Bayesian optimization of cortical neuroprosthetic vision using perceptual feedback

Journal of Neural Engineering

PAPER

Bayesian optimization of cortical neuroprosthetic vision using perceptual feedback

Burcu Küçükoğlu[1,*], **Leili Soo**[2], **David Leeftink**[1], **Fabrizio Grani**[2], **Cristina Soto Sanchez**[2], **Umut Güçlü**[1], **Marcel van Gerven**[1,3] and **Eduardo Fernandez**[2,3]

[1] Department of Machine Learning and Neural Computing, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands
[2] Institute of Bioengineering, University Miguel Hernández, Elche, Spain
[3] These authors contributed equally to this work.
[*] Author to whom any correspondence should be addressed.

E-mail: burcu.kucukoglu@donders.ru.nl

Keywords: Bayesian optimization, visual neuroprostheses, patient feedback, electrical stimulation, stimulation protocols, stimulation safety, neural prostheses

OPEN ACCESS

RECEIVED
26 January 2025

REVISED
1 June 2025

ACCEPTED FOR PUBLICATION
2 July 2025

PUBLISHED
24 July 2025

Abstract

Objective. The challenge in cortical neuroprosthetic vision is determining the optimal, safe stimulation patterns to evoke the desired light perceptions ('phosphenes') in blind individuals. Clinical studies gain insights into the perceptual characteristics of phosphenes through patient descriptions on provided stimulation protocols. However, the huge parameter space for multi-electrode stimulation makes it difficult to identify the optimality of the stimulation that lead to well-perceived phosphenes. A systematic search in the parameter space of the electrical stimulation is needed to achieve good perception. Bayesian optimization (BO) is a framework for finding optimal parameters efficiently. Using patient's perceptual feedback, a model of patient response based on iteratively generated stimulation protocols can be built to maximize perception quality. *Approach.* A patient implanted with an intracortical 96-channel microelectrode array in their visual cortex was iteratively presented with stimulation protocols, generated via BO vs. random generation (RG) in two separate experiments. Whereas standard BO methods do not scale well to problems with over a dozen inputs, we optimize a set of 40 electrode currents using trust region-based BO. The protocols determine the electrodes to stimulate and with how much current (0–50 μA), on a total current limit of 500 μA. The patient rated each stimulation's perceptual quality on a Likert scale, where 7 indicated the highest quality and 0 no perception. *Main results.* The patient ratings gradually converged on higher values with BO, compared to the RG experiment. BO gradually generated protocols with higher total current, in line with the patient preference for higher currents due to brighter phosphenes. Electrodes previously observed as effective in producing phosphene perception were chosen more by BO also with higher current allocation. *Significance.* This study demonstrates the power of BO in converging to optimal stimulation protocols based on patient feedback, providing an efficient search for stimulation parameters for clinical studies.

1. Introduction

Cortical neuroprostheses offer a promising approach to artificial restoration of vision in blind people. By direct electrical stimulation of the visual cortex through an electrode implant, light percepts called phosphenes can be evoked in the patient's visual field (Brindley and Lewin 1968, Dobelle and Mladejovsky 1974). The key challenge in evoking the desired perception lies in providing the optimal stimulation patterns in a safe manner. Finding optimal stimulation patterns that lead to meaningful and informative perception without over-stimulation of the cortex is of interest for both clinical and simulated phosphene vision studies.

Studies of simulated phosphene vision (Chen *et al* 2009a, 2009b) have emerged as a means to visualize the supposed perceptual effect of a proposed stimulation, ideally based on biological modeling of its consequences on the brain (Granley *et al* 2022b, Fine and Boynton 2024, van der Grinten *et al* 2024). By allowing for the evaluation of a stimulation pattern's visual outcome (de Ruyter van Steveninck *et al* 2022b, 2024) simulated phosphene vision also enables optimization of stimulation parameters virtually in a controlled test bed without the limitations from ethical and safety considerations of testing on patients. Therefore, relatively high resolution patterns through stimulation of multiple electrodes can be freely tested, based on optimizations of even complex end-to-end machine learning algorithms (Küçükoğlu *et al* 2022, 2025, Granley *et al* 2022a, de Ruyter van Steveninck *et al* 2022a). Such algorithms tend to have a high number of parameters to be optimized that require large amounts of data for training and possibly a long pre-training time. However, there is no way of knowing the exact perceived effect of stimulation on the patients unless testing with them directly.

Clinical studies, on the other hand, are focusing on understanding the perceptual effects of stimulation protocols administered based on descriptions from the patients regarding the shape, color, brightness, and localization of the phosphenes evoked in single electrode stimulation settings (Brindley and Lewin 1968, Dobelle *et al* 1974, Schmidt *et al* 1996, Fernandez *et al* 2021). These single electrode stimulations also help identify the threshold current needed for an electrode to induce perception (Chen *et al* 2020). Typically, an arbitrary multiplier to this threshold value is applied when administering current to ensure phosphene perception, albeit with no guarantee of good-quality perception, due to a mere focus on the existence of perceptual detection rather than quality. Whereas detection is based only on the visibility of phosphenes, the perceptual quality, thus how well the phosphenes are seen, may vary depending on various factors. For example, as stimulation currents are raised above the threshold, the size of phosphenes is found to reach a rapid saturation after an initial increase (Bosking *et al* 2017). While higher currents tend to increase the perceived brightness ratings for phosphenes, highest brightness ratings are not reached by the largest currents tested (Fernandez *et al* 2021). In fact, brightness is found to display a non-linear increase based on the stimulation amplitude, with small increases in amplitude at lower levels leading to noticeable changes in brightness, but additional increases in current at higher stimulation amplitudes resulting in smaller incremental changes in brightness (Evans *et al* 1979). When it comes to multi-electrode stimulation, with an increasing number of simultaneously stimulated electrodes, both the relative brightness and size of phosphenes are observed to increase (Fernandez *et al* 2021). Due to these complex relationships between stimulation parameters and phosphene characteristics, explaining the effect of stimulation on perceptual quality is not straightforward. As a more feasible point of investigation, multi-electrode stimulation studies tend to focus on pattern detection via phosphenes. In multi-electrode settings (Oswalt *et al* 2021, Bosking *et al* 2022), electrodes are arranged in various configurations, such as in a line, circle, or a letter, to assess the recognition of the intended pattern (Fernandez *et al* 2021), or its orientation (Chen *et al* 2020). Once a satisfactory pattern description is found, the corresponding stimulation protocol is recorded for future use. Previous research has also shown that recognizable shapes can be elicited through seemingly random electrode patterns (Fernandez *et al* 2021). In order to find meaningful phosphene shapes stimulated by a random selection of electrodes, a random number of electrodes with random stimulation parameters must be stimulated within the safety constraints. However, this method is typically used sparingly, as drawing conclusions about the optimality of stimulation protocols for phosphene perception is challenging due to the vast parameter space. A more systematic approach is necessary to streamline what would otherwise be a tedious parameter search process based solely on verbal descriptions.

The differences in approach, focus, and capability between the simulated and clinical phosphene vision studies reveal a gap between their current testing conditions, which can be reduced by meeting midway across their key differences. These differences indicate opposite levels of dependence on modeling capabilities, complexity of the stimulation protocols tested, the evaluation risks and costs, as well as opposites in testing limitations, training requirements, the number of parameters controlled, the relevance of available feedback, and finally the opposite focuses on systematic optimization vs. trial-based observation via the prioritization of phosphene interpretability vs. detection. This gap between the testing conditions of clinical and simulated phosphene vision studies demonstrates the need for a systematic approach for searching the space of parameters for multi-electrode stimulation in patients in an informed manner. This would enable clinical studies to take a step further towards testing more complex stimulation protocols that currently tend to emerge from simulated phosphene vision studies' safe yet virtual optimization conditions. As a result, the strengths of the testing conditions from both the simulated and

clinical phosphene vision studies can be combined for an effective search on neuroprosthetic vision parameters.

We envision an optimization setup based on direct patient feedback on stimulation protocols generated in an intelligent and autonomous manner, to search systematically through the space of parameters in a fast and efficient way given safety constraints. For this purpose, we propose Bayesian optimization (BO) as a fast and theoretically optimal method of parameter optimization in clinical settings (Gelman *et al* 2020, Garnett 2023). As an initial step towards this vision, here we present a procedure to optimize stimulation protocols involving a selection of electrodes to stimulate and their current levels, based on direct patient feedback on perceptual quality. To our knowledge, this is the first automated optimization procedure in the field of neuroprosthetic vision that is based on direct patient feedback. Such a systematic search procedure for optimal stimulation protocols could make early experimentation with a patient implanted more efficient and additionally informative on perception quality, especially for multi-electrode stimulation settings.

BO (Shahriari *et al* 2016, Garnett 2023) is an approach for solving black-box functions that converges quickly to globally optimal parameters with as few function evaluations as possible by searching within constraints given. This makes BO ideal for safety critical settings like neurostimulation optimization where evaluating the objective function is also expensive. It works by iteratively presenting data points and observing its quantitative evaluation in response, in order to fit a surrogate model capturing the relationship between the inputs and outputs of the black-box function. The model is updated at each iteration after the response to the newly generated data point, based on all past evaluations. The updated surrogate model in turn helps generate the next data point to be presented, using an acquisition function that reflects the surrogate's optimal or uncertain areas to ensure a balance between exploration and exploitation in the search space. The generated data points gradually converge to data points that maximize the evaluation as the model converges to optimal parameters.

BO has been previously used for various safety-critical contexts from robotics to neurostimulation (Sui *et al* 2015, Berkenkamp *et al* 2021, Cooper and Netoff 2022). In the context of neurostimulation, the use of BO involves optimization of parameters *in vivo* for deep brain stimulation in rats (Cole *et al* 2024) and in Parkinson's disease patients (Sarikhani *et al* 2022) within safety constraints, for tACS to evoke phosphenes in seeing human subjects steered by their preference ratings (Lorenz *et al* 2019), and for stimulation on motor cortex (Choinière *et al* 2024) in rats and monkeys (Bonizzato *et al* 2023). To our knowledge, no *in vivo* BO study for optimization of neurostimulation on humans or other animals has been previously conducted for visual prostheses. Therefore, the mentions of existing studies highlight the major differences from our setting of interest or their relevant details. BO has also been utilized for *in silico* parameter optimization for deep brain stimulation (Grado *et al* 2018, Cooper and Netoff 2022, Nagrale *et al* 2023), for offline recalibration of neuromodulation parameters over time in somatosensory neuroprosthesis (Aiello *et al* 2023), for offline optimization of peripheral motor neuroprosthesis stimulation from rat and monkey data (Losanno *et al* 2021), and for modeling spinal cord stimulation preferences from patient data to improve therapeutic outcomes (Zhao *et al* 2021). Preferential BO approaches have been used for *in silico* optimization of patient-specific parameters in retinal implants, based on the preferences of simulated patients (Granley *et al* 2023) or sighted subjects (Fauvel and Chalk 2022) over pairwise comparisons of predicted percepts. Linear encoders (Fauvel and Chalk 2022) or deep stimulus encoders (Granley *et al* 2023) were used to generate stimulation patterns (Fauvel and Chalk 2022) or electrical stimuli (Granley *et al* 2023) respectively. Candidate perceptions were modeled via a fixed decoder that receives the proposed patient-specific parameters as input, whereas previously learned deep stimulus encoders receive both the patient-specific parameters and the desired perception (Granley *et al* 2023). However, all these studies focused on a maximum of 5 parameters (Bonizzato *et al* 2023) for *in vivo* and 13 parameters (Granley *et al* 2023) for *in silico* optimization settings, without exploring the use of BO for high-dimensional parameter spaces, which are naturally relevant for especially the neuroprosthetic vision optimization problem setting.

In the context of neurostimulation optimization of our study, the objective function can be considered as a function of phosphene perception quality as defined by patient feedback, with its input being the stimulation protocol presented. Therefore, the goal is to find stimulation protocols that maximize perceptual quality (for details, see section 2.4.5). As a proof of concept, we conducted a clinical test on a blind patient implanted with an intracortical 96-channel microelectrode array in their visual cortex. Given the iterative nature of BO, stimulation protocols were presented to the patient iteratively, upon which the patient provided quantitative feedback for the quality of the phosphene perception raised based on a Likert scale, where a score of 7 indicated the highest quality and 0 no perception. To evaluate BOs performance against a benchmark, another experiment where stimulation protocols were randomly generated was conducted.

