held fixed). Second, data is divided in batches corresponding to the same stimulus amplitude, and the analysis for the $(j+1)$-th batch starts only after definite estimates $\hat{s}_{[j]}$ and $\hat{A}_{[j]}$ have already been produced ($[j]$ denotes the set $\{1,\ldots,j\}$). Moreover, this latter estimate of the artifact is used to initialize the estimate for A_{j+1} (intuitively, we borrow strength from lower stimulation amplitudes to counteract the more challenging effects of artifacts at higher amplitudes). We address each step of the algorithm in turn below. For simplicity, we describe the details only for the non-stimulating electrodes. Treatment of the stimulating electrode is almost the same but demands a slightly more careful handling that we defer to 2.4.4.

Given the batch Y_j and an initial artifact estimate A_j^0 (see 2.4.3) we alternate between neural activity estimation $\hat{s}_j$ given a current artifact estimate, and artifact estimation $\hat{A}_j$ given the current estimate of neural activity. This alternating optimization stops when changes in every $\hat{s}^n_j$ are sufficiently small, or nonexistent.

Matching pursuit for neural activity inference. Given the current artifact estimate $\hat{A}_j$ we maximize the conditional distribution for neural activity $p(s_j|Y_j,\hat{A}_j,\sigma^2) = \prod_{i=1}^{n_j} p(s_{j,i}|Y_{j,i},\hat{A}_j,\sigma^2)$, which corresponds to the following sparse regression problem (the set S embodies our constraints on spike occurrence and timing):

$$\min_{b^n_{j,i}\in S, n=1,\ldots,N} \sum_{i=1}^{n_j} \left\| (Y_{j,i} - \hat{A}_j) - \sum_{n=1}^{n} M^n b^n_{j,i} \right\|^2. \tag{9}$$

Intuitively, we seek to find the allocation of spikes that will lead the best match with the residuals $(Y_{j,i} - \hat{A}_j)$, leading to the smallest sum of squares. We use a standard greedy matching pursuit approach [12, 34] to locally optimize Eq(9).

Filtering for artifact inference. Given the current estimate of neural activity $\hat{s}_j$ we maximize the conditional of the artifact, that is, $\max_{A_j} p(A_j|Y_j,\hat{s}_j,\theta,\sigma^2)$, which here leads to the posterior mean estimator (again, the overline indicates mean across the n_j trials):

$$\hat{A}_j = E(A_j|Y_j,\hat{s}_j,\theta,\sigma^2,\phi^2) = K^{\theta}_{j,j}\left(K_{j,j}^{(\theta,\frac{\sigma^2}{n_j}+\phi^2)}\right)^{-1}(\bar{Y}_j - \bar{\hat{s}}_j). \tag{10}$$

This operation can be understood as the application of a linear filter. Indeed, by appealing to the eigendecomposition of $K_{j,j}^{(\theta,\sigma^2/n_j+\phi^2)}$ we see this operator shrinks the m-th eigencomponent of the artifact by a factor of $\kappa_m/(\kappa_m + \sigma^2/n_j + \phi^2)$ (κ_m is the m-th eigenvalue of $K_{j,j}^{(\theta,\sigma^2/n_j+\phi^2)}$), exerting its greatest influence where κ_m is small. Notice that in the extreme case that $\sigma^2/n_j + \phi^2$ is very small compared to the κ_m then $\hat{A}_j \approx (\bar{Y}_j - \bar{\hat{s}}_j)$.

Convergence. Remarkably, often only a few (e.g. 3) iterations of coordinate ascent (neural activity inference and artifact inference) are required to converge to a stable solution $(s^n_j)_{\{n=1,\ldots N\}}$. However, we stress this number can vary, depending e.g. on the number of neurons or the signal-to-noise (EI strength versus noise variance).

2.4.3 Iteration over batches and artifact extrapolation

The procedure described in 2.4.2 is repeated in a loop that iterates through the batches corresponding to different stimulus strengths, from the lowest to the highest. Also, when doing $j \rightarrow j+1$ an initial estimate for the artifact A^0_{j+1} is generated by extrapolating from the current, faithful, estimate of the artifact up to the j-th batch. This extrapolation is easily implemented as the mean of the noise-free posterior distribution in this GP setup, that is:

$$A^0_{j+1} = E(A_{j+1}|\hat{A}_{[j]}\theta,\phi^2) = K^{\theta}_{(j+1,[j])}\left(K^{(\theta,\phi^2)}_{([j],[j])}\right)^{-1}\hat{A}_{[j]}. \tag{11}$$

Importantly, in practice this initial estimate ends up being extremely useful, as in the absence of a good initial estimate, coordinate ascent often leads to poor optima. The very accurate initializations from extrapolation estimates help to avoid these poor local optima (see Fig 8).

We note that both for the extrapolation and filtering stages we still profit from the scalability properties that arise from the Kronecker decomposition. Indeed, the two required operations — inversion of the kernel and the product between that inverse and the vectorized artifact — reduce to elementary operations that only involve the kernels K_e, K_t, K_s [32].

2.4.4 Integrating the stimulating and non-stimulating electrodes

Notice that the same algorithm can be implemented for the stimulating electrode, or for all electrodes simultaneously, by considering equivalent extrapolation, filtering, and matched pursuit operations. The only caveat is that extrapolation across stimulation amplitude breakpoints does not make sense for the stimulating electrode, and therefore, information from the stimulating electrode must not be taken into account at the first amplitude following a breakpoint, at least for the first matching pursuit-artifact filtering iteration.

2.4.5 Further computational remarks

Note the different computational complexities of artifact related operations (filtering, extrapolation) and neural activity inference: while the former depends (cubically) only on T, E, J, the latter depends (linearly) on the number of trials n_j, the number of neurons, and the number of electrodes on which each neuron's EI is significantly nonzero. In the data analyzed here, we found that the fixed computational cost of artifact inference is typically bigger than the per-trial cost of neural activity inference. Therefore, if spike sorting is required for big volumes of data ($n_j \gg 1$) it is a sensible choice to avoid unnecessary artifact-related operations: as artifact estimates are stable after a moderate number of trials (e.g. $n_j = 50$), one could estimate the artifact with that number, subtract that artifact from traces and perform matching pursuit for the remaining trials. That would also be helpful to avoid unnecessary multiple iterations of the artifact inference - spike inference loop.

3 Results

We start by showing, in Fig 5, an example of the estimation of the artifact A and spiking activity s from single observed trials Y. Here, looking at individual responses to stimulation provides little information about the presence of spikes, even if the EIs are known. Thus, the estimation process relies heavily on the use of shared information across dimensions: in this example, a good estimate of the artifact was obtained by using information from stimulation at lower amplitudes, and from several trials.

3.1 Algorithm validation

We validated the algorithm by measuring its performance both on a massive dataset with available human-curated spike sorting and with ground-truth simulated data (we avoid the term ground-truth in the real data to acknowledge the possibility that the human makes mistakes). Regarding the latter, simulations with synthetic data constitute a powerful methodology for strengthening the findings obtained through comparison with human results: despite the obvious downside — simulations are based on a simplified version of reality that does not represent all the phenomena — by the use of simulations we are able to craft specific scenarios that provide crucial tests for the algorithm, and to determine to what extent its different features are responsible for performance. Also, unlike with human spike sorting data, simulated ground truth is error-free.

3.1.1 Comparison to human annotation

The efficacy of the algorithm was first demonstrated by comparison to human-curated results from the primate retina. The available dataset was heterogeneous, coming from nine different 512-electrode recordings from retinal preparations obtained in different

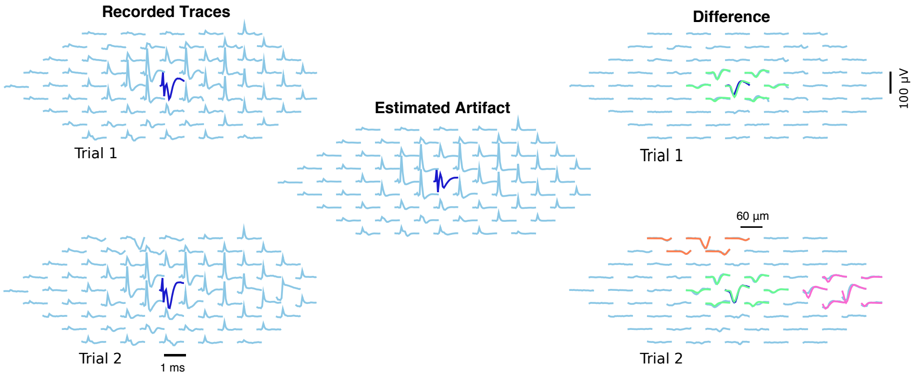

Fig 5. Example of neural activity and artifact inference in a neighborhood of the stimulating electrode. *Left:* Two recordings in response to a 2.01 μA stimulus. *Center:* estimated artifact (as the stimulus doesn't change, it is the same for both trials). *Right:* Difference between raw traces and estimated artifact, with inferred spikes in color. In one case (above) three spiking neurons were detected, while in the other (below) there was only one. The algorithm separates the artifact A and spiking acitivity s effectively here.

experiments. In total, 827 stimulating electrode and neuron pairs were available, with multiple trials at multiple stimulation amplitudes for each pair; we hereafter refer to each dataset from a single stimulating electrode and neuron pair as an *amplitude series*. These 827 *amplitude series* gave rise to 704,984 trials with available human results after combining all analyzed neurons, stimulating electrodes, stimulus amplitudes and preparations. We refer the reader to the supporting information for details on both experimental protocols and further information about the retinal preparations.

