2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Unsupervised Channel Compression Methods in Motor Prostheses Design

Abdullah Alothman[1], Vikash Gilja[*,1]

Abstract— **The development of high performance brain machine interfaces (BMIs) requires scaling recording channel count to enable simultaneous recording from large populations of neurons. Unfortunately, proposed implantable neural interfaces have power requirements that scale linearly with channel count. To facilitate the design of interfaces with reduced power requirements, we propose and evaluate an unsupervised-learning-based compressed sensing strategy. This strategy suggests novel neural interface architectures which compress neural data by methodically combining channels of spiking activity. We develop an entropy-based compression strategy that models the population of neurons as being generated from a lower dimensional set of latent variables and aims to minimize the loss of information in the latent variables due to compression. We evaluate compressed features by inferring the latent variables from these features and measuring the accuracy with which the activity of held out neurons and arm movements can be estimated. We apply these methods to different cortical regions (PMd and M1) and compare the proposed compression methods to a random projections strategy often employed for compressed sensing and to a supervised regression based channel dropping strategy traditionally applied in BMI applications.**

I. Introduction

Brain machine interfaces (BMIs) have the potential to help individuals with functional impairments, such as loss of motor control, due to neurological disease or spinal cord injury [2], [7], [8]. BMIs map brain signals acquired in relevant brain regions to patient intent to enable functional restoration. In previous studies, BMIs have enabled patients to control robotic arm movements [1], and type by translating brain signals directly into text [3]. Intracortical BMIs record and sample brain signals from relevant regions of the brain at rates high enough to process both local field potentials (LFP) and action potentials (spikes). In particular, spike counts in single-unit [1] and multi-unit [3], [6], [7] extracellular recordings have been used as input features in high-performance BMIs. To that end, increasing the number of simultaneously recorded electrodes is important for advancing prosthetic performance and for studying the dynamics of neural population with increasing precision. Thus, over the last few decades translational and exprimental neuroscience studies have sought to record from larger populations of neurons and technologies to enable basic studies have advanced to address this need. For example, Neuropixel 2.0 can simultaneously record spiking activity from 758 sites out of 10,240 available sites across two 4-shank probes [10]. The Argo, designed with 65,536 channels, has been applied to record spiking activity from 791 neurons and cortical surface LFP activity from over 30,000 channels [11]. While the number of available recording sites in recent technological advances continues to increase exponentially [12], the number of simultaneous recording channels of spiking activity in clinical applications continues to lag behind. Broad clinical application will require integrated device designs that facilitate implantation by minimizing or eliminating the need for wired connections to devices, necessitating integration with active electronics. However, simultaneous high resolution measurement of spiking activity is power intensive, as it typically requires signal filtering, amplification, digitization, and telemetry. The power requirements introduce a critical constraint on the number of recording sites due to the challenges of power delivery and the negative impacts of brain tissue heating [13].

Several studies proposed strategies to decrease power requirements of each recording channel by relaxing signal processing requirements. Nason et al. showed that system power consumption is reduced by replacing the spike count feature traditionally employed for neural decoding with a spike band power feature; for motor decoding they demonstrate performance comparable to that achieved with spike count features [2]. Even-chen et al. showed that power requirements can be reduced by relaxing design specifications of more traditional architectures, which results in added noise to spike count features, but has negligible impact on decoding performance for motor BMI applications [7].

In this study, we present and evaluate a complimentary strategy in which input neural feature channels can be combined or dropped in a principled and unsupervised manner to reduce the number of output channels without sacrificing application specific BMI performance. Such a strategy suggests opportunities for the redesign of invasive neural interfaces which could enable higher neural feature channel count recording with potential power savings through sub-linear scaling of power requirements. As a starting point for this strategy, we develop and apply methods to single-unit spike count features, which form a set of initial sensing channels. Specifically, the goal is to apply unsupervised-learning-based channel compression methods to combine these channels while minimizing loss of information. By simply adding spike count features to simulate combined channels, we reduce the channel count while maintaining good scoring on multiple evaluation metrics that estimate

The authors declare the following financial interests which may be considered as potential competing interests with the work reported in this paper: VG holds shares in Neuralink Corp. and is an advisor/options holder at Paradromics, Inc.

[1] Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92092, USA.