In order to apply BO in the context of stimulation protocols, we parameterized the problem space by electrode location indices where a current value was assigned for each index that was later mapped to a location in the electrode array based on a predetermined list of electrodes. Hence, at each iteration, a current value between [0–50] μA was allocated for each electrode, without surpassing the maximum total current constraint of 500 μA. Whereas standard BO methods do not scale well to problems with over a dozen parameters, here we utilized high-dimensional BO with trust regions to optimize for 40 parameters, thus allocating current only to a predetermined set of 40 electrodes in the array. Due to the high-dimensional nature of the cortical neuroprosthetic vision optimization problem, scaling BO to a higher dimensional parameter space was a crucial aim of the study, for which the trust regions provided an efficient solution to reap the benefits of BOs real-time optimization.

Overall, our study provides two unique contributions by proposing an optimization procedure of neural stimulation via cortical neuroprostheses for the blind based on direct patient feedback informing on perceptual quality, and via the use of BO that can scale well to a higher dimension of parameters than the standard BO, through the use of trust regions.

2. Methods

2.1. Bayesian optimization for personalized phosphene vision

Here, we propose an experimental design to find optimal patient-specific stimulation parameters for cortical neuroprosthetic vision using high-dimensional BO with trust regions. The design of the experimental process for Bayesian optimization of phosphene vision (BOPhos) (figure 1) starts with a vector format of the stimulation parameters to be generated by the BO, with each element corresponding to a stimulation current between [0–50] μA for an electrode, subject to a maximum total current constraint of 500 μA for the total of 40 electrodes. This makes BO operate in the high-dimensional space of 40 parameters, corresponding to stimulation currents for each electrode. The stimulation parameters are later mapped to the electrode array based on the predetermined location indices for the electrodes. The generated stimulation protocol is administered through the implant to the patient's visual cortex, upon which the patient is asked to provide a score between [0–7] for its perception, where 0 indicates no perception, and 7 corresponds to the highest perceptual quality on a Likert scale. This patient feedback is used via BO to update a surrogate model of the patient response, which in turn helps generate a new stimulation protocol at each iteration via an acquisition function to search the parameter space intelligently by balancing exploration and exploitation. Over time, as the surrogate model learns to capture the patient response effectively, it converges on generating more optimal stimulation protocols that maximize patient response. To analyze its performance, the automated and intelligent search process via BO is compared against a random search process as a benchmark.

2.2. Bayesian optimization

BO is a framework for solving black-box optimization problems using as few function evaluations as possible (Shahriari *et al* 2016). When evaluating the objective function is expensive—as is the case with cortical implant stimulation patterns—it is crucial to carefully choose where to query next. Efficiently optimizing the objective requires effectively balancing between exploration and exploitation.

Like other forms of optimization, the goal is to find the optimum of a function $f(\mathbf{x})$ on a bounded set $\Omega \subset \mathbb{R}^d$. What sets the Bayesian approach apart is the use of a global surrogate model of the objective function, that captures both the predicted objective value and their associated uncertainties. Deciding where to query the objective function next involves a trade-off between exploiting the best function value observed so far and exploring new inputs that might yield even better results. This balance is captured by the acquisition function $\alpha(\mathbf{x};\mathcal{D}_n)$, which determines the next optimal point by explicitly weighing exploration against exploitation. Instead of using a local gradient or Hessian approximation, the acquisition function leverages *all* the information available from previous iterations when selecting a new input. Together, the probabilistic surrogate model and the acquisition function form the two key components of the BO framework, enabling the optimization of black-box objective functions with a minimal number of evaluations.

2.2.1. Gaussian process (GP) surrogate model

The GP is a prior probability distribution over functions commonly used as a surrogate model in BO due to its flexibility, data efficiency and ability to quantify uncertainty (Williams and Rasmussen 2006, Brochu *et al* 2010b, Snoek *et al* 2012, Garnett 2023). GP regression can predict a function value $f(\mathbf{x})$ and its uncertainty at an unknown location using a set of past observations $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

Assuming noisy measurements, the mean and variance of the prediction at a new input $\mathbf{x}$ are given by:

$$\mu_n(\mathbf{x}) = \mu_0(\mathbf{x}) + \mathbf{k}(\mathbf{x})^\top \left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1} (\mathbf{y} - \mathbf{m}) \quad (1)$$

$$\sigma_n^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \mathbf{k}(\mathbf{x})^\top \left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1} \mathbf{k}(\mathbf{x}) \quad (2)$$

where $\mu_0(\mathbf{x})$ is the prior mean function, k is the covariance function and $\mathbf{k}(\mathbf{x}) = [k(\mathbf{x},\mathbf{x}_1), \ldots, k(\mathbf{x},\mathbf{x}_n)]$ is

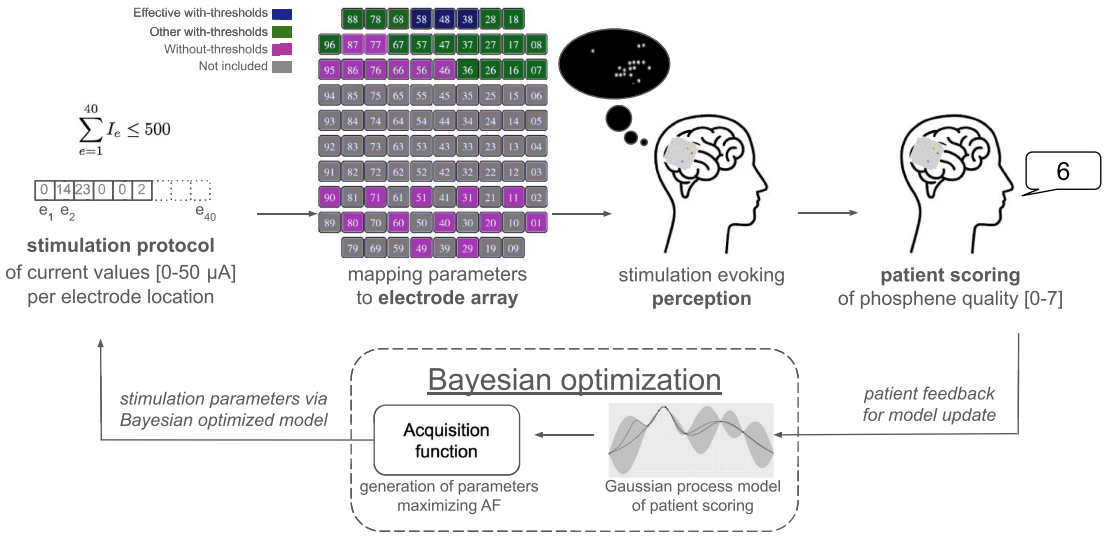

Figure 1. Experimental setup of BOPhos involving stimulation protocol generation and stimulation, followed by patient scoring of phosphenes' perceptual quality (for details, see section 2.4.5), which can be used as feedback for updating the Bayesian optimization procedure as opposed to the benchmark experiment of random generation of stimulation parameters. The different colors on the electrode array reflect the electrodes' categorization based on the characteristics of perceptions produced upon their stimulation, whereas the gray electrodes were excluded from the stimulation generation for simplification (for details, see section 2.4.3).

the covariance vector between the new point and the observed data points. The elements of the prior mean vector $\mathbf{m}$ and $n \times n$ covariance matrix $\mathbf{K}$ are defined as $m_i := \mu_0(\mathbf{x}_i)$ and $K_{i,j} := k(\mathbf{x}_i, \mathbf{x}_j)$ respectively.

The posterior mean $\mu_n(\mathbf{x})$ and variance $\sigma_n^2(\mathbf{x})$ evaluated at a point $\mathbf{x}$ represent the model's prediction and its uncertainty after n observations respectively. The latter is particularly valuable in BO, as the global optimization approach leverages the uncertainty estimates to balance exploration and exploitation effectively.

The covariance function k determines the structural properties of the function modeled by the GP, such as smoothness and periodicity. It acts as a prior over function space, determining how the function behaves between observed data points (Williams and Rasmussen 2006). Common choices for covariance functions include the squared exponential (SE) kernel and the Matèrn 5/2 kernel:

$$k_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2} r^2\right) \tag{3}$$

$$k_{\mathrm{Matèrn5/2}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \left(1 + \sqrt{5} r + \frac{5}{3} r^2\right) \exp\left(-\sqrt{5} r\right) \tag{4}$$

where $r^2 = \sum_{j=1}^{d} (x_j - x_j')^2 / \ell_j^2$ and ℓ_j represent the kernel length scales that act as smoothness parameters, and d is the dimensionality of the input.

2.2.2. Acquisition function and next query selection

The acquisition function in BO determines where to query the objective function next. Various acquisition functions have been proposed in the literature, each effective under different conditions (Brochu *et al* 2010b). Here, we focus on Thompson sampling (Thompson 1933), a Bayesian approach rooted in posterior sampling, which selects input points by drawing samples from the posterior distribution of the objective function modeled by a GP:

$$\mathbf{x}_{n+1} = \underset{\mathbf{x} \in \mathcal{F}}{\arg\max} f^{(n)}(\mathbf{x}) \text{ where } f^{(n)} \sim \mathcal{GP}(\mu_0, k \mid \mathcal{D}_n) \tag{5}$$

where $f^{(n)}$ represents the sampled function from the GP posterior after n iterations.

To select the next input point $\mathbf{x}_{n+1}$, a set of r candidate points is sampled. For each candidate point $\mathbf{x}_i$, a function value $f^{(n)}(\mathbf{x}_i)$ is drawn from the model's posterior, with $1 \leqslant i \leqslant r$. The candidate with the highest sampled function value is chosen as the next query point. This approach implicitly balances exploration and exploitation by naturally favoring input points that are both promising (high posterior mean) and uncertain (high posterior variance). This increases the chances of exploring areas with potential for improvement while reducing focus on less promising regions.

2.2.3. Generation of candidate points within input constraints

To ensure patient safety, the total current from all stimulated electrodes must remain below a chosen threshold. This allowable maximum total current is

often considerably lower than the possible maximum, leading to a trade-off between applying a high current to a single electrode or distributing lower currents across multiple electrodes. We express this as a linear inequality constraint:

$$c(\mathbf{x}) := \sum_{i=1}^{d} x_i \leqslant b \tag{6}$$

where d is the number of electrodes and b is the maximum total current threshold. The set of all electrode configurations that satisfy this constraint is given by $\mathcal{F} = \{x \in \Omega \mid c(\mathbf{x})\}$. In our experiment, $d = 40$ and $b = 500\,\mu$A.

To confine the Thompson sampling method to this feasible set, we apply rejection sampling when generating candidate solutions. Whereas Thompson sampling directly samples from the posterior distribution, rejection sampling repeatedly samples candidates until the acceptance criteria are met. By sampling r candidate points $\mathbf{x}_i$ from the full compact set $\Omega \subset \mathbb{R}^d$ and discarding those that fall outside $\mathcal{F}$, we obtain a set of candidate solutions within the feasible region.

2.2.4. Trust region BO for high dimensional search spaces

Traditionally, BO has been applied to problems of relatively low dimensionality, typically involving up to a dozen input parameters. While BO has demonstrated highly competitive performance on such low-dimensional search spaces, its application to black-box functions with high-dimensional search spaces remains challenging (Binois and Wycoff 2022). This difficulty arises due to the exponential growth of the search space with increasing dimensionality and because high-dimensional functions often have varying characteristics across different regions of the input space, making it more difficult to fit an accurate global objective function.

To address these challenges, trust region BO (TurBO) combines principles of local optimization with the global exploration capabilities of BO (Eriksson *et al* 2019). This method restricts the search space to a confined region called the trust region, which dynamically moves around the objective landscape. The trust region is centered around the best evaluation found thus far and adapts dynamically by expanding or shrinking depending on whether successive observations are improving or declining, respectively (see section 2.4.6 for further details). Typically, the trust region is chosen as a hypercube, although recent work has explored alternatives, such as using a truncated normal distribution to generate candidate solutions (Rashidi *et al* 2024).

In this work, we leverage the trust region approach to scale the optimization to a problem with $d = 40$ input parameters representing the electrode currents, which is significantly more than the number of input dimensions that can effectively be handled by traditional BO. The local search employed within an adaptive trust region enables efficient exploration of the high-dimensional search space.