We assessed the agreement between algorithm and human annotation based on two types of comparison. The first and most elementary was on a trial-by-trial basis, by comparing presence or absence of spikes (and their latencies). Although this provides a good first-order account of algorithm performance, it can conceal more complex scenarios: for example, in cases where the human indicates that a neuron gets suddenly activated at the highest amplitude of stimulation (i.e, spiking has high probability only at that highest amplitude but very low otherwise), accuracy could still be very high if the algorithm detects no spiking at all, while in reality, it completely failed to detect the onset of neural activation.

The above stresses the need to also make comparisons based on the presence or absence of neural activation, analyzing responses from the entire *amplitude series*, instead of individual trials. In detail, given an *amplitude series* we conclude that neural activation is present if the sigmoidal activation function fit (specifically, the CDF of a normal distribution) to the empirical activation curves —the proportion of trials where spikes occurred as a function of stimulation amplitude — exceeds 50% within the ranges of stimulation. In the positive cases, we define the stimulation threshold as the current needed to elicit spiking with 0.5 probability. This number provides an informative univariate summary of the activation curve itself.

Given either of the two above types of comparisons (spiking on trials or activation on *amplitude series*), the algorithm's inferences were compared to human annotation, and the usual three types of errors measurements were considered: false negative (FN) rate — the proportion of failures in detecting truly existing spikes— false positive (FP) rate — the proportion of misidentified spikes over the cases of no spiking — and error rate — a weighted average of the two previous, with the weights being the proportion of trials with and without spikes, respectively. Our baseline was a simple reference method:

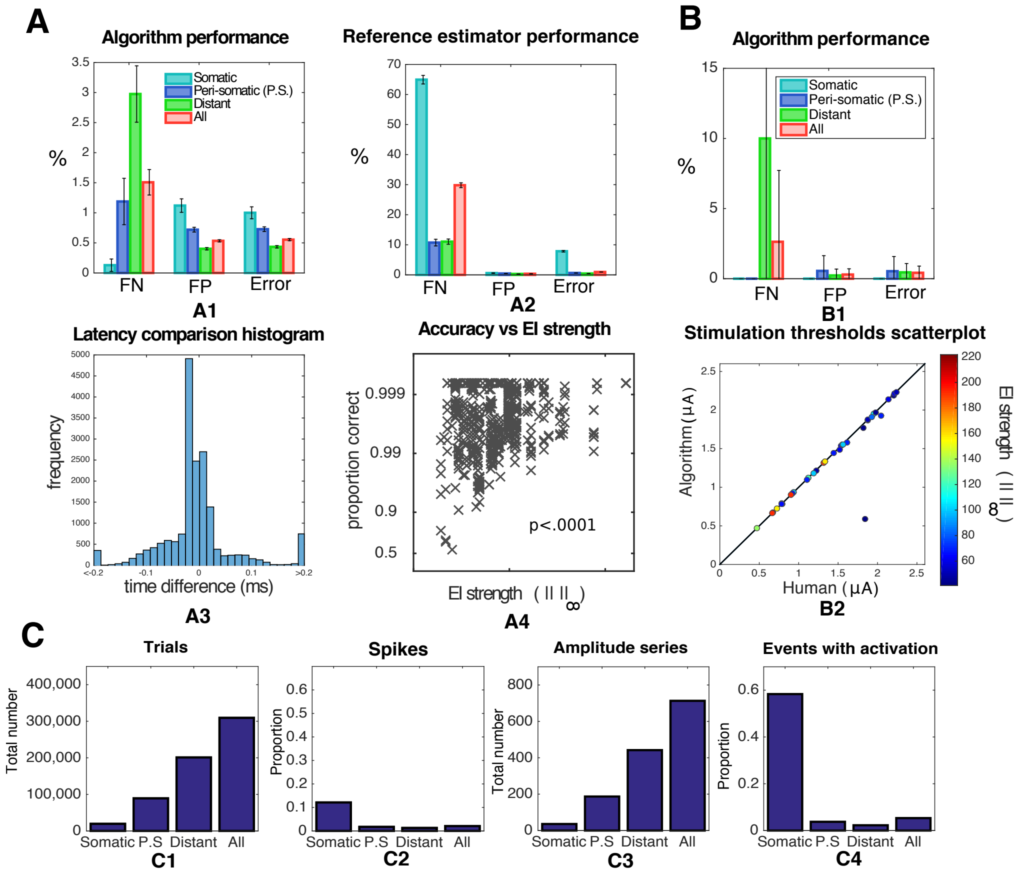

Fig 6. Population results from nine retinal preparations reveal the efficacy of the algorithm A) trial-by-trial analysis. *A1-A2*) performance measures for the algorithm and a reference estimator, broken down by distance between neuron and stimulating electrode. *A3*) for the true positives, histogram of the differences of latencies between human and algorithm. *A4*) measured accuracies (log scale) as a function of strength of the EI (using the $||\cdot||_\infty$ norm). A spearman correlation test revealed a significant positive correlation. **B)** *amplitude series* based analysis. *B1*) performance measures broken down by distance between neuron and stimulating electrode. *B2)* for the true positives, scatterplot of activation thresholds inferred by human and algorithm (the scale of colors represents the strength of the EI) **C)** population statistics: for the trial-by-trial analysis, total number of trials *C1*) and frequency of spiking *C2*). For the *amplitude series* based analysis, number of *amplitude series C3*) and frequency of activation *C4*).

it estimates the artifact as the mean of traces across trials, and after subtracting that estimate from the traces it looks for spikes by greedy template matching, as in Eq (9) (note that this simple baseline estimator typically fails when the stimulus amplitude is high enough to drive spiking with high probability, since then subtracting off the trial-averaged response will usually also subtract away neural activity, leading to false negatives).

Results of the trial-by-trial based analysis are shown in Fig 6A. Overall, they are satisfactory with error rates bounded by 1%, an order of magnitude smaller than of the reference estimator, which suffered from high error rates due to many false negatives. We investigated two covariates that could modulate performance: distance between targeted neuron and stimulating electrode, and strength of the neural signals (EI).

Regarding the former, we divided data by somatic stimulation (stimulating electrode is the closest to the soma), peri-somatic stimulation (stimulating electrode neighbors the closest electrode to the soma) and distant stimulation (neither somatic nor peri-somatic). As expected, accuracy was the lowest when the neural soma is close to the stimulating electrode (somatic stimulation), presumably a consequence of artifacts of larger magnitude in that case. Regarding the latter, we found that accuracy increases with strength of the EI, indicating that our algorithm benefits from strong neural signals. Finally, we also compared the latencies of correctly identified spikes, finding that big discrepancies were rare, and that in the vast majority of cases (>95%) inferred spike times were shifted by less than 0.1 ms.

Similarly, results of the *amplitude series* based analysis are shown in Fig 6B. We still obtained satisfactory results although this time the error rates were slightly higher. However, notice that the number of available events here was much smaller — a fact reflected in the larger error bars. Also, in the case of correctly detected events we compared the activation thresholds (Fig 6B2) and found little discrepancy between human and algorithm (with the exception of one outlier, which happened to have the smallest EI of the cells shown here and was therefore a particularly challenging example; we discuss this outlier at more length below in section 4.3.4).

3.1.2 Simulations

Synthetic datasets were generated by adding artifacts measured in the presence of TTX (to eliminate spiking activity s), real templates, and white noise, in an attempt to faithfully match basic statistics of neural activity in response to electrical stimuli, i.e., the frequency of spiking and latency distribution as a function of distance between stimulating electrode and neurons (see supporting information for details). These simulations were aimed to determine the extent to which the algorithm's main features were necessary. Specifically, two main operations arise from the use of the GP modeling framework: kernel-based artifact filtering (Eq 10) and extrapolation (Eq 11). It is not obvious that those features are actually needed; perhaps, similar or better results could be obtained if those operations were avoided or replaced by simpler, less computationally expensive ones. To address this issue, we considered both the omission and simplification of the filter (Eq 10), and the replacement of the kernel-based extrapolation (Eq 11) by a naive extrapolation estimator that guesses the artifact at the j-th amplitude of stimulation simply as the artifact at the $j-1$ amplitude of stimulation.