[*] Correspondence: vgilja@ucsd.edu

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

information in the original and compressed channels. The channel compression algorithm we develop is predicated on the assumption that a set of lower-dimensional (low-D) latent variables can effectively describe the activity of a higher-dimensional (high-D) population of neurons. Previous work demonstrated that low-D latent variables can capture neural variability effectively, relate better to behavioral features, and enable better performance in online BMIs than traditional features based directly on (high-D) neural observations [6], [8]. Trautmann et al. reported that the neural population dynamics, as described by low-D latent variables, can be accurately estimated from unsorted multi-units [5]. That study connects unsorted multi-units, which are mixtures of single units, to the theory of random projections and compressed sensing. This provides motivation for us to study whether informed mixtures of single-unit features, based on designed compression methods, can yield high accuracy estimates of low-D latent variables with even lower numbers of channels. This would allow neural features to be compressed to a fewer number of channels, enabling the design of neural prostheses that utilize such compression to potentially reduce power requirements, without compromising the application specific performance requirements. To quantify the loss of information due to compression, we introduce two evaluation metrics to assess the latent variables inferred from compressed features. The first metric measures how well the latent variables can represent the original neural features by predicting the activity of a neuron left out during the variables' inference. The second metric evaluates how accurately the latent variables can explain and relate to behavioral features; in the case of the empirical motor physiology data tested in the current study, the behavioral features are arm movement kinematics. Lastly, as the proposed method is an unsupervised method for low-D feature generation, we compare it to supervised feature selection methods similar to those that have been employed in previous motor BMI studies.

II. Methods

A. Dataset

The feature compression methods described were empirically evaluated with the publicly available ”pmd-1” dataset from the Collaborative Research in Computational Neuroscience [1]. This dataset includes extracellular recordings from two Utah Multi-electrode Arrays (MEAs) implanted in dorsal premotor cortex (PMd) and primary motor cortex (M1) of a monkey engaged in an upper limb reaching task. During the task, the monkey controlled a cursor on screen to acquire sequentially appearing targets. Targets appeared at random locations within a 5.15 cm radius of the previous target. The workspace was 20 cm x 20 cm with 2 cm square targets. There are 496 sequential reach trials in the dataset, split 80% for training the model utilized in this study and 20% for testing and measuring the evaluation metrics. From the MEAs, 93 neurons from PMd and 67 from M1 were manually sorted [1]. Before any compression method is applied, each input channel is the spike count of a single neuron across time. Once a compression method is applied, multiple neurons can be combined and mapped to a single output channel. Hence, in the methods and results we will use the term channels to refer to output channels that individually are spikes counts from one or more combined neurons across time.

B. Channel-wise compression by combination

In this work, spiking features are compressed by summing spike counts from multiple neurons. We define the uncompressed neural spike counts by $Y \in \Re^{N\times T}$ such that N is the number of initial sorted neurons and T as the number of timepoints within the dataset. We also define $Y_D \in \Re^{D\times T}$ as the compressed feature set with D representing the number of channels after compression. Therefore, we define the M matrix such that:

$$Y_D = M * Y \tag{1}$$

M matrix formulation is the main goal of this work. The M matrix compresses the original dataset by combining input channels and is formulated by different strategies. The goal is to develop strategies that maintain information about the spiking of the original population of single neurons in Y while reducing D.

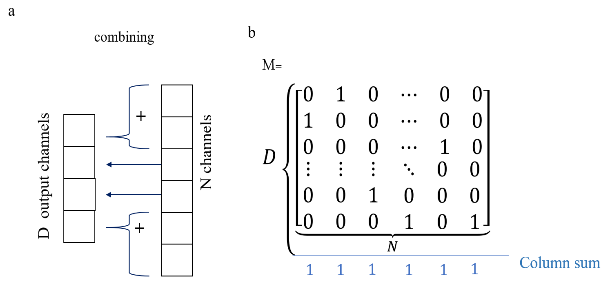

Fig. 1: a) Graphical illustration of channel compression as defined in this work, b) Example M matrix structure.

We impose several constraints on the design of the M matrix:

- The M matrix is applied to combine spike counts in channels, thus M must be binary.
- Channels from the N neurons in Y can be mapped to only one of the D channels in Y_D.
- All channels must be mapped to an output channel. No channels are dropped in the compression, thus the column sum of the M matrix has to be 1. Note: we consider dropping channels later in the Discussion.