2.3. BOPhos algorithm

The BOPhos algorithm can be summarized as follows:

1. Evaluate a set of n_{init} initial stimulation parameters using space-filling design such as a Sobol sequence (Sobol' 1967, Owen 1998), and initialize the trust region at the configuration with the highest patient rating.
2. Repeat until the number of patient evaluations is exhausted:
   (a) Fit a GP model to the observed patient ratings.
   (b) Generate a set of r candidate points $x_1, \ldots, x_r \in \Omega$ within the trust region, and discard the samples that fall outside the feasible set (see equation (6)).
   (c) For each remaining candidate, sample a realization $f(\mathbf{x}_i)$ to determine the electrode configuration that produces the highest sampled function value.
   (d) Present the stimulation protocol and receive the patient's perception quality rating.
   (e) Move the center of the trust region to the highest evaluation and adjust its size based on the success and failure counters.

2.4. Experimental setup

2.4.1. Subject

The patient tested was a 27 year-old male, who was participating in a 6 month clinical study aimed at developing a cortical visual neuroprosthesis for the blind. The patient was blinded as a result of a traumatic head injury and underwent bilateral enucleation of the eyes one year prior to the experiment. The study received approval from the Clinical Research Committee of the General University Hospital of Elche and was registered on ClinicalTrials.gov (NCT02983370). It adhered strictly to all applicable ethical standards, including clinical trial regulations (EU No. 536/2014, which replaced Directive 2001/20/EC), the Declaration of Helsinki, and the European Commission Directives (2005/28/EC and 2003/94/EC). The patient gave his informed written consent prior to participating. The data was systematically and securely stored.

2.4.2. Electrode array and implant location

An intracortical 96-channel microelectrode Utah array, supplied by BlackRock Microsystems (Salt Lake City, UT, USA), was implanted via surgery into the patient's early visual cortex, between V1 and V2. The procedure followed a standard surgical approach,

enhanced by robot assistance (Fernandez *et al* 2021, Rocca *et al* 2023). Figure 1 includes a picture of the electrode array. A predetermined set of 40 electrodes were included in the study for stimulation (represented in non-gray colors) in order to restrict the optimization to 40 parameters. For the BO implementation, the electrode locations were provided in a fixed order for the mapping of location indices to electrodes as follows: [18, 28, 38, 48, 58, 68, 78, 88, 8, 17, 27, 37, 47, 57, 67, 77, 87, 96, 7, 16, 26, 36, 46, 56, 66, 76, 86, 95, 1, 20, 40, 60, 80, 11, 31, 51, 71, 90, 29, 49].

2.4.3. Selection of electrodes and their characteristics

The selected electrodes composed two equally sized main electrode categories: those with measurable phosphene thresholds when stimulated in isolation and those without individual phosphene thresholds. The categorization was made before our experimentation, based on other experiments during the 6 month clinical study with the patient, in order to identify phosphene thresholds. Electrode thresholds were measured using a binary search procedure, starting with stimulation at 1 μA. If no phosphene was perceived, the current was doubled. If a phosphene was detected, a half-interval search was performed between the perceived and previous unperceived intensities until two consecutive correct identifications were made. This process continued until a threshold was determined or the maximum intensity of 128 μA was reached. While threshold experiments determine the stimulation parameters which induce a perceivable phosphene, the perceptual quality of the perceived phosphene is not known based on these experiments. All 96 electrodes were tested, and 10% of trials were control trials with no stimulation. The selected electrodes and the categorization reflect the general characteristics of all electrodes in the array. However, evaluating thresholds in multi-electrode stimulation settings is complex due to the numerous possible combinations and parameters, and as a result, the interactions between stimulation on multiple electrodes are not fully understood.

The first group of electrodes ($N = 20$) have a known individual phosphene threshold, meaning they are known to produce phosphenes when stimulated individually above the threshold. The individual phosphene perception thresholds for these electrodes, based on averages over multiple recordings with the patient prior to the current study, can be seen in figure 2. This group can be further divided into two subgroups. The first subgroup comprises the most effective electrodes ($N = 3$) due to their significantly lower thresholds and the distinguishable shapes of the phosphenes produced compared to the remaining electrodes, based on the observation of the experimenters in previous clinical studies. The second subgroup includes the rest of the electrodes in the group

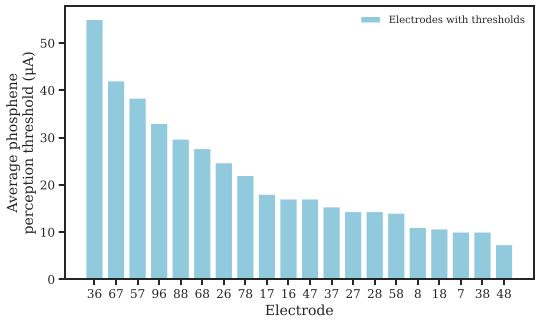

Figure 2. Average phosphene thresholds for electrodes belonging to the category of electrodes with individual phosphene thresholds. The averages are based on the mean of three sessions with the patient during the clinical experiments prior to the Bayesian optimization experimentation where all electrodes were tested.

($N = 17$), as shown in figure 1 by colors navy and green respectively.

The second group of electrodes ($N = 20$), colored in pink in figure 1, do not have individual phosphene thresholds, hence they do not produce phosphenes when stimulated individually. However, they can play a role in producing perception when stimulated along with other electrodes, as they were previously observed to produce a conscious perception when stimulated concurrently.

The selected electrodes were primarily located in the top rows of the electrode array, as this area contained the most effective electrodes for inducing phosphenes. To enhance spatial coverage, every second electrode from the bottom row was included, while the middle rows were excluded due to their failure to elicit phosphenes or record neural spikes.

2.4.4. Electrical stimulation

We stimulated the visual cortex using Ripple Neuromed's Explorer Summit processor, which had 3 Micro2+Stim front ends with 32 channels each. The stimulation step size was set to 1 μA. The electrical stimulus lasted for a total duration of 167 ms with 50 pulses at a frequency of 300 Hz. The cathodic-first biphasic pulses had a pulse width of 170 μs with 60 μs interphase duration. The current intensity for a stimulated electrode is generated via the experimental protocol of BOPhos. In order to reduce the total charge injected, we interleaved each stimulation pulse from different electrodes with a time delay of 0.5 ms between the onset of each consecutive pulse inside a train. With a frequency of 300 Hz and a pulse duration of 0.4 ms, up to 6 pulses can be sent separately. For stimulation patterns consisting of more than 6 electrodes, some of the pulses were sent at the same time. The consecutiveness was based on the order of the mapped electrode indices, where the

sequence of delays are repeated after every 6 electrodes. Thus, the pulses were sent simultaneously for the 1st and 7th electrodes, followed after a 0.5 ms delay by the 2nd and 8th electrodes to be stimulated, ordered according to the indices list.

Each BOPhos iteration corresponded to a single stimulation trial, in which the generated stimulation protocol was presented to the patient, followed by the patient's perceptual rating. The next stimulation followed after the experimenter had entered the score on the computer keyboard, thus without a fixed inter-stimulation interval. A fixed delay of 1 s was applied after the keyboard response, before the subsequent stimulation was delivered.

2.4.5. Perceptual quality ratings

In response to electrical stimulation, the patient saw circular phosphenes, according to patient descriptions from previous experiments with the same patient during the 6 month clinical study. Although the induced phosphenes appeared as small dots, their brightness and clarity varied. After the delivery of each stimulation trial, the patient was asked to rate the perceptual quality of the phosphenes perceived with the following instructions on the definitions of the scores: 0 = no perception, 1 = very very bad, 2 = very bad, 3 = bad, 4 = normal/neutral, 5 = good, 6 = very good, 7 = very very good perception. Thus, the quality ratings could be interpreted as how well the patient sees the phosphenes. It is important to note that in some instances, phosphenes could be perceived as 'too bright', which the patient considered poor quality. Therefore, while brightness was likely the primary factor in the patient rating, the quality was not fully defined as having the highest possible brightness.

2.4.6. Hyperparameters and algorithmic details of BO

For the GP model, we use a Matèrn 5/2 kernel with learnable length scale parameters for each dimension, through a procedure known as automatic relevance determination (ARD) (Neal 2012). The hyperparameters are determined by optimizing the log marginal likelihood, which is given by:

$$\log p\left(\mathbf{y} \mid \mathbf{X}, \theta\right) \propto \underbrace{-\frac{1}{2}\left(\mathbf{y}-\mathbf{m}^{\theta}\right)^{\top}\left(\mathbf{K}^{\theta}+\sigma^{2}\mathbf{I}\right)^{-1}\left(\boldsymbol{y}-\mathbf{m}^{\theta}\right)}_{\text{Model fit}} \underbrace{-\frac{1}{2}\log\left|\mathbf{K}^{\theta}+\sigma^{2}\mathbf{I}\right|}_{\text{Complexity penalty}} \tag{7}$$

where the trainable parameters θ consist of the kernel hyperparameters ℓ_i, the kernel variance σ_f^2 and the (constant) mean function μ_n, and σ_n^2 is the noise level that is added to the diagonal of the covariance matrix.

The initial set of $n_{\text{init}} = 12$ data points are selected using a Sobol sequence on the full input domain Ω. After observing the patient ratings, the trust region size was initialized at 80% of the input range. A single trust region was utilized due to the limited sample budget, as employing multiple trust regions simultaneously would have required a higher number of model evaluations. At each iteration, a set of $r = 5000$ candidate points (Eriksson *et al* 2019) were generated within a hyperrectangular trust region before going for the constraint feasibility check. The number of candidate points were chosen to balance the effectiveness of the sampling procedure with the computational demands. The trust region size was updated across the optimization process for generation of more suitable candidates from a constrained input range, by expanding after successive improvements over the best evaluation found so far or shrinking after successive failures to improve. To adjust the trust region size, a success tolerance of 3 was used, meaning expansion updates were done after 3 consecutive successes, where expansion implied doubling the size as long as it is kept below 1.6 (Eriksson *et al* 2019). The failure tolerance was set automatically after the initialization as 40, following after the dimensionality of the input. The shrinking implied halving the trust region size (Eriksson *et al* 2019). The experiment was continued until the trust region size reached below 0.045. Rather than the use of a predetermined number of iterations, the use of a trust region size-based stopping criterion ensured that the optimization is finalized at a point of convergence, meaning when the results could not be improved anymore within the given trust region size, after multiple iterations resulting in consecutive failures. Given the nature of the trust region-based optimization, this point of convergence can be considered as a 'good-enough' local optimum, whereas claims on global optimality can not be made due to use of finite number of iterations.

2.4.7. Software

All code was implemented in Python, and is available at https://github.com/burcukoglu/BOPhos.git. A Python version of 3.8.12 was used. For BO implementation BoTorch library version 0.11.1, and for GP implementation GPyTorch 1.11 was used, along with PyTorch 2.3.1. The software was run on a machine operating Ubuntu (20.04, Linux kernel version 5.15), with an Intel Core i9-7900X CPU (3.30GHz), 16GB RAM, and an NVIDIA GeForce RTX 3090 GPU.

2.4.8. Baseline experiment

To compare the performance of BO in converging to the generation of optimal stimulation parameters, a second experiment was conducted as a benchmark with random generation (RG) of stimulation parameters. This choice of benchmark followed after the use of random electrode selection and current allocation in clinical studies (Fernandez *et al* 2021). The number of electrodes to be stimulated was randomly generated from a uniform distribution, followed by

the stimulation currents per electrode. Electrode locations were randomly selected with a uniform probability, after which the generated currents were assigned. The maximum total current limit was ensured, as well as the selection of at least one electrode for stimulation, where a minimum total current of 1 μA was enforced for the stimulation protocol to be presented. The current values were regenerated until the total current constraints are satisfied, before the stimulation protocol can be presented to the patient. Similar to the BO experiment, an iteration corresponded to a single stimulation trial. The experiment was conducted for the same number of iterations as the BO experiment to ensure consistent conditions.

3. Results

3.1. Convergence properties

BO converged to optimal parameters for generating stimulation protocols that lead to highly rated perceptions in only 230 iterations within 20 minutes of patient testing, demonstrating its efficiency in the search process. Convergence of BO was evaluated based on a predetermined stopping criterion related to the performance of the optimization process as detailed in section 2.4.6. The BO experiment was terminated once the condition for the stopping criterion emerged during the experimentation. Although running for many more iterations may have improved the results as the number of iterations approaches infinity, this is infeasible for patient testing, making it conceptually difficult to comment on the global optimality of the results. Therefore, when referred to the optimality of parameters or stimulation protocols found by BO, these could be considered as 'good-enough' solutions, hinting at least at local optimality. For comparison, the RG experiment was also run for 230 iterations.