As the number of trials n_j goes to infinity, or as the noise level σ goes to zero, the influence of the likelihood grows compared to the GP prior, and the filtering operator converges to the identity (see Eq 10). However, applied on individual traces, where the influence of this operator is maximal, filtering removes high frequency noise components and variations occurring where the localization kernels do not concentrate their mass (Fig 4A), which usually correspond to spikes. Therefore, in this case filtering should lead to less spike-contaminated artifact estimates. Fig 7B confirms this intuition with results from simulated data: in cases of high σ^2 and small n_j the filtering estimator led to improved results. Moreover, a simplified filter that only consisted of smoothing kernels (i.e. for all the spatial, temporal and amplitude-wise kernels the localization terms $d_{\alpha,\beta}$ in Eq 5 were set equal to 1, leading to the Matérn kernel in Eq 4) led to more modest improvements, suggesting that the localization terms (Eq 5) — and not only the smoothing kernels — are a sensible and helpful modeling choice.

Likewise, we expect that kernel-based extrapolation leads to improved performance if the artifact magnitude is large compared to the size of the EIs: in this case, differences between the naive estimator and the actual artifact would be large enough that many spikes would be misidentified or missed. However, since kernel-based extrapolation

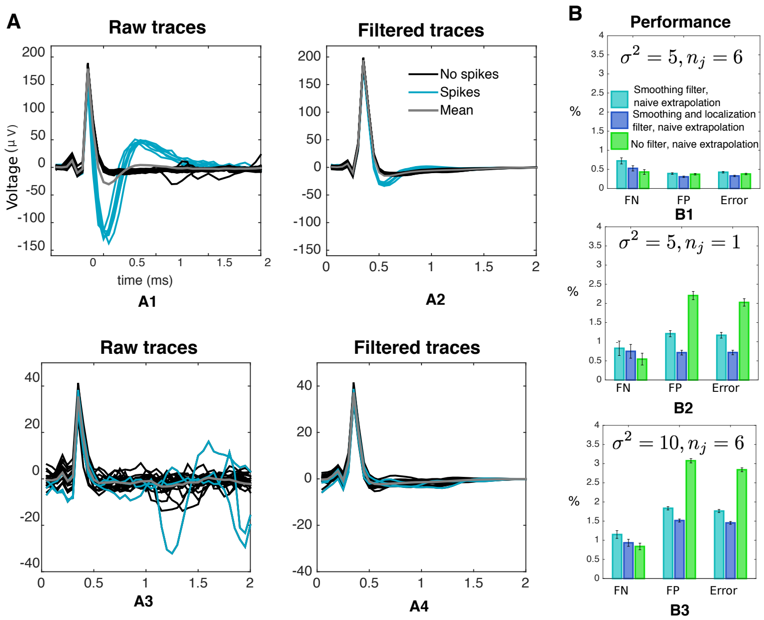

Fig 7. Filtering (Eq 10) leads to a better, less spike-corrupted artifact estimate in our simulations. A) effect of filtering on traces for two non-stimulating electrodes, at a fixed amplitude of stimulation ($2.2\mu A$). *A1,A3*) raw traces, *A2,A4*) filtered traces. Notice the two main features of the filter: first, it principally affects traces containing spikes, a consequence of the localized nature of the kernel in Eq (2). Second, it helps eliminate high-frequency noise. **B)** through simulations, we showed that filtering leads to improved results in challenging situations. Two filters — only smoothing and localization + smoothing — were compared to the omission of filtering. In all cases, to rule out that performance changes were due to the extrapolation estimator, extrapolation was done with the naive estimator. *B1*) results in a less challenging situation. *B2*) results in the heavily subsampled ($n_j = 1$) case. *B3*) results in the high-noise variance ($\sigma^2 = 10$) case.

produces better artifact estimates (see Fig 8A-B), the occurrence of those failures should be diminished. Indeed, Fig 8C shows that better results are attained when the size of the artifact is multiplied by a constant factor (or equivalently, neglecting the noise term σ^2, when the size of the EIs is divided by a constant factor). Moreover, the differential results obtained when including the filtering stage suggest that the two effects are non-redundant: filtering and extrapolation both lead to improvements and the improvements due to each operation are not replaced by the other.

3.2 Extension: analysis of responses to two-electrode stimulation

So far we have focused our attention on the case that only one stimulating electrode is active at a time. Next we examined the generalization to two-electrode stimulation data. TTX experiments indicate that responses to two-electrode stimulation are well explained by the linear combination of their corresponding single-electrode stimulation counterparts (Fig 9). Therefore, given responses to two-electrode stimulation, one can consider as an initial estimate of the artifact the linear sum of the artifacts that result from stimulation at each single electrode, and subtract this estimate from the raw traces. As the resulting traces now have a diminished artifact magnitude (see Fig 9B),

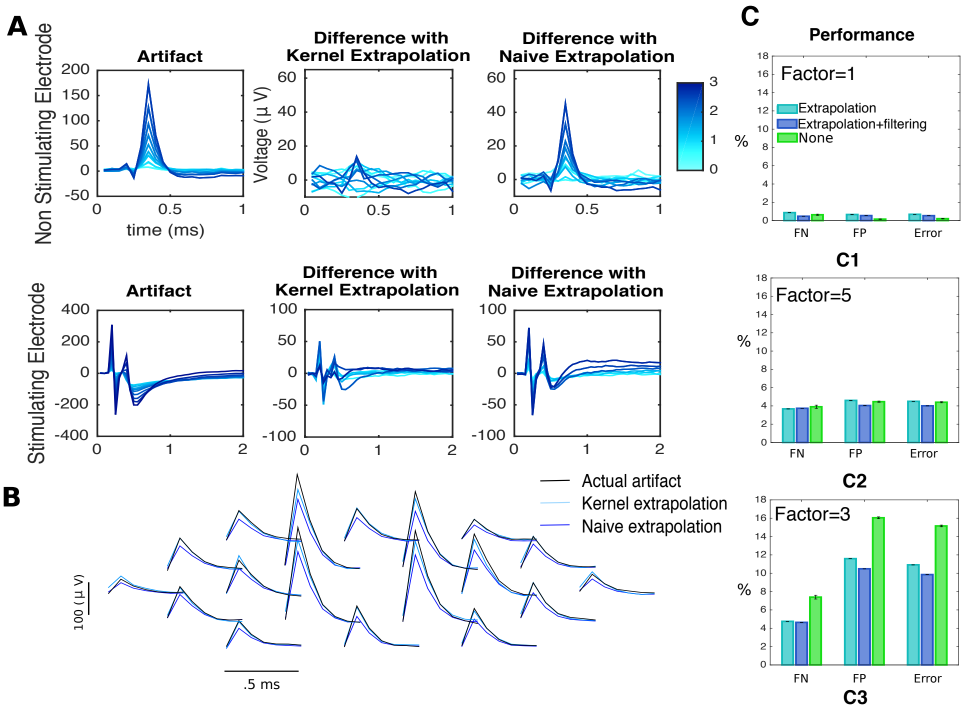

Fig 8. Kernel-based extrapolation (Eq 11) leads to more accurate initial estimates of the artifact. A) comparison between kernel-based extrapolation and the naive estimator, the artifact at the previous amplitude of stimulation. For a non-stimulating (first row) and the stimulating (second row) electrode, left: artifacts at different stimulus strengths (shades of blue), center: differences with extrapolation estimator (Eq 11), right: differences with the naive estimator. **B)** comparison between the true artifact (black), the naive estimator (blue) and the kernel-based estimator (light blue) for a fixed amplitude of stimulus ($3.1\mu A$) on a neighborhood of the stimulating electrode (not shown). **C)** Through simulations we showed that extrapolation leads to improved results in a challenging situation. Kernel-based extrapolation was compared to naive extrapolation. *C1*) results in a less challenging situation. *C2-C3*) results in the case where the artifact is multiplied by a factor of 3 and 5, respectively.

they are more amenable for treatment using simple methods. Indeed, we found that spike detection using the naive extrapolation estimator described in 3.1 for these artifact subtracted traces leads to discrepancies with human-curated inferred spikes (Fig 9C) of the order of 0.5% (comparable to responses to single electrode stimulation).

3.3 Applications: high resolution neural prosthesis

A prominent application of our method relates to the development of high-resolution neural prostheses (particularly, epiretinal prosthesis), whose success will rely on the ability to elicit arbitrary patterns of neural activity through the selective activation of individual neurons in real-time [28, 35, 36]. For achieving such selective activation in a closed-loop setup, we need to know how different stimulating electrodes activate nearby neurons, information that is easily summarized by the activation curves, or even the activation thresholds. Unfortunately, obtaining this information in real time — as required for prosthetic devices — is currently not feasible since estimation of thresholds requires the analysis of individual responses to stimuli.