C. Greedy dimensionality reduction

As an initial examination of compression strategies, we consider taking a greedy approach in which channel count is iteratively reduced. Channels are initialized as sorted single neurons and at each iteration one pair of channels is combined to reduce the channel count by 1. The pair is selected by evaluating a metric defined by a designed compression method. Channels are combined iteratively until a desired number of channels remain.

D. Latent neural trajectory estimation

After each M matrix is produced for a compression method, we extract trajectories of the latent variable from compressed channels to be evaluated. Trajectories are estimated by applying Gaussian Process Factor Analysis (GPFA). GPFA is a generative probabilistic model developed by Yu et al. [4] for application to neural spiking data that unifies learning temporal smoothing and dimensionality reduction parameters with the goal of extracting latent variable trajectories that describe shared variability in the original high-dimensional data. The latent variables in GPFA is defined here as $Z \in \Re^{Q \times T}$, where $Q < N$. Therefore, the relationship between the observations and the latent variable at time $t \in T$ is given by:

$$y_{:}^{t} | z_{:}^{t} \sim \mathcal{N}(Cz_{:}^{t} + \mu, R) \tag{2}$$

Where $z_{:}^{t}$ represents all the latent variables at timepoint t, $C \in \Re^{N \times Q}$ is the factor loading matrix, $\mu \in \Re^{N \times 1}$ is the mean of the neural observations, and $R \in \Re^{N \times N}$ is the covariance. Each latent variables is correlated with itself across time through a Gaussian process (GP):

$$z_{i}^{:} \sim \mathcal{N}(0, K_i) \tag{3}$$

where $K_i \in \Re^{T \times T}$ is the covariance matrix of the ith Gaussian process. The smoothing properties is determined by the choice of the form of the GP covariance. The form of K chosen is a squared exponential covariance function. The neural states can be inferred by:

$$E[\bar{z} \mid \bar{y}] = \bar{K}\bar{C}'(\bar{C}\bar{K}\bar{C}' + \bar{R})^{-1} * (\bar{y} - \bar{\mu}) \tag{4}$$

In equation 4, $\bar{y}$ and $\bar{z}$ are concatenations of features of neural observations and neural states respectively across all timepoints T. $\bar{C}$ and $\bar{R}$ are T blocks of C and R, and $\bar{K}$ is composed of Q sub-matrices where each sub-matrix along the diagonal is the covariance between two timepoints in the neural state i, $\bar{K}_{i,(t_m,t_n)}$ (more details in [4]). Using GPFA, we infer the low-D trajectory (i.e. the latent variables across time) based upon Y and the compressed Y_D. To facilitate direct comparison of models, GPFA parameters are learned once from the original uncompressed dataset Y and the compression matrix M is applied to the learned parameters to find the parameters of the reduced dataset Y_D:

$$y = Cz + \mu + \epsilon \rightarrow y_D = My = MCz + M\mu + M\epsilon \tag{5}$$

Therefore the parameters of the GPFA model of the reduced dataset are given by: $C_D = MC$ and $\mu_D = M\mu$. The covariance $R_D = cov(M\epsilon) = MRM'$. The M matrix compresses neural observations only, while the latent space is unchanged, thus the compression does not change K. We employ 4-fold cross-validation to the dataset and train four separate GPFA models. The parameters for each model are used in the compression methods, described below, to learn 4 different sets of M matrices for each method.

E. Evaluation Metrics

After extracting the trajectories of the latent variables for the compressed dataset $E[z \mid y_D]$ and the original dataset $E[z \mid y]$, we assess the impact of compression by evaluating two performance metrics as a function of the number of compressed channels. The metrics evaluate how well the latent variables can represent the high-D neural features (original channels) and the behavioral features.

Leave-out neuron goodness-of-fit (prediction error): This evaluation metric is adapted from Yu et al. [4]. The metric estimates the activity of neuron j from the latent variables inferred from all other neurons in the testing set. That is, neuron j is dropped, then the remaining neurons are used to infer the latent variables which are then used to predict the activity of neuron j. Both inference and prediction of the test set are done with the model parameters learned from the training set. The prediction error is computed by measuring the sum of squared error between the original neuron and its prediction. This metric evaluates how much the low-D latent variables represent and predict the neural observations. As we iteratively reduce the number of channels, we expect this error metric to monotonically increase.

Kinematic reconstruction error: Kinematics reconstruction is an important evaluation metric for motor prostheses design. We seek evaluate how well the low-D population encodes the behavior. The behavior in this work is the hand position as the monkey is engaged in the reaching task. We use optimal linear estimators to build a simple linear relationship between the behavior and the trajectories of the latent variables inferred from either the original or compressed channels [9].