3.2. Patient scoring in response to stimulation protocols

The stimulation protocols generated by BO are compared to those randomly generated across the iterations for each experiment, where each iteration corresponds to a single stimulation trial. Patient scoring across trials are displayed in figure 3, which shows higher scores are reached by the BO experiment also with higher frequencies compared to the RG experiment. The direct display of exact patient scoring demonstrates the unsystematic search of the RG approach over the parameter space, which produces lower and inconsistent scores with an initial dip of scores at the beginning. Stimulation protocols that could lead to scores beyond the mean of the possible score range are barely generated, suggesting that only suboptimal stimulation protocols had the opportunity to be tested for patient perception. In contrast, BO yields a steady increase to reach higher scores soon after the start of the experiment, and maintains the high scores with consistency across many iterations, yet with regular drops to its value. The patient scoring demonstrates that BO is effective in generating stimulation protocols whose perception quality is evaluated at the highest levels by the patient. However, due to its nature as an iterative optimization procedure, it is prone to effects of adaptation due to repeated stimulation, as suggested by the regular patient scoring drops across later iterations despite the overall consistency in consecutive scores.

3.3. General characteristics of the stimulation protocols

Total currents across trials for the two experiments are compared in figure 4. In the RG experiment, the stimulation protocols generated produce inconsistent total current values, yet reaches the maximum total current constraint of 500 μA a few times. BO, however, converges to the maximum total current soon after the iterations start. While high levels of current are not necessarily desirable for repeated stimulation due to the risk of overstimulation, it was observed in clinical studies that patients have a preference for brighter stimuli, which is attributed to higher stimulation currents. This explains why the BO converged on stimulation protocols with the highest total current possible, within the safety limits provided for it. It is worth noting that in the RG experiment, despite receiving the highest total current a few times throughout the iterations, the patient has not evaluated these stimulation protocols with higher scores (e.g. iterations 142 and 160 both received a score of 2 as seen in figure 3). This emphasizes the importance of the choice of electrodes in reaching optimal stimulation protocols, which the BO has clearly succeeded in, given the difference in patient scoring to the stimulation protocols with the highest total currents compared to the RG experiment. As electrodes could have different phosphene perception thresholds, beyond which a stimulation could produce a perception only, it could be the case that despite high total currents administered, some of it has not contributed directly to patient perception, hence causing a lower patient scoring. However, BO seems to overcome this by generating stimulation protocols of not only high total current but also with a choice of electrodes that have low phosphene thresholds, hence more effective in producing perception. This is later further analyzed based on the knowledge of individual perception thresholds for electrodes acquired in previous clinical studies.

As depicted in figure 4, the number of electrodes stimulated is also inconsistent for the RG experiment, whereas for BO it converges to around 30 out of 40 possible electrode choices, soon after initial iterations. The mean current per stimulated electrodes in a generated protocol is similarly inconsistent for the RG experiment, yet converges to around 16 μA for BO. While the number of electrodes stimulated is in

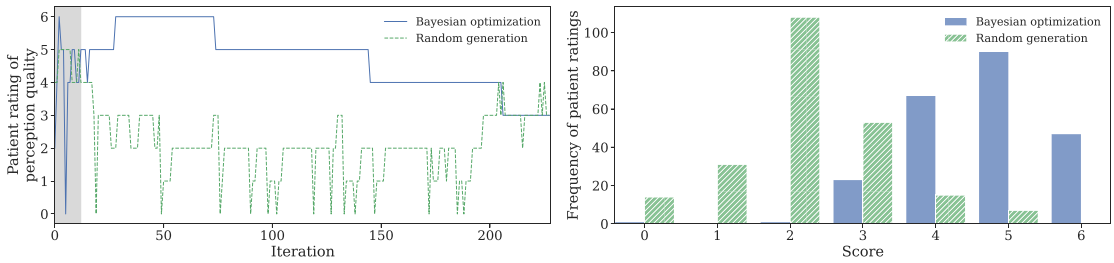

Figure 3. Patient scoring across iterations of stimulation protocol generation shows the exact patient rating in response to each stimulation trial (left), where ratings range between 0–7, with the higher values implying higher perceptual quality and 0 signifying no perception. Each iteration corresponds to a single stimulation trial, where a new stimulation protocol is generated, through the selection of a potentially different subset of electrodes for stimulation and their current allocation, based on the Bayesian optimized model or through random generation depending on the experiment. The gray area indicates the initial set of data points tested for the Bayesian optimization experiment, before the optimization begins. Total frequency of scores displays the distribution of patient ratings for each experiment (right).

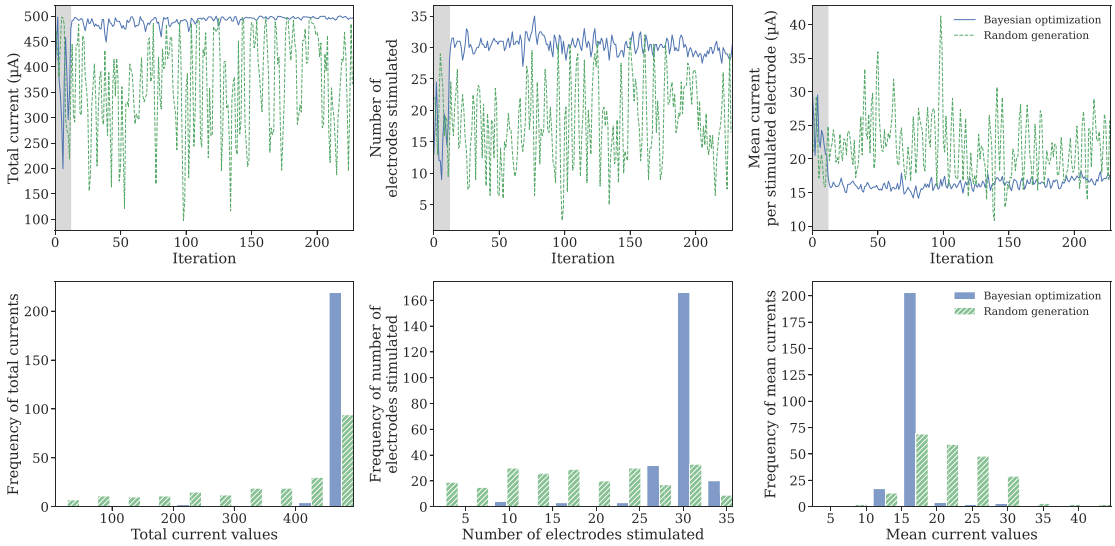

Figure 4. Top row displays the total current across iterations (left), number of electrodes stimulated per generated stimulation protocol (middle), and mean current per stimulated electrode (right). Plotted values show the moving average with a window size of two. The gray area indicates the initial set of data points tested for the BO experiment, before the optimization begins. Each iteration corresponds to a single stimulation trial. Bottom row displays the respective total frequencies of the values shown in the top row, binned into ten equal intervals.

general higher for BO, the mean current per stimulated electrodes is also generally lower, despite higher total current values on average. This may suggest an advantage of the BO framework for minimizing the mean current per electrode stimulated, despite not being explicitly steered to optimize for it.

3.4. Analysis of electrode categories

The choice of electrodes by the generated stimulation protocols and their currents are further investigated based on the categorization of electrodes. As figure 5 shows, the choice of the most effective subcategory of electrodes and also their current allocation is much higher in the BO compared to the RG experiment, demonstrating the success of BO in learning to generate the most optimal electrodes for phosphene perception. Similarly, the selection and the current allocation for the main category of electrodes with individual phosphene perception thresholds were significantly higher in the BO than in the RG experiment. Within this main category, BO especially learns to allocate most of the current it distributes to the subcategory of most effective electrodes, to the extent that the second subcategory is allocated similar current levels in BO to that in RG, yet still with an increased count of stimulated electrodes. This demonstrates BOs effectiveness not only in the selection of more effective electrodes but also in their allocation of higher current values. On the other hand, the electrodes without individual phosphene thresholds seem to be selected more in BO compared to RG, and also with higher current allocation than any other group, despite being selected less by BO compared to the other main category that contains equal number of electrodes.

Non-cumulative scoring of electrode category stimulation counts and the corresponding currents

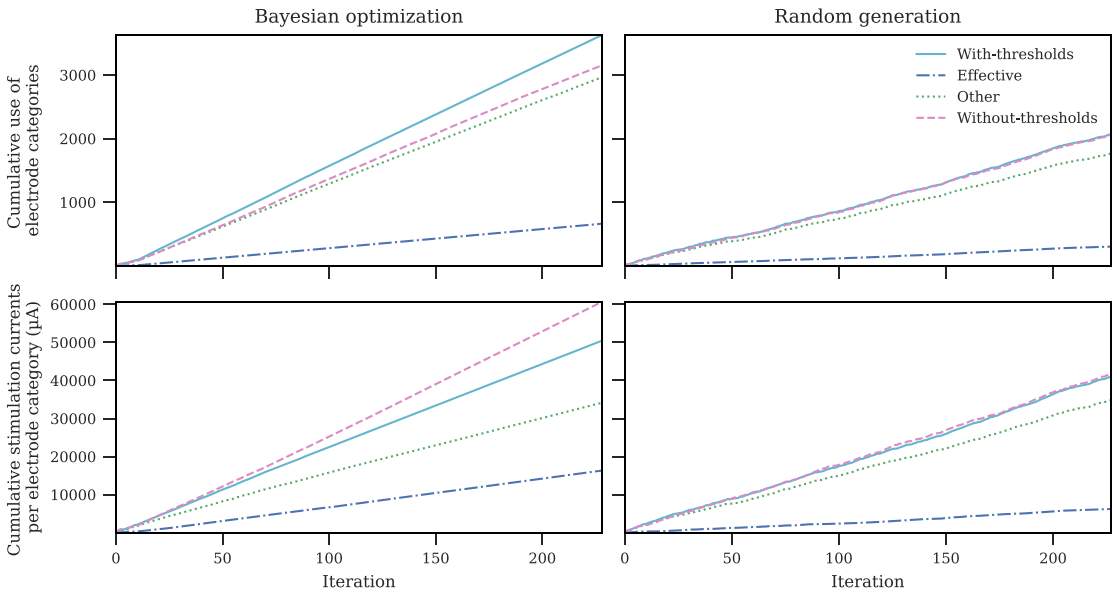

Figure 5. Cumulative count of electrodes stimulated per electrode category (top) and their cumulative current (bottom). The two main categories are electrodes with individual phosphene thresholds (*with-thresholds*), and without individual phosphene thresholds (*without-thresholds*). The first category (*with-thresholds*) is further divided into two subcategories, with the first (*effective*) including the most effective electrodes due to lower individual phosphene thresholds and distinct phosphene perception characteristics, and the second (*other*) including the rest of the electrodes in the category. Note that the number of electrodes across categories differ.

are depicted in figure 6. The RG experiment reflects the randomness in the generation of stimulation protocols, where electrode selection and current allocation per category is inconsistent across iterations and merely reflect the number of electrodes available in each category in their ranking of selection count and total current. In contrast, the three most effective electrodes were always chosen by BO from right after the start, with the highest mean current per electrode overall across categories. This shows BOs tendency to select and allocate higher current to the electrodes from the most effective electrodes subcategory. The count of electrodes stimulated for the rest of the categories rather reflects the size of the categories, with on average 16 out of 20, 13 out of 17, and 14 out of 20 electrodes being chosen for remaining groups, respectively for the main group with threshold, the subgroup of those with threshold, and finally those without threshold. Still, the group of electrodes without individual thresholds are highlighted, with the group's total current surpassing those of the group of electrodes with individual thresholds, despite the lower selection ratio for the electrodes in the former. As a result, the group of electrodes without individual thresholds reaches to the second highest levels of mean current allocation per electrode, following after the group of most effective electrodes.

3.5. Analysis of individual electrodes

An individual analysis for each electrode within each category is provided in figure 7. While each electrode is selected about equally often with similar allocated current in the RG experiment, a clear preference for or aversion to certain electrodes is apparent in the BO experiment, reflecting their unique contributions to perception or phosphene thresholds. Among the most effective electrodes, all are chosen roughly equally, with electrode 38 being allocated the most current by BO and electrode 48 the least. As for their stance in the bigger group of electrodes, there are 12 other electrodes in the with-thresholds category, and 12 in the without-thresholds category, that are selected to be stimulated as often. Among the electrodes with individual thresholds, the three most effective electrodes are all in the top six (along with 8, 17 and 78) of the most current-allocated ones, with 38 having the highest current allocation of the main category. Together, these six electrodes constitute the electrodes with lowest individual phosphene thresholds (see figure 2), thus suggesting their effectiveness in inducing phosphenes also in multi-electrode stimulation setting. Surprisingly, two other electrodes (60, 51) have higher current allocation than 38, which are part of the category of electrodes without individual thresholds, implying potentially a high contribution in perception on their part in the multi-electrode settings.