Figures 10,11,12 and 13 show pictorial representations of different features of the

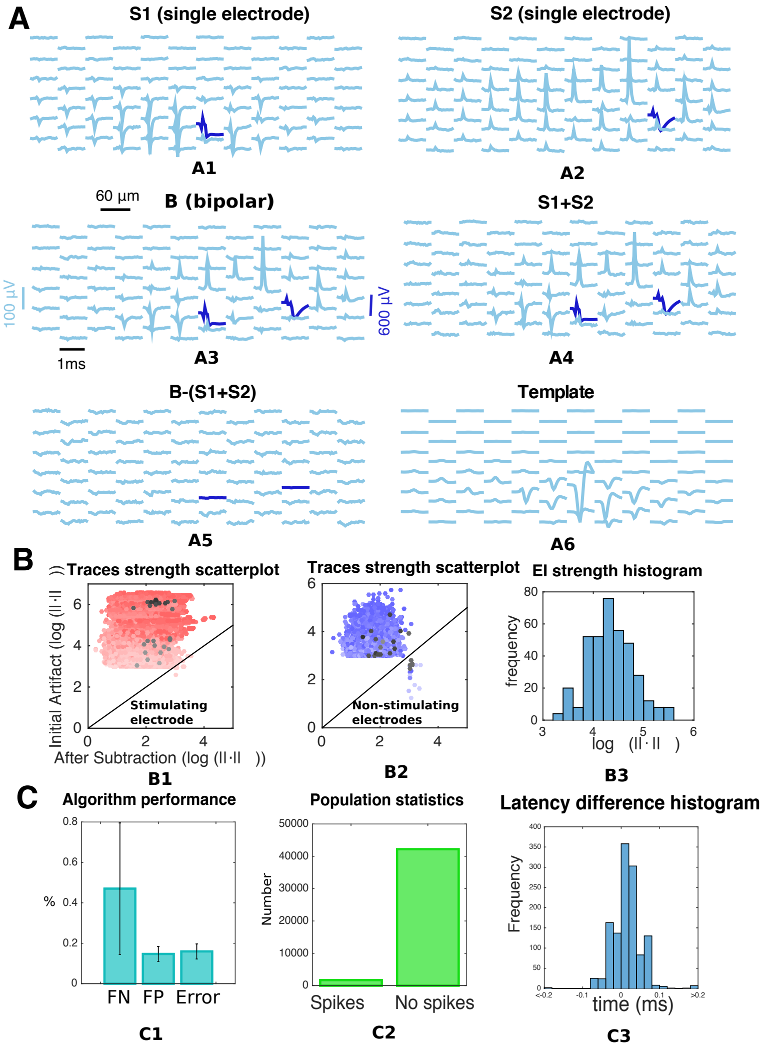

Fig 9. Analysis of responses in the two-electrode stimulation case reduces to the analysis of single-electrode stimulation. A) example of observed linearity: *A1-A2*) artifacts for single electrode stimulation at two different stimulating electrodes with same strength (3.1 μ A) and opposite polarities. *A3*) corresponding two-electrode stimulation. *A4*) sum of *A1*) and *A2*). *A5*) difference between *A3*) and *A4*). *A6*) for reference, the EI of a typical neuron in shown in the same scale. **B)** population-based generalization of the finding in **A)** from thousands of stimulating electrode pairs, collapsing stimulating amplitudes and electrodes. *B1-B2*) scatterplots of the log $|| \cdot ||_\infty$ norm for two-electrode stimulation artifacts at different stimulus strengths (strength of the color) before and after subtracting the sum of single electrode artifacts. Points in the gray-scale are the ones shown in **A)**. In the vast majority of cases (99%, points above the diagonal) subtracting the linear sum of individual artifacts is a sensible choice as it decreases the strength of the artifacts (histogram in *B3*). **C)** Algorithm's performance in the two-electrode stimulation dataset. Data comes from a single preparation with $n = 43,890$ responses to stimulation of twelve neurons. Here, the filtering stage was omitted and extrapolation was done using the naive estimator. Error rates are bounded by 0.5%.

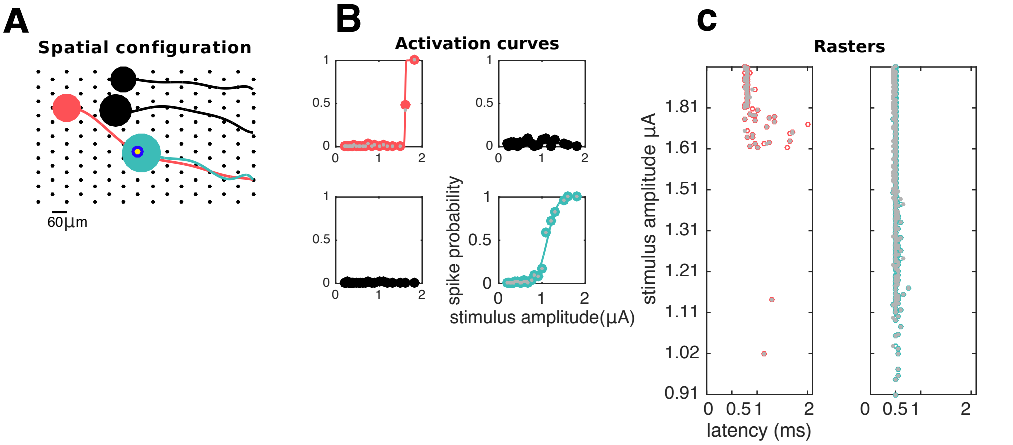

Fig 10. Analysis of responses of neurons in a neighborhood of the stimulating electrode. A) Spatial configuration: stimulating electrode (blue/yellow annulus) and four neurons on its vicinity. Soma of green neuron and axon of pink neuron overlap with stimulating electrode. **B)** Activation curves (solid lines) along with human-curated and algorithm inferred spike probabilities (gray and colored circles, respectively) of all the four cells. Stimulation elicited activation of green and pink neurons; however, the two other neurons remained inactive. **C)** Raster plots for the activated cells, with responses sorted by stimulation strength in the y axis. Human and algorithm inferred latencies are in good agreement (gray and colored circles, respectively). Here, direct somatic activation of the green neuron leads to lower-latency and lower-threshold activation than of the pink neuron, which is activated through its axon.

results obtained with the algorithm, and their comparison with human annotation. Each of these figures provides particular insights to inform and guide the large-scale closed-loop control of the neural population. Importantly, generation of these maps took only minutes on a personal computer, compared to many human hours, indicating feasibility for clinical applications and substantial value for analysis of laboratory experiments [28, 36].

Figure 10 focuses on the stimulating electrode's point of view: given stimulation in one electrode, it is of interest to understand which neurons will get activated within the stimulation range, and how selective that activation can be made. This information is provided by the activation curves, i.e, their steepness and their associated stimulation thresholds. Additionally, latencies can be informative about the spatial arrangement of the system under study, and the mode of neural activation: in this example, one cell is activated through direct stimulation of the soma, and the other, more distant cell is activated through the indirect and antidromic propagation of current through the axon [37]. This is confirmed by the observed latency pattern.

Figure 11 depicts the converse view, focusing on the neuron. Here we aim to determine the cell's electrical receptive field [38, 39] to single-electrode stimulation; that is, the set of electrodes that are able to elicit activation, and in the positive cases, the corresponding stimulation thresholds. These fields are crucial for tailoring stimuli that selectively activate sub-populations of neurons.

Figure 12 shows how the algorithm enables the analysis of two-electrode stimulation; particularly, the study of differential patterns of activation due to the additional stimulation of a neighboring electrode with a current of the same strength of different polarity (bipolar stimulation). This strategy has been suggested to enhance selectivity [40], by differentially shifting the stimulation thresholds of the cells so the

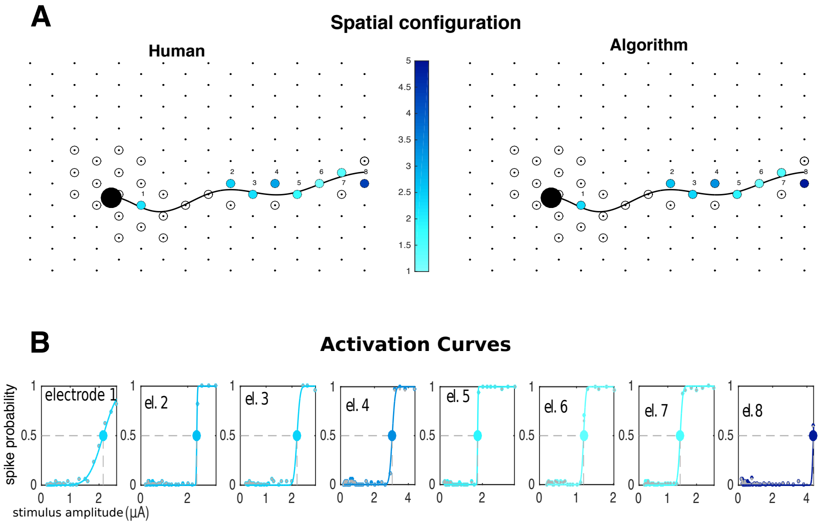

Fig 11. Electrical receptive field of a neuron. **A)** spatial representation of the soma (black circle) and axon (black line) over the array. Electrodes where stimulation was attempted are represented by circles, with colors indicating the activation threshold in the case of a successful activation of the neuron within the stimulation range. **B)** For those cases, activation curves (solid lines) are shown along with with human and algorithm inferred spike frequencies (gray and colored circles, respectively). Large circles indicate the activation thresholds represented in **A)**. In this case, much of the activity is elicited through axonal stimulation, as there is a single electrode close to the soma that can activate the neuron. Human and algorithm are in good agreement.

range of currents that lead to activation of a single cell is widened.