F. Compression methods

Finally, we discuss the compression methods used to combine channels. These methods vary between operating on the neural observations alone or both the latent variables and observations. In the current work our compression is based upon the relationship between the neural population and the latent variables without considering correlations across time. That is, the current methods do not leverage relationships learned by the GP component of GPFA. The joint distribution between the observations and the latent states is given in equation 6.

$$\begin{bmatrix} y \\ z \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} d \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & C' \\ C & CC' + R \end{bmatrix}) \tag{6}$$

Private noise: In this method, we combine channels based on their private noise as given by the covariance matrix R. Channels with the highest private noise are combined together in order to minimize the effect of multiple private noisy channels in latent variable estimations.

Conditional entropy: Entropy is the measurement of uncertainty in a random variable. We utilize conditional entropy to evaluate how much of the latent variable entropy is affected by the choice of compression for the observation channels. That is, we measure the conditional entropy, given in equation 7, of the latent variable given the current channel

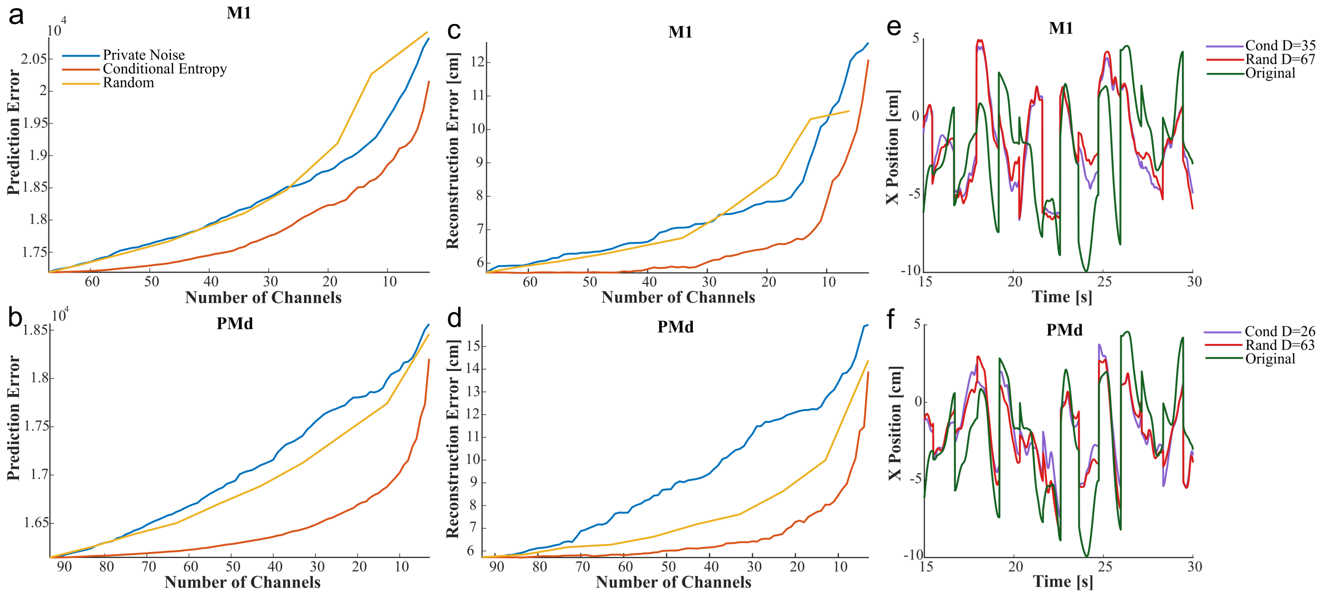

Fig. 2: Prediction error is shown for a) M1 and b) PMd through iterative greedy combinations for private noise and conditional entropy methods, as well as application of a random projections base approach. c) and d) Kinematic reconstruction error is shown for M1 and PMd, respectively. e) and f) Reconstructed x-position for two similarly performing M Matrices is shown, one for random projections and one for the conditional entropy method for both M1 and PMd, respectively.

set. In each greedy iteration, we evaluate which combination of channels has the lowest conditional entropy.

$$H(Z \mid Y) = H(Z,Y) - H(Y) \tag{7}$$

We know that by combining channels, we lose mutual information and thus the conditional entropy is expected to increase monotonically. At each iteration, we select the channel pair that minimizes this increase in entropy. This allows us to reduce the channel count at each iteration while minimizing loss in mutual information as defined by the GPFA model.