There is a clear aversion of BO to certain electrodes, which have been allocated very little current across iterations and only infrequently selected. These electrodes seem to have minimal contribution to perception and thus can be considered rather ineffective. Among the electrodes without individual thresholds, there are more electrodes BO is averse to (46, 20, 77, 66, 95, 56, 1, 76) based on both current allocation and

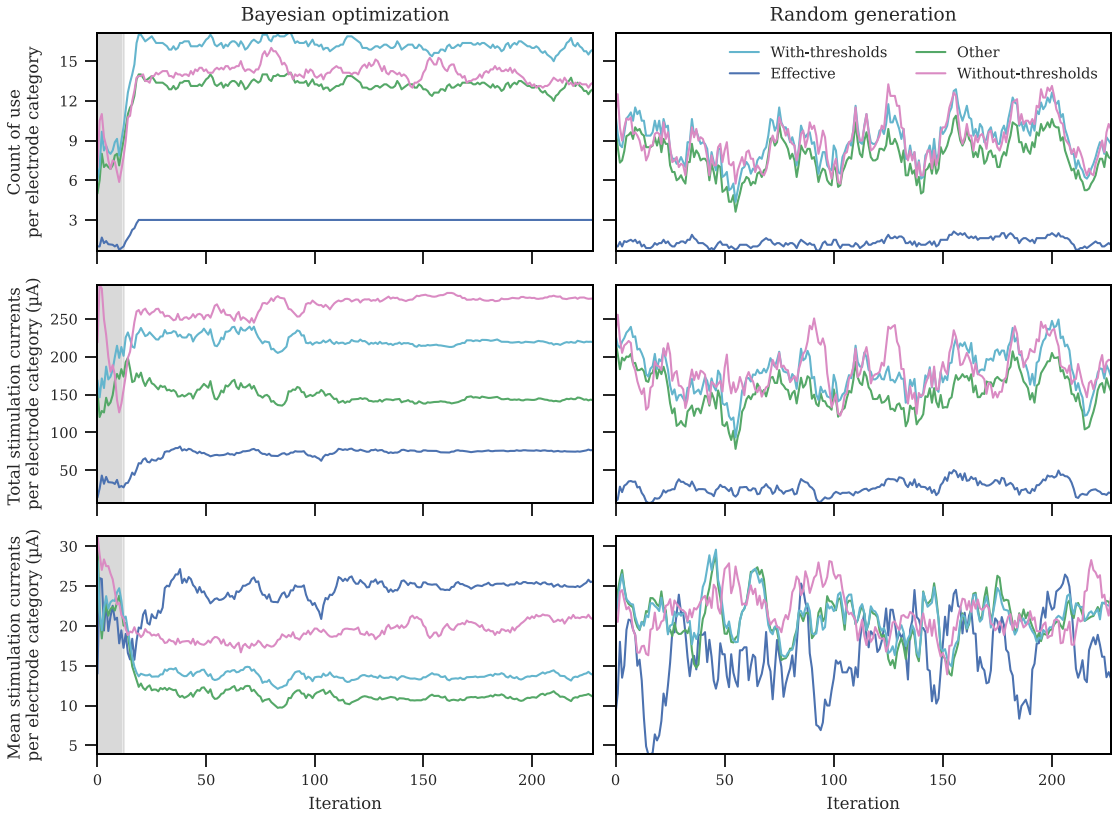

Figure 6. Number of electrodes stimulated per electrode category (top), their total currents (middle) and the mean current per electrode category (bottom). Number of electrodes available in the two main categories, and later for the two subcategories of the first group are as follows: *with-thresholds* = 20, *without-thresholds* = 20, *effective* = 3, *other* = 17. Plotted values show the moving average with a window size of 8. The gray area indicates the initial set of data points tested for the BO experiment, before the optimization begins.

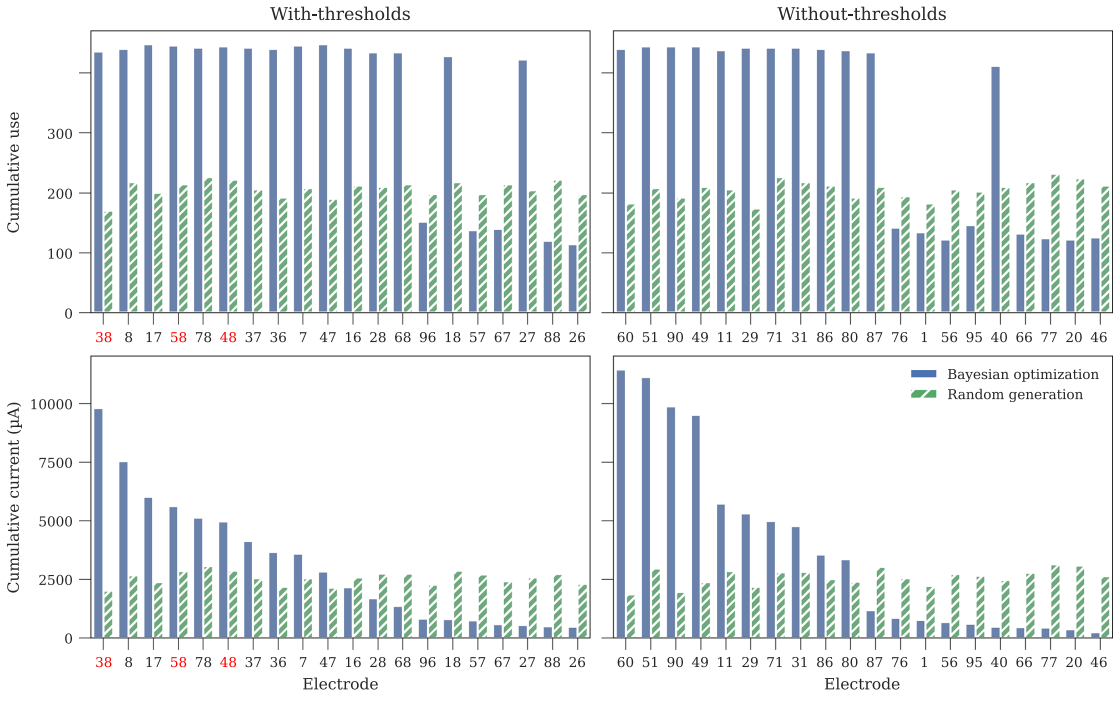

Figure 7. Cumulative count of stimulation per electrode (top) and their final cumulative current (bottom) grouped based on electrode categories. Electrode labels are ranked based on cumulative current values for BO. Red labels highlight the electrodes belonging to the most effective subcategory.

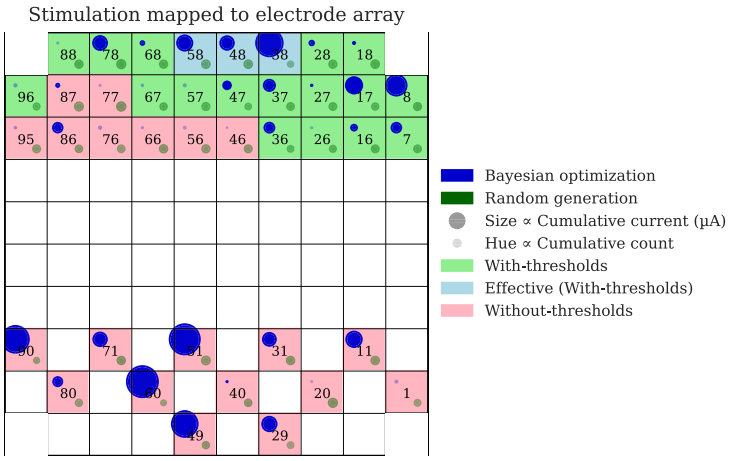

Figure 8. Final cumulative stimulation values mapped to the electrode array where numbers represent specific electrodes and the background colors represent different electrode categories. For each electrode, the upper blue circle represents the results of the Bayesian optimization experiment, whereas the lower green circle represents the random generation experiment. Circle sizes and hues are proportional to the total stimulation current and total count of electrode use for stimulation across iterations respectively.

electrode selection, compared to the with-thresholds category. The electrodes with individual thresholds show the most dispersed current allocation levels, with half of them being allocated rather low current levels, and five having both the lowest current and selection levels (26, 88, 67, 57, 96). These five electrodes are also among the ones with higher individual phosphene thresholds (see figure 2), which might explain BOs aversion to them, provided that their ineffectiveness due to requiring high levels of current to induce perception continues in multi-electrode settings. Minimal selection of such electrodes requiring high levels of current to induce perception might also explain the lower mean current per stimulated electrode observed for the BO compared to the RG experiment.

Several electrodes stand out for having both the lowest current and the highest selection levels. These are electrodes 40 for the without-thresholds, 27 and 18 for the with-thresholds categories. This may imply a rather consistent contribution to perception by these electrodes in the multi-electrode settings, even with low current stimulation. The lower individual phosphene thresholds for electrodes 27 and 18 (see figure 2), and their subsequent effectiveness to induce phosphenes with lower currents, may explain their high selection levels and low current allocation, provided that these characteristics persist in multi-electrode stimulation settings. Overall, BO seems to provide rather a fast way of gathering insights on specific electrodes' contribution to perception (see figure 8 for a mapping of results to the electrode array).

The observed preference of BO to allocate higher total and mean currents for the group of electrodes without individual thresholds compared to the electrodes with individual thresholds (see figures 5 and 6), can be potentially explained by the higher current allocation levels for the top ranking electrodes without thresholds, as opposed to higher number of effective electrodes in the with individual thresholds category (see figure 7). While this preference for electrodes without individual phosphene thresholds in current allocation may reflect an important contribution to perception from them in multi-electrode stimulation settings, it cannot be concluded confidently due to a lack of an approach of clear analysis. As an alternative explanation, it could be the case that for multi-electrode stimulation settings, these electrodes have high phosphene perception thresholds, hence making the stimulation current really high, yet with minimal contribution to the brightness or perception. While the patient scoring seems to increase initially, it is observed to decrease later (see figure 3). This might be a result of the increase in the current allocation to the top ranking electrodes without individual phosphene thresholds around the iteration when the scores start to decrease (figure 9), in the case their stimulation is ineffective for perception. However, the decline in patient scores during later iterations of BO could also be attributed to adaptation effects or psychological factors, such as reduced motivation or attention, particularly given the regular and gradual nature of the decline that rather maintains a consistency in the scores in between. This may also be supported by the lack of a major decrease in the currents allocated to individual electrodes across later iterations (see figure 9). In fact, the currents per electrode seem to rather converge on certain values with BO, potentially hinting at their phosphene thresholds under multi-electrode stimulation. The observed tendency for electrodes with individual phosphene thresholds to exhibit higher or lower final current levels than