Finally, figure 13 shows a large-scale summary of the responses to single-electrode stimulation. There, a population of ON and OFF parasol cells was stimulated at many different electrodes close to their somas, and each of those cells was then labeled by the lowest achieved activation threshold. These maps provides a proxy of the ability to activate cells with single-electrode stimulation, and of the different degrees of difficulty in achieving activation. Since in many cases only as few as 20% of the neurons can be activated [41], the information of which cells were activated can provide a useful guide for the on-line development of more complex multiple electrode stimulation patterns that activate the remaining cells.

4 Discussion

Now we discuss the main features of the algorithm in light of the results and sketch some extensions to enable the analysis of data in contexts that go beyond those analyzed here.

4.1 Simplifications

In 3.1 we considered estimators that arose from the omission and/or simplification of the filtering and extrapolation stages. Although we showed that the full method provided better results in stressed situations (e.g. sub-sampled, high variance and large artifact regimes), in non-stressed cases both methods achieved good performance.

It is worth noting that the “naive" extrapolation estimator can also be understood in terms of the GP framework, as it corresponds to the predicted artifact if a Brownian

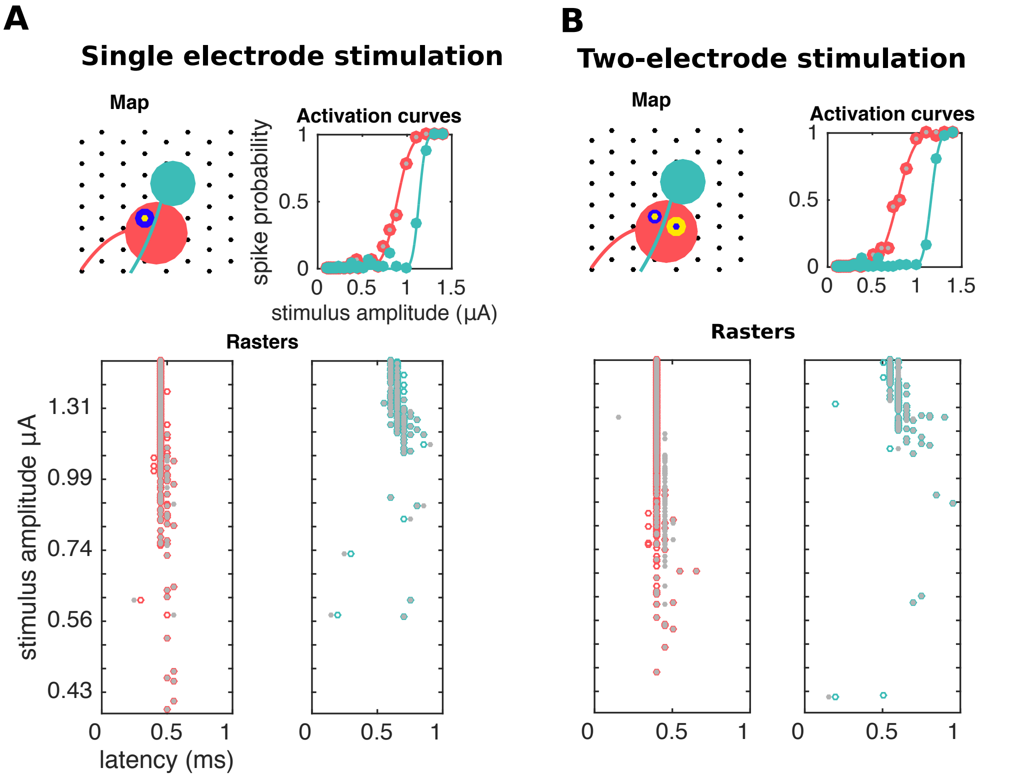

Fig 12. Analysis of differential responses to single A) and two-electrode B) stimulation. Gray and colored dots indicate human and algorithm inferences, respectively. In both cases activation of the two neurons is achieved. However, shape of activation curves is modulated by the presence of a current with the same strength and opposite polarity in a neighboring electrode (yellow/blue annulus in **B**): indeed, in this case bipolar stimulation leads to an enhanced ability to activate the pink neuron without activating the green neuron. The algorithm is faithfully able to recover the relevant activation thresholds.

motion kernel is assumed for the stimulus coordinate. Therefore the GP framework has allowed us to introduce an entire family of estimators that exhibit a certain trade-off between computational complexity and performance. In practice, it will be a task for the experimenter to decide which kernel is most suitable for a given application. For example, regarding the number of trials n_j we have showed that the full model can significantly improve experimental capability: since experiments are performed in living systems, experimental time is limited. The ability to analyze fewer trials without loss of accuracy (using the kernel-based estimators) opens up the possibility for new experimental designs that may not have been otherwise feasible. However, the simplified method can be applied to datasets with large n_j if desired.

4.2 Comparison to other methods

We avoided explicit comparisons to other methods because none of the existing techniques to remove stimulation artifacts fit our applications. Specifically, we needed to detect spikes elicited by electrical stimulation which is characterized by sub-ms latencies. However, [20–22] all present methods to detect spikes with latencies greater or equal to 2 ms. Further, the method of [22] requires that the artifact duration remains smaller than the duration of a typical neural response, which is not the case for our

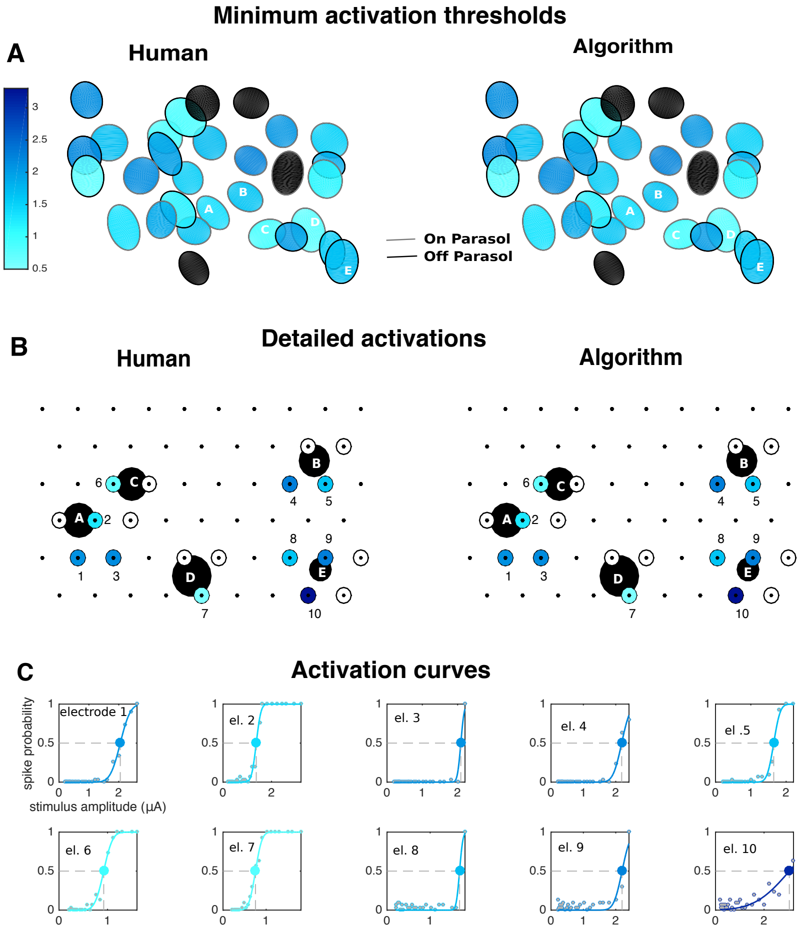

Fig 13. Large-scale analysis of the stimulation of a population of parasol cells. For each neuron, one or more stimulating electrodes in a neighborhood of neural soma were chosen for stimulation. **A)** Receptive fields colored by the lowest achieved stimulation threshold (black if activation was not achieved). **B)** Inferred somas (big black circles) of the neurons labeled A-E in **A**), showing which electrodes were chosen for stimulation (small circles) and whether activation was achieved (colors). **C)** Activation curves (solid lines) of the neurons in **B)** for the successful activation cases. Gray and colored dots represent human and algorithm results, respectively, and large circle indicates stimulation thresholds.

data. Low-latency neural activity tends to be synchronized in time across trials, and so if it occurs in >50% of trials, is easily confused with electrical artifact. Existing methods fail to detect low-latency spikes primarily because they estimate artifacts in a way that combines neural activity with true electrical artifact, either by using locally fitted cubic polynomials to model the artifact ([21]) or by simple averaging over all trials ([20]). Both methods inadvertently subtract neural information at low-latencies when subtracting the modeled artifact. Indeed, the method in [20] corresponds —up to subtle differences — to the reference estimator we included in section 3.1; therefore, this reference estimator can be deemed as a proxy for the performance of existing techniques.

4.3 Extensions

4.3.1 Beyond the retina: absence of available electrical images

We stress the generalizability of our method to neural systems beyond the retina, as we expect that the qualitative characteristics of this artifact, being a general consequence of the electrical interactions between the neural tissue and the MEA [16], is replicable up to different scales that can be accounted for by appropriate changes in the hyperparameters.