III. Results

Based on the complexity of the task, we chose the latent dimension Q = 20 as previous studies with similar recordings suggest that it is sufficient to capture sufficient neural variance from the population [4], [8]. To evaluate how well the proposed compression methods performed, we simulated a random projections approach as a control. This control analyses generated 100 randomized M matrices for all reduced output dimension D. These randomized M matrices are assessed through all of the trained models in the 4-fold cross validation. Figure 2 shows both evaluation metrics for M1 (a,c) and PMd (b,d) data. The plots shown for private noise and conditional entropy are the average of the 4-fold cross-validation, while the random plot is the average per reduced dimension of the 4-fold and across all random M matrices. The plots suggest that combining channels based on private noise is a poor choice as it often performs worse than random projections with respect to both metrics while the conditional-entropy-based method outperforms all of the other methods. Sub figures e) and f) show reconstruction of the x-axis of hand position for two similarly performing M matrices in both M1 and PMd respectively. Qualitatively, it can be observed that conditional entropy has similar reconstruction to randomized M matrices with almost half the number of channels required.

IV. Discussion

We used two different strategies to compress the data and evaluate how well the latent variables can be inferred after compression. These inferred variables are assessed through evaluation metrics that measure their ability to 1) represent the original single unit neural observations and 2) encode behavioral features. Overall, the conditional entropy method, which optimizes compression using an iterative greedy approach, achieves good results compared to a random projections based approach.

To further evaluate the efficacy of compression that combines single unit channels with the conditional entropy method, we compare its performance to that of more traditional neuron dropping approaches (Figure 3). We apply the novel conditional entropy approach to drop the least informative neurons at each iteration (retaining the subset of neurons that increase conditional entropy the least at each iteration without forming multiunit channels). The goal of this approach is to retain the most informative single unit channels at each iteration. We also mix the single unit combination and channel dropping approaches in an algorithm that chooses which method increases conditional entropy the least at each iteration.

We compare these unsupervised single unit combination and single unit dropping approaches to a single unit dropping approach that utilizes a supervised regression model; by applying a regression model between the hand position and the neural observations, we drop the least significant neural

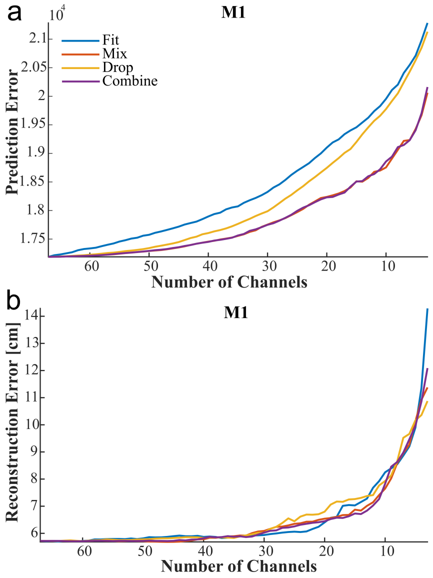

Fig. 3: a) Prediction error and b) reconstruction error in M1 for the unsupervised combination, unsupervised feature dropping, and supervised regression models.

features (based upon p-values from the regression model). This approach resembles supervised feature selection methods often employed for BMI applications. Surprisingly, with respect to kinematic reconstruction error, the unsupervised method yields similar performance to the supervised method. Critically, the unsupervised approach, unlike the supervised approach, operates without any knowledge of the behavioral features. Figure 3 also suggests that latent variables estimated through combinations of single units are better representative of the original single unit observations than the single unit dropping approach and the supervised method. Lastly, with respect to both metrics, our analysis of motor control data suggests that unsupervised combination outperforms methods that retain single units by a channel dropping approach. Note the channel dropping approach resembles strategies that are readily employable by conventional interface architectures, which often utilize multiplexers to facilitate channel down selection. This result suggests that if output channel dimensionality is limited, neural interface architectures that facilitate data-driven combination of single unit channels may yield superior performance to system designs that are limited to single unit channel down selection only.