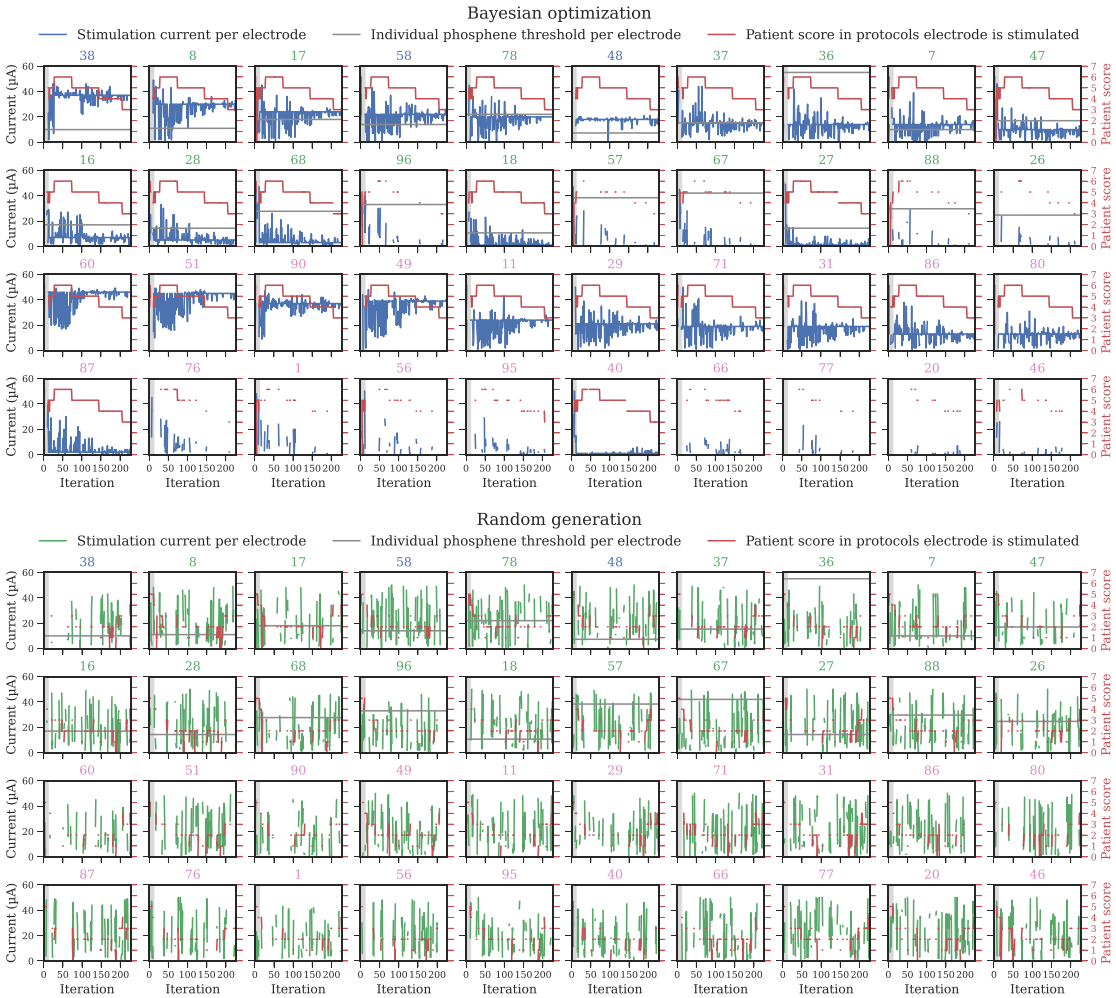

Figure 9. Stimulation currents (scaled by left y-axis) per electrode across iterations and the patient scores (scaled by right y-axis in red) for the respective protocols the electrode is stimulated as part of, within the Bayesian optimization (upper subplots) vs. random generation (lower subplots) experiments. For ease of inspection, when the electrode is not stimulated in an iteration, the zero current and the corresponding patient score for that protocol is not displayed. Individual phosphene thresholds for electrodes in the with-thresholds category are provided with gray horizontal lines. For each subplot, the electrode number is indicated in the title, with the title colors highlighting the category the electrode belongs to, as in figures 1 and 8. Within the subplots for each experiment, the upper two rows correspond to electrodes with individual thresholds, and the lower two rows correspond to electrodes without individual thresholds, where the electrodes are ordered based on their final cumulative current rank as seen in figure 7.

their individual thresholds, may provide insight into the current levels required for effective phosphene perception when stimulating an electrode along with other electrodes. Ideally, future analysis should support the current observations with an analysis of the impact of adaptation and psychological effects, and also the testing of BO-generated stimulation protocols in the later iterations before any exposure to prior stimulation.

In conclusion, aligned with the findings of previous clinical studies, BO learnt to converge on choosing and allocating the most current to the most effective electrodes. Furthermore, it enabled a way to test for optimal stimulation protocols in the multi-electrode setting in a fast and efficient manner, which would have required an exhaustive search if it was to be done purely via clinical studies. This could make the onboarding of a patient newly implanted with a cortical neural prosthesis faster, allowing for efficient and systematic initial experimentation to reach conclusions about the effectiveness of electrodes and their phosphene perception thresholds in a patient-specific manner. Moreover, going beyond merely quantifying perception thresholds and assuming sufficient perception with currents above, BO helps inform about the quality of above-threshold perception, which can be good or poor, irrespective of the level of current administered even in the case of a low threshold, thus bringing additional value to the current capabilities of clinical studies.

3.6. Model evaluation

In order to provide statistical support for our findings, we next provide further analysis on patient ratings and stimulation data.

3.6.1. Statistical analysis of patient ratings

We first investigated whether patient ratings obtained from the RG and BO experiments differ systematically at the population level. Ignoring the sequential nature of the protocol and modeling the patient ratings as draws from a latent probabilistic model, we seek to determine if the BO method provides improved patient ratings compared to the RG baseline. We adopt a Bayesian approach to infer the marginal likelihood (or evidence) of the observed data under different latent probabilistic models that could have generated the patient ratings. The null hypothesis posits that the scores from both trials share the same underlying population distribution, while the alternative hypothesis states that a population-level difference exists between the RG and BO patient ratings.

To compare these hypotheses, we conduct a Bayesian model comparison by constructing separate probabilistic graphical models for the null hypothesis ($\mathcal{M}_0$) and the alternative hypothesis ($\mathcal{M}_1$). Bayesian model comparison aims to quantify the relative evidence in the data for these two models using the Bayes factor, defined as:

$$\mathrm{BF}_{10} = \frac{p(\mathcal{D} \mid \mathcal{M}_1)}{p(\mathcal{D} \mid \mathcal{M}_0)} \tag{8}$$

where $\mathcal{D}$ represents the observed patient ratings from both trials.

To model patient ratings, we acknowledge that while the perceptual quality can be expected to exist on a continuous scale, it is observed experimentally through discrete, ordinal categories $\{0, 1, \ldots, 7\}$. We therefore construct a hierarchical ordinal regression model, where perceptual quality follows a latent Gaussian distribution defined over the interval $[0, 7]$ and the observations are generated by applying fixed thresholds to this distribution. Specifically, the likelihood of each ordinal rating is computed from the differences in cumulative distribution function (CDF) of the latent Gaussian between adjacent thresholds (Gelman *et al* 2020). This modeling strategy reflects both the discrete and ordinal characteristics of the patient ratings. This results in the following generative model:

$$\begin{aligned}
\mu &\sim \mathcal{N}\left(m_0, s_0^2\right) \\
\ln\sigma &\sim \mathcal{N}(0,1) \\
z_i &\sim \mathcal{N}(\mu,\sigma)\big|_{[0,7]} && i = 1,\ldots,N \\
\tau &= (\tau_0 < \tau_1 < \cdots < \tau_8) \\
p_{i,k} &= \Phi\left(\tfrac{\tau_{k+1}-\mu_i}{\sigma}\right) - \Phi\left(\tfrac{\tau_k-\mu_i}{\sigma}\right) && i = 1,\ldots,N,\ \ k = 1,\ldots,7 \\
y_i &\sim \text{Categorical}\left(p_{i,0},\ldots,p_{i,7}\right) && i = 1,\ldots,N
\end{aligned}$$

where $m_0 = 3$, $s_0^2 = 2$, and the fixed thresholds are given by $\tau = (0, 0.5, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7)$, with z symbolizing the latent continuous rating, Φ CDF, and y observations.

In the null hypothesis model ($\mathcal{M}_0$), the latent Gaussian distribution has a single mean μ and standard deviation σ across both RG and BO datasets. In the alternative hypothesis model ($\mathcal{M}_1$), each trial (RG and BO) has its own latent Gaussian parameters $\mu_{\mathrm{RG}}, \sigma_{\mathrm{RG}}$ and $\mu_{\mathrm{BO}}, \sigma_{\mathrm{BO}}$, while all other model components (thresholds and ordinal likelihood) remain identical. To quantify the difference between the two protocols, we define the effect size $\delta = \mu_{\mathrm{RG}} - \mu_{\mathrm{BO}}$. This difference captures the magnitude by which the latent patient ratings differ on average between the RG baseline and the BO method. Under the null model ($\mathcal{M}_0$), which assumes a common mean across trials, this effect size implicitly equals zero ($\delta = 0$). Lastly, since in the BO experiment the initial $n_{\mathrm{init}} = 12$ observations were generated using a Sobol sequence, we exclude these measures from the analysis to ensure a fair comparison of protocol-driven outcome. To estimate the marginal likelihoods (evidences) required for Bayesian model comparison, we employ sequential Monte Carlo (SMC) inference (Hinne 2025b). SMC efficiently approximates the marginal likelihood by sequentially sampling and re-weighting particles, making them well-suited for hierarchical Bayesian models. For numerical inference, we utilize the BaMoJax library (Hinne 2025a), an efficient and scaling Bayesian modeling framework designed for hierarchical probabilistic models. We run the SMC inference with 1000 particles and 2000 SMC iterations, chosen to ensure stable convergence and reliable estimates of the marginal likelihood.

Figure 10 shows the posterior latent Gaussian distributions inferred under the RG model (green), BO model (blue), and the null hypothesis model (gray), using the posterior mean estimates of parameters $\mu_{\mathrm{RG}}, \sigma_{\mathrm{RG}}$ and $\mu_{\mathrm{BO}}, \sigma_{\mathrm{BO}}$. Observed patient ratings are overlaid as bar plots for comparison, while the posterior distribution of the effect size (δ) is displayed

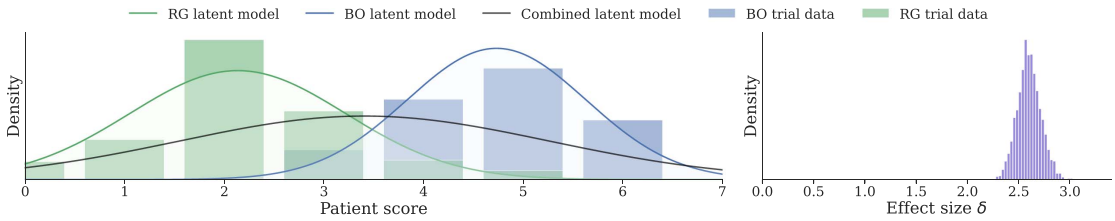

Figure 10. Latent continuous model fit of $\mathcal{M}_0$, in gray, and $\mathcal{M}_1$, in green and blue, with the ordinal data from the RG and BO trials respectively (left), and the distribution of effect size ($\delta = \mu_{BO} - \mu_{RG}$) conditioned on $\mathcal{M}_1$ in purple (right). A strong evidence in favor of the alternative hypothesis is found (logBF = 199.7).

separately in the right panel. Our Bayesian model comparison strongly favors the alternative hypothesis ($\mathcal{M}_1$) over the null ($\mathcal{M}_0$), yielding a log Bayes factor of approximately 199.7. Such a high Bayes factor constitutes decisive evidence that the latent distributions of patient ratings differ significantly between the BO and RG trials. Furthermore, the posterior distribution has a mean $\mathbb{E}[\delta] = 2.567$ and a standard deviation $\sqrt{\mathbb{V}[\delta]} = 0.155$, indicating that patient ratings under the BO protocol are consistently higher compared to the RG baseline.

3.6.2. Cross-validation analysis

In BO, GP models are widely adopted as surrogate probabilistic models due to their non-parametric nature, flexibility, and effectiveness at capturing non-linear relationships (Shahriari *et al* 2016), motivating its choice of prior distribution for modeling neural stimulation and patient ratings. However, the validity of the GP prior in this application warrants further investigation through a comparative cross-validation analysis against simpler Bayesian alternatives. Although patient experimentation time was limited, this analysis leverages data previously collected from both BO and RG experiments to evaluate model performance. Specifically, we compared the GP surrogate model against Bayesian generalized linear models (GLM) using Ridge regularization, which assigns a zero mean Gaussian prior on all weights with a common precision parameter, and ARD, which assigns a zero mean Gaussian prior on each weight but with its own precision parameter. Both Bayesian GLM variants are established benchmarks within Bayesian literature (MacKay 1992), and offer additional interpretability due to their linear design.

We apply k-fold cross-validation adapted to time series data (Arlot and Celisse 2010), due to the sequential nature by which the data is collected. In time series k-fold cross-validation, data is sequentially partitioned into k equal-length intervals, where, starting from the first interval, the first k intervals are cumulatively used as the training set to predict the subsequent interval. This preserves the temporal structure inherent to the collected data, and aims to infer how well a model predicts the patient ratings given the stimulation protocol input, which we measure via the mean squared error (MSE) and mean negative log likelihood (MNLL).