In this work we have assumed that the electrical images (EIs) of the spiking neurons are available. If this is not the case, we propose stimulation at low amplitudes so that the elicited cell activity is variable and therefore an initial crude estimate of the artifact can be initalized by the simple median over many repetitions of the same stimulus. Then, after artifact subtraction EIs could be estimated with standard spike sorting approaches.

More generally, this additional EI estimation step could be stated in terms of an outer loop that iterates between EI estimation, given current artifact and neural activity estimates, and neural activity and artifact estimation given the current EI estimate — that is, our algorithm. This outer loop would be especially helpful to deal with EI mis-specification due to biases in EI estimation [12], and to enable the online update of the EI in order to counteract the effect of tissue drift [42, 43], which could lead to problematic changes in EI shape over the course of an experiment.

4.3.2 Accounting for correlated noise

We assumed that the noise process (ϵ) was uncorrelated in time and across electrodes, and had a constant variance. This is certainly an overly crude assumption: noise in recordings does exhibit strong spatiotemporal dependencies [12, 44], and methods for properly estimating these structured covariances have been proposed [12, 45]. To relax this assumption we can consider an extra, pre-whitening stage in the algorithm, where traces are pre-multiplied by a suitable whitening matrix. This matrix can be estimated by using stimulation-free data (e.g. while obtaining the EIs) as in [12]. In the case of the data considered here this pre-whitening step did not lead to improvements in artifact estimation accuracy.

4.3.3 Saturation

Amplifier saturation is a common problem in electrical stimulation systems [14, 16, 19], and arises when the actual voltage (comprising artifacts and neural activities) exceeds the saturation limit of the stimulation hardware. Although in this work we have considered stimulation regimes that did not lead to saturation, we emphasize that our method would be helpful to deal to saturated traces as well: indeed, in opposition to naive approaches that would lead to no other choice than throwing away entire saturated recordings, our model-based approach enables a more efficient treatment of saturation-corrupted data. We can understand this problem as an example of inference in the context of partially missing observations, for which methods are already available in the GP framework [31].

4.3.4 Automatic detection of failures and post-processing

Since errors cannot be fully avoided, in order to enhance confidence in neural activity estimates provided by the algorithm, we propose to consider diagnostic measures to flag suspicious situations that could be indicative of an algorithmic failure. We consider two measures that arise from a careful analysis of the underlying causes of discrepancies between algorithm and human annotation.

The first comes from the activation curves: at least in the retina, it has been widely documented that these be smoothly increasing functions of the stimulus strength [25, 35]. Therefore, deviations from this expected behavior — e.g., non-smooth activation curves characterized by sudden increases or drops in spiking probability — are indicative of potential problems. For example, the outlier cell in Fig 6B2 is a clear case of an incorrectly inferred sudden increase of spiking from one stimulus amplitude to the next (not shown). Therefore, the application of this simple post-processing criterion would mark this cell for revised analysis, even without the comparison to human annotation here.

The second relates to the residuals, or the difference between observed data and the sum of artifact and neural activity. Cases where those residuals are relatively large could indicate a failure in detecting spikes, perhaps due to a mismatch between a mis-specified EI and observed data.

In either case, these diagnostic measures can be implemented as an automatic procedure based on goodness-of-fit statistics (e.g. the deviance [46]).

4.3.5 Larger and denser arrays, multi-electrode stimulation

In this work the computationally limiting factor is E, the number of electrodes, as this dominates the (cubic) computational time of the GP inference steps. Recent advances in the scalable GP literature [47–49] should be useful for extending our methods to even larger arrays as needed; we plan to pursue these extensions in future work.

In addition to these computational enhancements, we note that an extension to denser arrays would also require a careful revision of the current model: indeed, preliminary results with denser arrays ($30\mu m$ spacing between electrodes, not shown) revealed that due to the increased proximity between the stimulating electrode and its neighboring electrodes, those electrodes also possessed large artifacts and were subject to the effect of breakpoints. Thus a reasonable path forward it to consider one model for the stimulating electrode and its neighbors (instead of a model for the stimulating electrode solely) and a separate model for the rest. Then, both models could be integrated into a single algorithm using a similar strategy as the one developed in this paper.

Finally, following 3.2, we suggest an obvious extension to be explored: the analysis of responses to stimulation in many electrodes. Although promising, in that case special care would have to be taken to guarantee that the interactions between artifacts induced by stimulation in each electrode remain linear.

4.3.6 Online data analysis, closed-loop experiments,

The present findings open a real possibility for the development of closed-loop neural stimulation experiments [10, 50] featuring online data analysis at a much larger scale than was previously possible. A straightforward modification of the algorithm would be particularly useful for online contexts: if artifact estimates are already good there is no need to alternate between artifact and neural inference. Therefore, one can use a subset of the data to estimate the artifact, then stop the artifact updates and just infer artifact-subtracted neural activity as new data come in.

Additionally, we propose parallelization as the main mechanism to obtain further computational speed-ups needed for closed-loop experimentation. Specifically, current processing times are in the order of 30 seconds per amplitude series (with $J \approx 35, n_j \approx 50$) in a 2.2 GHz Intel Core i7 (quadcore) personal computer. Since all the relevant signals are derived from the analysis of responses to the totality of the stimulating electrodes (or multi-electrode stimulation patterns), and since after parameter learning (Eq 7) each of the amplitudes series can be analyzed in parallel, the overall processing time could be heavily amortized if a suitable computational architecture (e.g. GPU, cloud computing) was used. Finally, we note the additional parallelization potential that could be exploited in the neural activity inference stage of the algorithm (Eq 9): inference of activity on each trial of a given amplitude of stimulation can be analyzed separately since it does not depend on inferred activities for the $n_j - 1$ remaining trials.

5 Conclusion

We have developed a method to automate spike sorting in electrical stimulation experiments using large MEAs, where artifacts are a concern. We believe our developments will be useful to enable closed-loop neural stimulation at a much larger scale than was previously possible, and to enhance the ability to actively control neural dynamics. Also, our algorithm has the potential to constitute an important computational substrate for the development of future neural prostheses, particularly epiretinal prostheses. Code is available from the first author upon request.

6 Acknowledgments

LEG received funding from NIH Grant 1F32EY025120. EJC received funding from NEI Grant EY021271. LP received funding by NSF BIGDATA IIS 1546296. JPC received funding from Sloan Foundation and the McKnight Foundation fellowships. We thank David Blei for useful discussions and Mariano Gabitto and Ella Batty for their comments on the manuscript.

7 Author Contributions

Conceived and designed the methods/experiments: GEM, LP, JPC, LEG, EJC. Performed the experiments: LEG. Analyzed the data: GEM, LP. Contributed reagents/materials/analysis tools: AL, PH, EJC. Wrote the paper: GEM, LP, EJC, LEG, JPC.

References

1. Wagenaar DA, Madhavan R, Pine J, Potter SM. Controlling bursting in cortical cultures with closed-loop multi-electrode stimulation. The Journal of neuroscience. 2005;25(3):680–688.

2. Middlebrooks JC, Snyder RL. Selective electrical stimulation of the auditory nerve activates a pathway specialized for high temporal acuity. The Journal of Neuroscience. 2010;30(5):1937–1946.

3. Meacham KW, Guo L, DeWeerth SP, Hochman S. Selective stimulation of the spinal cord surface using a stretchable microelectrode array. 2011;.

4. Bakkum DJ, Frey U, Radivojevic M, Russell TL, Muller J, Fiscella M, et al. Tracking axonal action potential propagation on a high-density microelectrode array across hundreds of sites. Nature Communications. 2013;4(2181).

5. Kim R, Joo S, Jung H, Hong N, Nam Y. Recent trends in microelectrode array technology for in vitro neural interface platform. Biomedical Engineering Letters. 2014;4(2):129–141.

6. Jorgenson LA, Newsome WT, Anderson DJ, Bargmann CI, Brown EN, Deisseroth K, et al. The BRAIN Initiative: developing technology to catalyse neuroscience discovery. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2015;370(1668). doi:10.1098/rstb.2014.0164.

7. Barry MP, Dagnelie G. Use of the Argus II Retinal Prosthesis to Improve Visual Guidance of Fine Hand MovementsArgus II Retinal Prosthesis. Investigative ophthalmology & visual science. 2012;53(9):5095–5101.

8. Goetz GA, Palanker DV. Electronic approaches to restoration of sight. Reports on Progress in Physics. 2016;79(9):096701.

9. Franke F, Jakel D, Dragas J, Muller J, Radivojevic M, Bakkum D, et al. High-density microelectrode array recordings and real-time spike sorting for closed-loop experiments: an emerging technology to study neural plasticity. Frontiers in Neural Circuits. 2012;6(105). doi:10.3389/fncir.2012.00105.

10. Potter SM, El Hady A, Fetz EE. Closed-Loop Neuroscience and Neuroengineering. Frontiers in Neural Circuits. 2014;8(115). doi:10.3389/fncir.2014.00115.

