Below are the responses to major issues raised by the reviewers; other issues (figures, equations, nomenclature, implementation and run-time details, code release) will be addressed in revision:

Primary contribution of this paper (reviewer 2 and 3) The reviewer comments were helpful here. We think the most important and unique contribution of this work is to leverage the distribution of model parameters from previous experiments for efficient characterization of the neural interface. Importantly, we focus on a unique large-scale interface operating at single-cell resolution. Our work unifies and extends previous studies on closed loop experiments [Paninski et al., 2007] and exploration of model architectures [Real et al., 2017] by incorporating a prior on model parameters from previous experiments in many animals, recorded over many years. As large-scale and high-resolution devices become more common, similar multi-animal datasets will likely be available. However, we are not aware of other work that matches the present work in resolution or scale of data. We would emphasize these points in revision.

Application to other systems (reviewers 1 and 3) Although the results are presented in the context of primate retina, the methods do not rely on specifics of the retinal circuitry, and we expect they would be useful in other neural systems as well. Specifically, the similarity of artifact shape across experiments is likely governed by impedance at the tissue interface. Also, the relationship between spike amplitude and stimulation threshold may be general, and may depend only on the spatial configuration of the electrode and the cell. We think that these methods are relevant to Intra-Cortical Micro-Stimulation (ICMS) [Salzman et al., 1990] for proprioceptive feedback in somatosensory cortex for motor prostheses [Salas et al., 2018], Optogenetics [Shababo et al., 2013] or as reviewer 1 suggests, for DBS (though current devices do not approach cellular resolution).

Problem statement (reviewer 2) In the context of a bi-directional retinal prosthesis, this work addresses one of the major outstanding problems regarding characterization of electrical response properties using a small number of measurements of electrical stimulation (i.e., efficiency - reviewer 3). In our lab prototype, identification of location and type for around 500 cells requires a few minutes of spontaneous activity recordings, performed in parallel across 512 electrodes. However, for electrical stimulation, each electrode needs to be stimulated in isolation to avoid nonlinear interactions ($\sim$ 1.5 hours for 512 electrodes); this measurement thus scales linearly with the number of electrodes. The problem of electrical response calibration has not been addressed previously, primarily due to the much fewer number of stimulating electrodes in most existing devices (reviewer 2). However, with advent of larger arrays that can stimulate 1000s of electrodes [Dragas et al., 2017] with multi-electrode current patterns [Fan et al., 2018], naive response calibration in the clinic may be far too time-consuming. To make these devices usable, it will be necessary to substantially reduce the calibration time, making methods such as the ones presented here crucial.

Relationship to prior work on neural interfaces (reviewer 2) Even though we mention most of the prior works in the context of the methods, we failed to include enough information about prior works on spike sorting. Mena et al., 2017 use previously recorded spike waveforms to jointly estimate the cellular activity and artifacts, with multiple trials of a single stimulation current. A Gaussian process prior is used for smoothly extrapolating the artifact across current values. O'Shea et al., 2018 only estimates the stimulation artifact (they do not assign spikes to cells), exploiting artifact similarity for a given stimulation electrode, across different pulses, trials and different recording electrodes. In contrast, this work performs joint estimation of artifact and spikes, exploits the similarity of the artifact across experiments and stimulating electrodes, and does not require an increasing sequence of current values.

Shortcomings (reviewer 1 and 3) We agree that the limitation of the study should be addressed directly in the manuscript. (1) We focus on minimizing the number of electrical stimulations, it will be important in future to minimize computational runtime as well. (2) For spike sorting, linear super-position of spikes and artifacts fails when recording amplifiers saturate. (3) The artifact is characterized using stimulation of a given current pattern in previous experiments, thus, the method is inapplicable to novel stimulation patterns. (4) For response modeling, the relationship between single electrode spike amplitude and stimulation threshold must be generalized to use spike amplitudes and simultaneous stimulation from multiple electrodes. (5) The method should account for differences in activation curve slopes for axons and somatic activation. (6) Analysis of linear decoding [Brackbill et al., 2018, Warland et al., 1997, Stanley et al., 1999] for estimation of prosthesis performance should be modified to incorporate nonlinear methods, which can yield higher performance [Parthasarthy et al., 2017]. (7) Response modeling and adaptive stimulation are validated only in simulation, where ground truth is available. (8) Each algorithm is analyzed in isolation, but the combined improvement from using all three should be evaluated (this is difficult due to lack of ground truth activation probabilities). These caveats will be included in the paper, subject to space limitations.

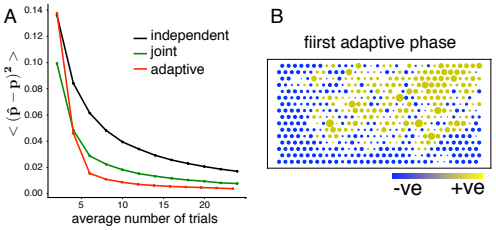

Figure 1: **Another dataset (reviewer 3).** A, B Same as Figure 3B, 4C in paper.