Table 1 summarizes the average performance metrics across all folds. For the BO data, the GP achieves a lower MSE out of the considered models, but a higher MNLL in comparison to both Bayesian GLM variants. The ARD and Ridge models perform comparable in terms of MSE, while the Ridge model achieves lowest MNLL on the BO dataset. Although the GP model demonstrates superior predictive accuracy (lower MSE), it has poorer probabilistic calibration (higher MNLL) compared to linear variants, likely stemming from increased complexity and non-linearity. For the RG dataset, we find that the GP model and the Ridge model perform comparably in terms of MSE, but the Ridge model achieves favorable probabilistic calibration. Finally, the ARD model performs worse on the RG data in terms of MSE and MNLL. The performance difference between the Bayesian GLM variants may result from the Ridge model requiring estimation of fewer precision parameters, making it more robust in the sparse-data regime typical of early experimental stages. The results with BO data suggest that while the GP model is a valid predictor of patient rating data in response to neural stimulation, the Ridge Bayesian GLM model achieves competitive performance on both datasets, and can thus be a relevant consideration as a probabilistic surrogate model in future works. Still, GP provides the benefit of advanced capabilities to model the expected non-linear relationship between neural stimulation and patient ratings through the lack of a linearity assumption that may prevent an oversimplification in the modeling.

4. Discussion

This study proposed BOPhos, an experimental procedure to optimize stimulation protocols for cortical neuroprosthetic vision based on direct patient feedback on perceptual quality using BO, in order to efficiently and systematically search through the huge parameter space of multi-electrode stimulation in safety-critical clinical settings. By converging within as little as 20 min, BO has learned to generate optimal

Table 1. K-fold cross-validation analysis for time series ($k = 5$) for evaluating the performance of GP regression against Bayesian linear regression with automatic relevance detection (ARD) and Ridge variants. We compare the cross-validated scores in terms of mean squared error (MSE) and mean negative log likelihood (MNLL) for the data from both the BO and RG experiments. The downward arrows indicate that a lower score is better for the metrics. Bold font indicates best performance.

	BO		RG	
Model	MSE (↓)	MNLL (↓)	MSE (↓)	MNLL (↓)
Gaussian process	**0.776**	4.308	**1.056**	1.906
Bayesian GLM (ARD)	1.223	1.798	1.734	3.767
Bayesian GLM (Ridge)	1.226	**1.398**	1.062	**1.579**

stimulation protocols that lead to highly rated patient perceptions, through determining which of the 40 electrodes to stimulate and with how much current, under safety constraints. This proof-of-concept study demonstrates the power of high-dimensional BO with trust regions in quickly converging to optimal stimulation protocols based on direct patient feedback, providing a more efficient search for stimulation parameters in clinical settings. Compared to random search, BO has learnt to allocate less current per stimulated electrode, despite higher total current allocation and a higher number of electrodes stimulated on average, suggesting a potential advantage of BO in minimizing the mean current per electrode without being explicitly steered towards such a goal. Additionally, BO has learnt to allocate higher current for more effective electrodes. Thus, BO has been shown to be useful, especially for identifying effective electrodes and informing about perception quality in multi-electrode stimulation for fast onboarding of patients in clinical settings, given its alignment with the findings of the previous clinical studies. Therefore, BO provides an efficient method for initial systematic experimentation with a patient for testing the effectiveness of multi-electrode stimulation protocols, which could have otherwise been an exhaustive and time-consuming search process.

It is important to highlight that the optimization here was geared toward maximization of the patient's perceptual quality rating of the phosphenes perceived upon the stimulation presented. Therefore, the convergence to optimal stimulation protocols in this study refers to the fact that the stimulation protocols generated towards the end of the optimization process are based on an improved modeling of the relationship between the applied stimulation and the observed patient response. While the use of phosphene perception quality ratings was a design choice for the optimization objective, other objectives can also be used, all of which have their own trade-offs. In this study, we focused on the objective of perceptual quality since a general understanding of the effectiveness of electrodes and their current levels for effective perception in the multi-electrode stimulation setting was sought through a quick, efficient search. An alternative objective could be maximizing the interpretability ratings of induced percepts, which would aim for meaningful pattern recognition, again subject to the personal judgment of the patient. This could be a logical next experiment, building on the general insights collected on electrodes. It would introduce target visual inputs for recognition that the current study lacked, at the cost of a more detailed search for the meaningful perception of specific visuals, hence requiring more experimentation time to gain insights on a range of specific visual inputs. An alternative approach would be to use neural rather than behavioral responses as an optimization criterion. Here, one may consider neural correlates of perceptual judgments or direct reconstructions of visual input from neural responses in downstream visual areas as targets (Chen *et al* 2020, Dado *et al* 2024). This would enable automated closed-loop control via an objective metric not prone to human misjudgment, although without a direct explicit connection to the perceptual experience of the patient. This could be preferred for full automation of the optimization pipeline, albeit at a loss of informativeness and interpretability of the metric driving the optimization. Future work should investigate the impact of BO in conjunction with these alternative objectives.

Regardless of the chosen optimization objective, it is important to consider the nature of the relationship between stimulation and the response that will be collected as feedback to drive the BO process. While a GP model is used in this study to better capture the expected non-linear relationship between the cortical stimulation and the patient rating on perceptual quality, the competitive performance of the Ridge GLM variant in model evaluation suggests that linear models could also be helpful in capturing seemingly complex relationships. For choosing the most appropriate model, one approach to consider could be the decomposition of non-linear processes into linear components, which may enable a choice for simpler models. For the cortical visual prostheses, however, the nature of processing is not yet clear. As cortical

neuroprosthetic vision targets cortical neurons, non-linearities may emerge due to neural adaptation, spatial summation, or interaction effects between cortical areas, especially at higher intensities or with prolonged stimulation. Phosphene perception may also not scale linearly with stimulus intensity, as suggested for brightness perception in previous research (Evans *et al* 1979). Since the context of stimulation via the occipital cortex goes beyond the natural visual processing through the retina and the visual pathway, more research is needed to uncover the nature of these expected non-linearities.

While the impact and benefits of the BO are clear in this proof-of-concept study, due to the iterative nature of the algorithm, the patient is required to be stimulated repetitively. This repetition, though aligned with the inherent design of the cortical visual neuroprosthesis, could be a challenge in interpreting the performance of the BO framework in the context of neural stimulation, where model updates are based on patient feedback. Due to adaptation effects as a result of repetitive stimulation or psychological factors such as motivation and attention, patient ratings of the stimulus may decline over time, reflecting a decrease in the responsiveness of the sensory system or a diminished interest from the patient in continuing the experiment. While these effects may make it harder for BO to get accurate feedback on patient perception, thus slowing down its convergence or causing it to converge to a suboptimal model of the patient response, it still manages to converge towards the right direction due to an overall decline in patient ratings despite consistent total current levels. Ideally, however, such adaptation effects should be taken into account by adjusting the stimulation over time so that sufficient current is administered to ensure the perception of the same stimuli with a higher current over time. Alternatively, a follow-up study could be conducted on electrodes found to be more effective based on BO in order to test their contribution to perception more accurately, based on a selection from a smaller subset of electrodes and before any exposure to prior stimulation. Another means to deal with adaptation and psychological effects could be by administering the iterative stimulations with a greater time interval in between, or with regular breaks between multiple consecutive stimulations, to minimize the side effects of repeated stimulation.

In order to reach more definitive conclusions about the effectiveness of electrodes and current allocation required for highly rated perception, some limitations should be taken into account and mitigated for future work, apart from the adaptation and psychological effects. First of all, due to the limited availability of implanted blind patients, the study was conducted only with a single patient; hence, whether the use of BO can generalize to more patients could not be investigated. Second, the time of testing available with this single patient was also limited, which allowed us to only run both experiments once, with BO followed by the RG experiment immediately after. Ideally, for each experiment, the same experimental design could be repeated multiple times with enough time difference in between so that more conclusive results could be reached about the effectiveness of electrodes and current allocation required for highly rated perception over multiple runs. Even if this were not possible, an alternative would be having both experiments again but this time with RG being tested before BO to eliminate the effects of order. However, limited patient testing time has not made these ideal scenarios possible, thus potentially causing a bias in our results more in favor of the BO experiment due to less exposure to prior stimulation, hence possibly leading to higher patient ratings or more attentiveness to perception.

Apart from these limitations, another means to reach more definitive conclusions could be through incorporating information on phosphene thresholds for the electrodes when stimulated simultaneously, so that it can be known roughly how much of the current allocated per electrode is effective, especially for electrodes that always have individual phosphene thresholds when stimulated alone. This could be interesting to investigate for future work.

In terms of the design of the BO procedure, different alternatives were considered and could provide a route for future work. First, preferential BO (Brochu *et al* 2010a, González *et al* 2017), where the patient is asked to provide a preference between two different stimuli at each iteration, can be used for incorporating feedback from the patient, as it is easier for human subjects to judge a relative comparison between two stimuli more accurately rather than judging one single stimulus across a series of stimuli in absolute terms (Miller 1956, Kahneman and Tversky 1979, Seymour and McClure 2008). While in the context of neural stimulation such preferential comparison will still give rise to order effects between the two stimuli, they could be potentially mitigated by the above-mentioned solutions. Second, instead of only maximizing patient perception quality based on patient feedback, a second objective can be incorporated to minimize total stimulation current, since minimizing stimulation current is both desirable for safety-critical settings like neural stimulation and can also compensate for the tendency of BO to converge to maximum total current possible when only optimizing for patient perception. A challenge in combining two separate objectives in this case comes from the patient rating being an ordinal variable, whereas the total current is a continuous variable. Whereas there are studies investigating BO for mixed variable multi-objective problems (Sheikh and Marcus 2022), an implementation of this in neural stimulation context needs further analysis. Future work can consider these alternatives to potentially improve the effectiveness of optimization pipelines for neural stimulation.

To our knowledge, this is the first study in the context of cortical neuroprosthetic vision that optimizes stimulation parameters *in vivo* directly based on patient feedback, by focusing on the informative value of perception quality for clinical studies. It is also special in optimizing for more than a dozen parameters in the context of neural stimulation, beyond which typical BO tends to fail. We successfully demonstrate a high-dimensional optimization of 40 parameters, by adopting trust region-based BO. It is worth noting that while TurBO provides efficiency benefits in searching through the high-dimensional parameter space, it does so by letting go of global exploration via the use of trust region only around the best evaluation so far. Hence, more optimal solutions might exist, albeit at the cost of efficiency in the search, especially when considering the full array of 96 electrodes as the parameter search space as an obvious next step.

BO can also bring the possibility for closed-loop control of cortical neuroprosthetic vision one step closer through a potential incorporation of neural responses as the feedback to update the surrogate model. Future work should also focus on the incorporation of a visual processing pipeline, where a camera image is processed to generate stimulation parameters for a targeted phosphene vision, to evoke the desired perceptions in the brain of the blind patient for a complete optimization of cortical neuroprosthetic vision via BO. This proof-of-concept study here lays the groundwork for any future work in this direction for optimizing neuroprosthetic vision based on direct patient feedback in a high-dimensional search space.

5. Conclusion

We have proposed BOPhos, an experimental design for finding optimal patient-specific stimulation parameters for cortical neuroprosthetic vision using high-dimensional BO with trust regions. We demonstrated the power of BO in quickly converging to optimal parameters *in vivo* by testing on a blind patient implanted with a Utah array in the early visual cortex and using direct patient feedback on perception quality for steering the optimization towards maximizing a perceptual score under safety constraints. BO proved useful, especially in allocating higher current to more effective electrodes and leading to higher perceptual quality ratings. While a complete BOPhos would also require a visual processing pipeline to evoke targeted perceptions, this study establishes a foundation for future research by incorporating direct patient feedback for phosphene optimization in a high-dimensional search space for the first time. Our study contributes towards an efficient and systematic search for the huge parameter space of multi-electrode stimulation for fast onboarding of patients in clinical settings via an initial experimentation on the effectiveness of electrodes and optimality of stimulation parameters for perceptual quality.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 899287 (project NeuraViPeR). This publication is also part of the project Innovative Neurotechnology for Society with Project Number 17619 of the Crossover programme. This work was supported in part by Grants PDC2022-133952-100 and PID2022-141606OB-I00 from the Spanish 'Ministerio de Ciencia, Innovación y Universidades', and by grant CIPROM/2023/25 from the Generalitat Valenciana (Spain). This publication is also part of the project Dutch Brain Interface Initiative with Project Number 024.005.022 of the research programme Gravitation which is (partly) financed by the Dutch Research Council (NWO). This publication is also part of the project ROBUST: Trustworthy AI-based Systems for Sustainable Growth with Project Number KICH3.LTP.20.006, which is (partly) financed by the NWO, ASMPT, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

We thank Dr Max Hinne for helping out with the implementation of the statistical analysis.