11. Lewicki MS. A review of methods for spike sorting: the detection and classification of neural action potentials. Network: Computation in Neural Systems. 1998;9(4):R53–R78.

12. Pillow JW, Shlens J, Chichilnisky EJ, Simoncelli EP. A Model-Based Spike Sorting Algorithm for Removing Correlation Artifacts in Multi-Neuron Recordings. PLoS ONE. 2013;8(5):e62123. doi:10.1371/journal.pone.0062123.

13. Rey HG, Pedreira C, Quiroga RQ. Past, present and future of spike sorting techniques. Brain research bulletin. 2015;119:106–117.

14. Merletti R, Knaflitz M, De Luca CJ, et al. Electrically evoked myoelectric signals. Crit Rev Biomed Eng. 1992;19(4):293–340.

15. Hottowy P, Dąbrowski W, Kachiguine S, Skoczen A, Fiutowski T, Sher A, et al. An MEA-based system for multichannel, low artifact stimulation and recording of neural activity. Proc 6th Int Meet Substrate-integrated Micro Electrode Arrays. 2008; p. 261–265.

16. Hottowy P, Skoczen A, Gunning DE, Kachiguine S, Mathieson K, Sher A, et al. Properties and application of a multichannel integrated circuit for low-artifact, patterned electrical stimulation of neural tissue. Journal of neural engineering. 2012;9(6):066005.

17. Brown EA, Ross JD, Blum RA, Nam Y, Wheeler BC, Deweerth SP. Stimulus-Artifact Elimination in a Multi-Electrode System. 2008;2(1):10–21.

18. Wichmann T, Devergnas A. A novel device to suppress electrical stimulus artifacts in electrophysiological experiments. Journal of Neuroscience Methods. 2011;201(1):1–8. doi:10.1016/j.jneumeth.2011.06.026.

19. Obien M, Deligkaris K, Bullmann T, Bakkum DJ, Frey U. Revealing neuronal function through microelectrode array recordings. Frontiers in neuroscience. 2015;8:423.

20. Hashimoto T, Elder CM, Vitek JL. A template subtraction method for stimulus artifact removal in high-frequency deep brain stimulation. Journal of Neuroscience Methods. 2002;113:181–186. doi:10.1016/S0165-0270(01)00491-5.

21. Wagenaar D, Potter SM. Real-time multi-channel stimulus artifact suppression by local curve fitting. Journal of Neuroscience Methods. 2002;120:113–120. doi:10.1016/S0165-0270(02)00149-8.

22. Heffer LF, Fallon JB. A novel stimulus artifact removal technique for high-rate electrical stimulation. Journal of Neuroscience Methods. 2008;170:277–284. doi:10.1016/j.jneumeth.2008.01.023.

23. Erez Y, Tischler H, Moran A, Bar-gad I. Generalized framework for stimulus artifact removal. Journal of Neuroscience Methods. 2010;191(1):45–59. doi:10.1016/j.jneumeth.2010.06.005.

24. Müller J, Bakkum DJ, Hierlemann A. Sub-millisecond closed-loop feedback stimulation between arbitrary sets of individual neurons. Closing the Loop Around Neural Systems. 2014; p. 38.

25. Sekirnjak C, Hottowy P, Sher A, Dabrowski W, Litke A, Chichilnisky E. Electrical stimulation of mammalian retinal ganglion cells with multielectrode arrays. Journal of neurophysiology. 2006;95(6):3311–3327.

26. Sekirnjak C, Hottowy P, Sher A, Dabrowski W, Litke AM, Chichilnisky E. High-resolution electrical stimulation of primate retina for epiretinal implant design. The Journal of neuroscience. 2008;28(17):4446–4456.

27. Litke A, Bezayiff N, Chichilnisky EJ, Cunningham W, Dabrowski W, Grillo A, et al. What does the eye tell the brain?: Development of a system for the large-scale recording of retinal output activity. IEEE Transactions on Nuclear Science. 2004;51(4):1434–1440.

28. Jepson LH, Hottowy P, Mathieson K, Gunning DE, Dabrowski W, Litke AM, et al. Spatially Patterned Electrical Stimulation to Enhance Resolution of Retinal Prostheses. J Neurosci. 2014;34(14):487–4881.

29. Ekanadham C, Tranchina D, Simoncelli EP. A blind sparse deconvolution method for neural spike identification. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira FCN, Weinberger KQ, editors. NIPS; 2011. p. 1440–1448. Available from: http://dblp.uni-trier.de/db/conf/nips/nips2011.html#EkanadhamTS11.

30. Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. MIT Press; 2006.

31. Wilson A, Gilboa E, Nehorai A, Cunningham JP. Fast kernel learning for multidimensional pattern extrapolation. In: Advances in Neural Information Processing Systems; 2014. p. 3626–3634.

32. Gilboa E, Saatçi Y, Cunningham JP. Scaling multidimensional inference for structured Gaussian processes. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2015;37(2):424–436.

33. Genton MG. Classes of kernels for Machine Learning: a statistics perspective. Journal of machine learning research. 2001;2(Dec):299–312.

34. Pachitariu M, Steinmetz N, Kadir S, Carandini M, Harris KD. Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels. bioRxiv. 2016; p. 061481.

35. Jepson LH, Hottowy P, Mathieson K, Gunning DE, Dabrowski W, Litke AM, et al. Focal electrical stimulation of major ganglion cell types in the primate retina for the design of visual prostheses. The Journal of Neuroscience. 2013;33(17):7194–7205.

36. Jepson LH, Hottowy P, Weiner GA, Dabrowski W, Litke AM, Chichilnisky EJ. High-Fidelity Reproduction of Spatiotemporal Visual Signals for Retinal Prosthesis. Neuron. 2014;83(1):87 – 92. doi:http://dx.doi.org/10.1016/j.neuron.2014.04.044.

37. Grosberg LE, Hottowy P, Jepson LH, Ito S, Kellison-Linn F, Sher A, et al. Axon activation with focal epiretinal stimulation in primate retina. Investigative Ophthalmology & Visual Science. 2015;56(7):780–780.

38. Fine I, Cepko CL, Landy MS. Vision research special issue: Sight restoration: Prosthetics, optogenetics and gene therapy. Vision Res. 2015;111(Pt B):115–23. doi:10.1016/j.visres.2015.04.012.

39. Maturana MI, Apollo NV, Hadjinicolaou AE, Garrett DJ, Cloherty SL, Kameneva T, et al. A Simple and Accurate Model to Predict Responses to Multi-electrode Stimulation in the Retina. PLoS Comput Biol. 2016;12(4):e1004849.

40. Grumet AE, Wyatt JL, Rizzo JF. Multi-electrode stimulation and recording in the isolated retina. Journal of neuroscience methods. 2000;101(1):31–42.

41. Grosberg LE, Ganesan K, Goetz GA, Madugula S, Bhaskhar N, Fan V, et al. Selective activation of ganglion cells without axon bundles using epiretinal electrical stimulation. bioRxiv. 2016; p. 075283.

42. Branchaud E, Burdick JW, Andersen R, et al. An algorithm for autonomous isolation of neurons in extracellular recordings. In: Biomedical Robotics and Biomechatronics, 2006. BioRob 2006. The First IEEE/RAS-EMBS International Conference on. IEEE; 2006. p. 939–945.

43. Franke F, Natora M, Boucsein C, Munk MH, Obermayer K. An online spike detection and spike classification algorithm capable of instantaneous resolution of overlapping spikes. Journal of computational neuroscience. 2010;29(1-2):127–148.

44. Fee MS, Mitra PP, Kleinfeld D. Automatic sorting of multiple unit neuronal signals in the presence of anisotropic and non-Gaussian variability. Journal of neuroscience methods. 1996;69(2):175–188.

45. Franke F, Quiroga RQ, Hierlemann A, Obermayer K. Bayes optimal template matching for spike sorting–combining fisher discriminant analysis with optimal filtering. Journal of computational neuroscience. 2015;38(3):439–459.

46. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S, et al. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in medicine. 1997;16(9):965–980.

47. Titsias MK. Variational learning of inducing variables in sparse Gaussian processes. In: International Conference on Artificial Intelligence and Statistics; 2009. p. 567–574.

48. Wilson AG, Nickisch H. Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP). CoRR. 2015;abs/1503.01057.

49. Hensman J, Matthews AG, Filippone M, Ghahramani Z. MCMC for Variationally Sparse Gaussian Processes. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015. p. 1648–1656. Available from: http://papers.nips.cc/paper/5875-mcmc-for-variationally-sparse-gaussian-processes.pdf.

50. Pais-Vieira M, Yadav AP, Moreira D, Guggenmos D, Santos A, Lebedev M, et al. A Closed Loop Brain-machine Interface for Epilepsy Control Using Dorsal Column Electrical Stimulation. Scientific Reports. 2016;6:32814.

51. Hottowy P, Beggs JM, Chichilnisky EJ, Dabrowski W, Fiutowski T, Gunning DE, et al. 512-electrode MEA system for spatio-temporal distributed stimulation and recording of neural activity. In: Proceedings of the 7th International Meeting on Substrate-Integrated Microelectrode Arrays, Reutlingen, Germany (Stett, A ed), June; 2010. p. 327–330.