This work describes a strategy for compressing the number of output spike count channels based upon summing groups of single neuron channels to combine them into multiunit channels. By applying this strategy to neurophysiological data, we demonstrate that output channel count can be reduced with an unsupervised approach by choosing channel combinations that minimize loss of information in an inferred low-D latent variable space. By relaxing the requirement to transmit single neuron spike counts, our strategy suggests neural interface system designs with the potential to achieve sub-linear power scaling with respect to recording channel count. Such power-efficiency gains can enable the development of clinically viable, fully implantable neural interfaces with increased application-specific performance, such as more accurate and robust functional motor restoration[14].

Moving forward, the models described and results presented motivate a number of future directions.

- In the future we aim to augment the methods by accounting for temporal dynamics of the latent variable.
- We used spike counts of sorted single-unit recordings as the features in this study. In the future we plan to study how well the methods developed here extend to neural features that enable power-efficient implementations through modifications of neural interface architectures.
- We will employ a variety of behavioral studies in multiple species to further evaluate the proposed techniques.

ACKNOWLEDGMENT

We would like to thank Krishna Shenoy and Eric Trautmann for conversations that helped shape the initial direction of this work.

REFERENCES

[1] Matthew G. Perich, Patrick N. Lawlor, Konrad P. Kording, Lee E. Miller (2018). Extracellular neural recordings from macaque primary and dorsal premotor motor cortex during a sequential reaching task. . CRCNS.org.http://dx.doi.org/10.6080/K0FT8J72.

[2] S. R. Nason et al., "A low-power band of neuronal spiking activity dominated by local single units improves the performance of brain–machine interfaces," Nat Biomed Eng, vol. 4, no. 10, pp. 973–983, Oct. 2020, doi: 10.1038/s41551-020-0591-0.

[3] F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy, "High-performance brain-to-text communication via imagined handwriting," bioRxiv, p. 2020.07.01.183384, Jul. 2020, doi: 10.1101/2020.07.01.183384.

[4] B. M. Yu, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, and M. Sahani, "Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity," Journal of Neurophysiology, vol. 102, no. 1, pp. 614–635, Jul. 2009, doi: 10.1152/jn.90941.2008.

[5] E. M. Trautmann et al., "Accurate Estimation of Neural Population Dynamics without Spike Sorting," Neuron, vol. 103, no. 2, pp. 292-308.e4, Jul. 2019, doi: 10.1016/j.neuron.2019.05.003.

[6] J. C. Kao, S. I. Ryu, and K. V. Shenoy, "Leveraging neural dynamics to extend functional lifetime of brain-machine interfaces," Scientific Reports, vol. 7, no. 1, Art. no. 1, Aug. 2017, doi: 10.1038/s41598-017-06029-x.

[7] N. Even-Chen et al., "Power-saving design opportunities for wireless intracortical brain–computer interfaces," Nat Biomed Eng, Aug. 2020, doi: 10.1038/s41551-020-0595-9.

[8] J. C. Kao, P. Nuyujukian, Stephen I. Ryu, M. M. Churchland, J. P. Cunningham, and K. V. Shenoy, "Single-trial dynamics of motor cortex and their applications to brain-machine interfaces," Nat Commun, vol. 6, no. 1, p. 7759, Nov. 2015, doi: 10.1038/ncomms8759.

[9] Salinas, E. Abbott, L. F. Vector reconstruction from firing rates.J. Comput.Neurosci.1,89–107 (1994).

[10] N. A. Steinmetz et al., "Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings," p. 44.

[11] K. Sahasrabuddhe et al., "The Argo: A 65,536 channel recording system for high density neural recording in vivo," bioRxiv, p. 2020.07.17.209403, Jul. 2020, doi: 10.1101/2020.07.17.209403.

[12] I. H. Stevenson and K. P. Kording, "How advances in neural recording affect data analysis," Nature Neuroscience, vol. 14, no. 2, Art. no. 2, Feb. 2011, doi: 10.1038/nn.2731.

[13] P. D. Wolf, "Thermal Considerations for the Design of an Implanted Cortical Brain–Machine Interface (BMI)," in Indwelling Neural Implants: Strategies for Contending with the In Vivo Environment, W. M. Reichert, Ed. Boca Raton (FL): CRC Press/Taylor and Francis, 2008.

[14] J. M. Carmena et al., "Learning to Control a Brain–Machine Interface for Reaching and Grasping by Primates," PLoS Biol, vol. 1, no. 2, p. e42, Oct. 2003, doi: 10.1371/journal.pbio.0000042.