ORCID iDs

Burcu Küçükoğlu 0009-0001-2247-6371
David Leeftink 0000-0002-9542-3334
Eduardo Fernandez 0000-0002-7052-6011

References

Aiello G, Valle G and Raspopovic S 2023 Recalibration of neuromodulation parameters in neural implants with adaptive Bayesian optimization *J. Neural Eng.* **20** 026037

Arlot S and Celisse A 2010 A survey of cross-validation procedures for model selection *Stat. Surv.* **4** 40–79

Berkenkamp F, Krause A and Schoellig A P 2021 Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics *Mach. Learn.* **112** 3713–47

Binois M and Wycoff N 2022 A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization *ACM Trans. Evol. Learn. Optim.* **2** 1–26

Bonizzato M *et al* 2023 Autonomous optimization of neuroprosthetic stimulation parameters that drive the motor cortex and spinal cord outputs in rats and monkeys *Cell Rep. Med.* **4** 101008

Bosking W H, Oswalt D N, Foster B L, Sun P, Beauchamp M S and Yoshor D 2022 Percepts evoked by multi-electrode stimulation of human visual cortex *Brain Stimul.* **15** 1163–77

Bosking W H, Sun P, Ozker M, Pei X, Foster B L, Beauchamp M S and Yoshor D 2017 Saturation in phosphene size with increasing current levels delivered to human visual cortex *J. Neurosci.* **37** 7188–97

Brindley G and Lewin W 1968 The sensations produced by electrical stimulation of the visual cortex *J. Physiol.* **196** 479–93

Brochu E, Brochu T and de Freitas N 2010a A Bayesian interactive optimization approach to procedural animation design *Proc. 2010 ACM SIGGRAPH/Eurographics Symp. on Computer Animation*, SCA'10 (Eurographics Association) pp 103–12

Brochu E, Cora V M and De Freitas N 2010b A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning (arXiv:1012.2599)

Chen S C, Suaning G J, Morley J W and Lovell N H 2009a Simulating prosthetic vision: I. Visual models of phosphenes *Vis. Res.* **49** 1493–506

Chen S C, Suaning G J, Morley J W and Lovell N H 2009b Simulating prosthetic vision: II. Measuring functional capacity *Vis. Res.* **49** 2329–43

Chen X, Wang F, Fernandez E and Roelfsema P R 2020 Shape perception via a high-channel-count neuroprosthesis in monkey visual cortex *Science* **370** 1191–6

Choinière L, Guay-Hottin R, Picard R, Lajoie G, Bonizzato M and Dancause N 2024 Gaussian-process-based Bayesian optimization for neurostimulation interventions in rats *STAR Protocols* **5** 102885

Cole E R, Connolly M J, Ghetiya M, Sendi M E S, Kashlan A, Eggers T E and Gross R E 2024 SAFE-OPT: a Bayesian optimization algorithm for learning optimal deep brain stimulation parameters with safety constraints *J. Neural Eng.* **21** 046054

Cooper S E and Netoff T I 2022 Multidimensional Bayesian estimation for deep brain stimulation using the SafeOpt algorithm *MedRxiv Preprint* (https://doi.org/10.1101/2022.01.30.22270042) (Accessed 23 September 2024)

Dado T, Papale P, Lozano A, Le L, Wang F, van Gerven M, Roelfsema P, Güçlütürk Y and Güçlü U 2024 Brain2GAN: feature-disentangled neural encoding and decoding of visual perception in the primate brain *PLoS Comput. Biol.* **20** e1012058

de Ruyter van Steveninck J, Güçlü U, van Wezel R and van Gerven M 2022a End-to-end optimization of prosthetic vision *J. Vis.* **22** 1–14

de Ruyter van Steveninck J, Nipshagen M, van Gerven M, Güçlü U, Güçlüturk Y and van Wezel R 2024 Gaze-contingent processing improves mobility, scene recognition and visual search in simulated head-steered prosthetic vision *J. Neural Eng.* **21** 026037

de Ruyter van Steveninck J, van Gestel T, Koenders P, van der Ham G, Vereecken F, Güçlü U, van Gerven M, Güçlütürk Y and van Wezel R 2022b Real-world indoor mobility with simulated prosthetic vision: the benefits and feasibility of contour-based scene simplification at different phosphene resolutions *J. Vis.* **22** 1

Dobelle W H and Mladejovsky M G 1974 Phosphenes produced by electrical stimulation of human occipital cortex and their application to the development of a prosthesis for the blind *J. Physiol.* **243** 553–76

Dobelle W H, Mladejovsky M G and Girvin J P 1974 Artificial vision for the blind: electrical stimulation of visual cortex offers hope for a functional prosthesis *Science* **183** 440–4

Eriksson D, Pearce M, Gardner J, Turner R D and Poloczek M 2019 Scalable global optimization via local Bayesian optimization *Advances in Neural Information Processing Systems* vol 32 (Curran Associates, Inc)

Evans J R, Gordon J, Abramov I, Mladejovsky M G and Dobelle W H 1979 Brightness of phosphenes elicited by electrical stimulation of human visual cortex *Sens. Process.* **3** 82–94

Fauvel T and Chalk M 2022 Human-in-the-loop optimization of visual prosthetic stimulation *J. Neural Eng.* **19** 036038

Fernandez E *et al* 2021 Visual percepts evoked with an intracortical 96-channel microelectrode array inserted in human occipital cortex *J. Clin. Invest.* **131** e151331

Fine I and Boynton G M 2024 A virtual patient simulation modeling the neural and perceptual effects of human visual cortical stimulation, from pulse trains to percepts *Sci. Rep.* **14** 17400

Garnett R 2023 *Bayesian Optimization* (Cambridge University Press)

Gelman A, Vehtari A, Simpson D, Margossian C C, Carpenter B, Yao Y, Kennedy L, Gabry J, Bürkner P-C and Modrák M 2020 Bayesian workflow (arXiv:2011.01808)

González J, Dai Z, Damianou A and Lawrence N D 2017 Preferential Bayesian optimization *Proc. 34th Int. Conf. on Machine Learning* (*Proc. Machine Learning Research* vol 70) ed D Precup and Y W Teh (PMLR) pp 1282–91

Grado L L, Johnson M D and Netoff T I 2018 Bayesian adaptive dual control of deep brain stimulation in a computational model of Parkinson's disease *PLoS Comput. Biol.* **14** 1–23

Granley J, Fauvel T, Chalk M and Beyeler M 2023 Human-in-the-loop optimization for deep stimulus encoding in visual prostheses *Proc. 37th Int. Conf. on Neural Information Processing Systems*, NIPS'23 (Curran Associates Inc)

Granley J, Relic L and Beyeler M 2022a Hybrid neural autoencoders for stimulus encoding in visual and other sensory neuroprostheses *Advances in Neural Information Processing Systems* vol 35, ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh (Curran Associates, Inc) pp 22671–85

Granley J, Riedel A and Beyeler M 2022b Adapting brain-like neural networks for modeling cortical visual prostheses *SVRHM 2022 Workshop @ NeurIPS*

Hinne M 2025a Bamojax: Bayesian modelling in JAX *Zenodo* https://doi.org/10.5281/zenodo.15038847

Hinne M 2025b An introduction to sequential Monte Carlo for Bayesian inference and model comparison-with examples for psychology and behavioural science *Behav. Res. Methods* **57** 125

Kahneman D and Tversky A 1979 Prospect theory: an analysis of decision under risk *Econometrica* **47** 263–91

Küçükoğlu B, Rueckauer B, Ahmad N, de Ruyter van Steveninck J, Güçlü U and van Gerven M 2022 Optimization of neuroprosthetic vision via end-to-end deep reinforcement learning *Int. J. Neural Syst.* **32** 2250052

Küçükoğlu B, Rueckauer B, de Ruyter van Steveninck J, van der Grinten M, Güçlütürk Y, Roelfsema P R, Güçlü U and van Gerven M 2025 End-to-end learning of safe stimulation parameters for cortical neuroprosthetic vision *J. Neural Eng.* **22** 046022

Lorenz R, Simmons L E, Monti R P, Arthur J L, Limal S, Laakso I, Leech R and Violante I R 2019 Efficiently searching through large tACS parameter spaces using closed-loop Bayesian optimization *Brain Stimul.* **12** 1484–9

Losanno E, Badi M, Wurth S, Borgognon S, Courtine G, Capogrosso M, Rouiller E M and Micera S 2021 Bayesian optimization of peripheral intraneural stimulation protocols to evoke distal limb movements *J. Neural Eng.* **18** 066046

MacKay D J C 1992 Bayesian interpolation *Neural Comput.* **4** 415–47

Miller G 1956 The magical number seven, plus or minus two: some limits on our capacity for processing information *Psychol. Rev.* **63** 81–97

Nagrale S S, Yousefi A, Netoff T I and Widge A S 2023 In silico development and validation of Bayesian methods for

optimizing deep brain stimulation to enhance cognitive control *J. Neural Eng.* **20** 036015
Neal R 2012 *Bayesian Learning for Neural Networks* (*Lecture Notes in Statistics*) (Springer)
Oswalt D *et al* 2021 Multi-electrode stimulation evokes consistent spatial patterns of phosphenes and improves phosphene mapping in blind subjects *Brain Stimul.* **14** 1356–72
Owen A B 1998 Scrambling Sobol' and Niederreiter–Xing points *J. Complex.* **14** 466–89
Rashidi B, Johnstonbaugh K and Gao C 2024 Cylindrical Thompson sampling for high-dimensional Bayesian optimization *Int. Conf. on Artificial Intelligence and Statistics* (PMLR) pp 3502–10
Rocca A, Lehner C, Wafula-Wekesa E, Luna E, Fernández-Cornejo V, Abarca-Olivas J, Soto-Sánchez C, Fernández-Jover E and González-López P 2023 Robot-assisted implantation of a microelectrode array in the occipital lobe as a visual prosthesis: technical note *J. Neurosurg.* **140** 1169–76
Sarikhani P, Ferleger B, Mitchell K, Ostrem J, Herron J, Mahmoudi B and Miocinovic S 2022 Automated deep brain stimulation programming with safety constraints for tremor suppression in patients with Parkinson's disease and essential tremor *J. Neural Eng.* **19** 046042
Schmidt E M, Bak M J, Hambrecht F T, Kufta C V, O'Rourke D K and Vallabhanath P 1996 Feasibility of a visual prosthesis for the blind based on intracortical microstimulation of the visual cortex *Brain* **119** 507–22
Seymour B and McClure S M 2008 Anchors, scales and the relative coding of value in the brain *Curr. Opin. Neurobiol.* **18** 173–8
Shahriari B, Swersky K, Wang Z, Adams R P and de Freitas N 2016 Taking the human out of the loop: a review of Bayesian optimization *Proc. IEEE* **104** 148–75
Sheikh H M and Marcus P S 2022 Bayesian optimization for mixed-variable, multi-objective problems *Struct. Multidiscip. Optim.* **65** 331
Snoek J, Larochelle H and Adams R P 2012 Practical Bayesian optimization of machine learning algorithms *Advances in Neural Information Processing Systems* pp 2951–9
Sobol' I M 1967 On the distribution of points in a cube and the approximate evaluation of integrals *USSR Comput. Math. Math. Phys.* **7** 86–112
Sui Y, Gotovos A, Burdick J and Krause A 2015 Safe exploration for optimization with Gaussian processes *Int. Conf. on Machine Learning* (PMLR) pp 997–1005
Thompson W R 1933 On the likelihood that one unknown probability exceeds another in view of the evidence of two samples *Biometrika* **25** 285–94
van der Grinten M, de Ruyter van Steveninck J, Lozano A, Pijnacker L, Rueckauer B, Roelfsema P, van Gerven M, van Wezel R, Güçlü U and Güçlüturk Y 2024 Towards biologically plausible phosphene simulation for the differentiable optimization of visual cortical prostheses *eLife* **13** e85812
Williams C K I and Rasmussen C E 2006 *Gaussian Processes for Machine Learning* vol 2 (MIT Press)
Zhao Z *et al* 2021 Optimization of spinal cord stimulation using Bayesian preference learning and its validation *IEEE Trans. Neural Syst. Rehabil. Eng.* **29** 1987–97