52. Chichilnisky E. A simple white noise analysis of neuronal light responses. Network: Computation in Neural Systems. 2001;12(2):199–213.

53. Field GD, Sher A, Gauthier JL, Greschner M, Shlens J, Litke AM, et al. Spatial properties and functional organization of small bistratified ganglion cells in primate retina. The Journal of Neuroscience. 2007;27(48):13261–13272.

8 Supporting information

8.1 Experimental procedures

All electrophysiology data were recorded from primate retinas isolated and mounted on an array of extracellular electrodes as described in previously published literature [35]. Eyes were obtained from terminally anesthetized macaque monkeys (Macaca species, either sex) used for experiments in other labs, in accordance with IACUC guidelines for the care and use of animals. After enucleation, the eyes were hemisected and the vitreous humor was removed. The hemisected eye cups containing the retinas were stored in oxygenated bicarbonate-buffered Ames solution (Sigma) at room temperature during transport (up to 2 hours) back to the lab. Patches of intact retina 3mm in diameter were isolated and placed retinal ganglion cell-side down on a 512-electrode MEA. Throughout the experiments, retinas were superfused with oxygenated bicarbonate-buffered Ames solution at 35°C.

In all experiments the raw voltage signals from each electrode were amplified, filtered, and multiplexed with custom circuitry [16, 51]. Electrodes had diameters of 10-15 μm and were separated by 60 μm. Data were acquired at 20 kHz on all electrodes and bandpass filtered between 43 and 5000 Hz. Charge-balanced, triphasic current pulses with relative amplitudes of 2:-3:1 and phase widths of 50 s were applied to each electrode, and reported current amplitudes correspond to the charge of the second, cathodal, phase. A platinum ground wire circling the perfusion chamber served as a distant ground in all one-electrode stimulation experiments. In some experiments, a 1 mM tetrodotoxin (TTX) solution in Ames' solution was perfused into the retina to inhibit all action potentials in order to directly measure the stimulus artifact in a retinal preparation.

8.1.1 Obtaining the EIs

Retinal ganglion cells (RGCs) were identified in the absence of electrical stimulation using previously described spike sorting techniques [27] and classified into types based on how they respond to a visual white noise stimulus projected onto the retina [52, 53]. For each RGC, thousands of voltage waveforms were averaged on all electrodes, resulting in a spatiotemporal voltage signature specific to that RGC. These signatures are used as templates in our sorting algorithm.

8.2 Estimation of mean

Regarding the mean parameter of the artifact kernels, μ, we follow the standard in the applied statistics community: μ is a centering parameter and all the non-random aspects of data should be captured by it. In our case this component is given by what we call the switching artifact, a waveform $A_0 = A_0(e, t)$ that is present regardless of the amplitude of stimulation. We estimate $\hat{\mu}$ by taking the mean of recordings at the lowest amplitude of stimulation (see Fig 14 for details on the characteristics of the switching artifact, and to see the effect of this mean-subtraction stage on recordings).

8.3 Dataset details

8.3.1 Real data

In table 1 we specify details of the nine retinal preparations for which human annotation was available. In each preparation there were characteristic numbers of stimulating electrodes and neurons being analyzed. Usually, given a stimulating

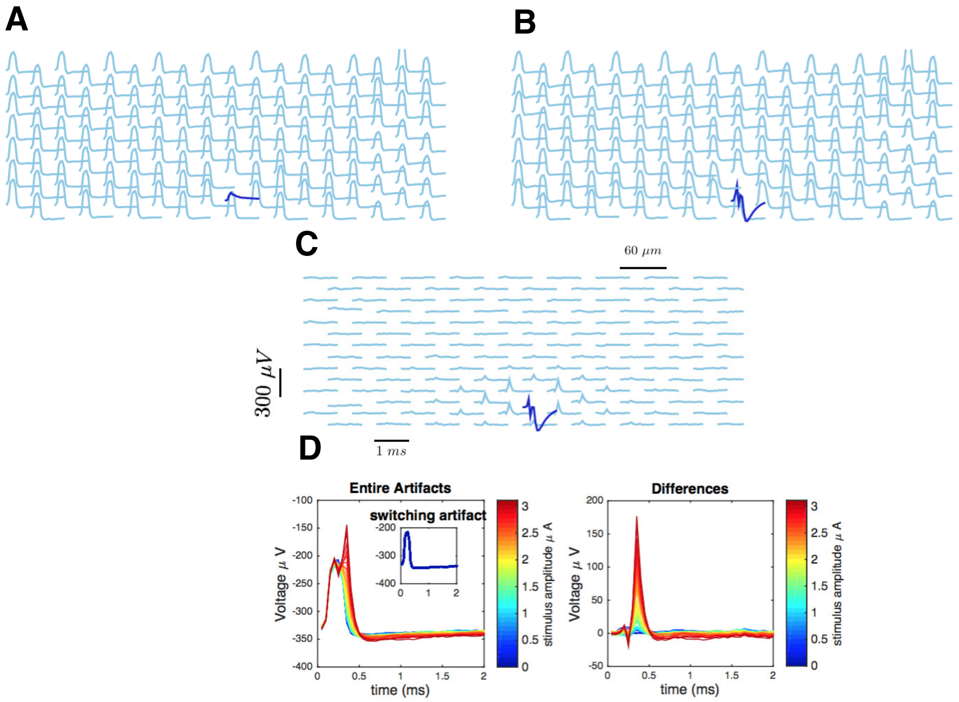

Fig 14. A) Raw artifact traces at the smallest amplitude of stimulation (0.1 μA), considered an estimate of μ, the switching artifact. **B)** Raw artifact traces at 0.99 μA of stimulus. **C)** Difference. Notice that the main text refers to this already mean-subtracted artifact. **D**) *Left*: Raw artifact at all different stimuli for a non-stimulating electrode (inset, switching artifact). *Right*: Differences.

electrode human annotation was available for only one, or at most a few neurons (e.g. two or three). However, to assess algorithm stability in the analysis of a large number of neurons, we considered the totality of EIs of neurons analyzed on each preparation (e.g. 24 in the first preparation), but restricted performance computations to the subsets of neurons for which human annotation was available.

Importantly, we restricted our analysis to the stimulation amplitudes that did not lead to gross contamination of recordings due to the activation entire axonal bundles in the retina (for a recent account of this pervasive phenomenon see [41]), as this would lead to a situation that is not accounted for by our model. For each amplitude series with available human annotation, we determined the maximum amplitude of stimulation that did not lead to activation of a bundle by looking for 'hot' electrodes, distant from the stimulating one, exhibiting high temporal variance in the artifact (here, for simplicity the artifact was estimated by the simple average over traces). Then, we did not consider any amplitude of stimulation beyond the onset of axonal bundle activation, the first amplitude where we identified such hot electrodes. We found that a robust method for estimating this threshold (equivalently, the presence of hot electrodes) was based on a Kolmogorov-Smirnov goodness-of-fit test on the empirical distribution of the (log) temporal variances of the artifact on distant electrodes, with the gaussianity null hypothesis. The appearance of hot electrodes created a new mode in the distribution, leading to a violation of the normality assumption. We found that by setting the cut-off p-value for this test as 10^{-12} we achieved the best match with axonal bundle activation onsets estimated by human experts (not shown).

	EIs	Recordings		
Preparation	Neurons	Trials	Stimulating electrodes	Average trials per stimulus
1	24	385, 661	269	51
2	5	38, 783	17	48
3	18	35, 658	15	21
4	18	29, 190	12	21
5	8	13, 596	26	22
6	12	60, 654	55	34
7	7	67, 799	67	25
8	1	1, 600	1	25
9	31	86, 130	80	30

Table 1. Details on the nine retinal preparations analyzed

8.3.2 Simulated data

Simulated data was created by artificially adding neural activity to TTX recordings, in an attempt to faithful mimic the phenomena observed in the real case [26, 35]. Specifically, we considered 83 neurons (the largest subset of the ones targeted in the real data analysis so that their EIs did not heavily overlap) and recordings to 380 stimulating electrodes (one at a time) in a TTX experiment with $n_j = 6$ trials to $J = 35$ different stimuli between 0.1 and $3.5\mu A$. Then, given a single stimulating electrode we sampled activation curves for all the neurons whose EI at the stimulating electrode was strong enough, indicating proximity. Activation curves were parametrized by their thresholds, chosen uniformly in the stimulation range, and their steepness, also sampled uniformly. Spikes of those neurons were then sampled from these activation curves with latencies chosen so they would match the human spike sorting results in the following two aspects: 1) they had same median latency as a function of the distance between the neuron and stimulating electrodes (spiking of nearby neurons has shorter latency) and 2) they had same variance in spike latency as a function of spike probability (in the steady spiking regimes, where the probability of firing is high, latencies are much less variable). Also, to obtain better estimates of false positive rates, we fed the algorithm with 'dummy' neurons (three per amplitude series, with EIs chosen at random from the available set of remaining neurons) with no spiking at all.

All the reported results involving simulations are based on 5000 samples of amplitude series following the above procedure.