Data Compression Versus Signal Fidelity Tradeoff in Wired-OR Analog-to-Digital Compressive Arrays for Neural Recording

Yan, Pumiao; Akhoundi, Arash; Shah, Nishal P.; Tandon, Pulkit; Muratore, Dante G.; Chichilnisky, E. J.; Murmann, Boris

DOI
10.1109/TBCAS.2023.3292058

Publication date
2023

Document Version
Final published version

Published in
IEEE transactions on biomedical circuits and systems

Data Compression Versus Signal Fidelity Tradeoff in Wired-OR Analog-to-Digital Compressive Arrays for Neural Recording

Pumiao Yan, *Graduate Student Member, IEEE*, Arash Akhoundi, *Student Member, IEEE*, Nishal P. Shah, *Member, IEEE*, Pulkit Tandon, Dante G. Muratore, *Senior Member, IEEE*, E. J. Chichilnisky, and Boris Murmann, *Fellow, IEEE*

***Abstract*—Future high-density and high channel count neural interfaces that enable simultaneous recording of tens of thousands of neurons will provide a gateway to study, restore and augment neural functions. However, building such technology within the bit-rate limit and power budget of a fully implantable device is challenging. The wired-OR compressive readout architecture addresses the data deluge challenge of a high channel count neural interface using lossy compression at the analog-to-digital interface. In this article, we assess the suitability of wired-OR for several steps that are important for neuroengineering, including spike detection, spike assignment and waveform estimation. For various wiring configurations of wired-OR and assumptions about the quality of the underlying signal, we characterize the trade-off between compression ratio and task-specific signal fidelity metrics. Using data from 18 large-scale microelectrode array recordings in macaque retina ex vivo, we find that for an event SNR of 7–10, wired-OR correctly detects and assigns at least 80% of the spikes with at least 50× compression. The wired-OR approach also robustly encodes action potential waveform information, enabling downstream processing such as cell-type classification. Finally, we show that by applying an LZ77-based lossless compressor (gzip) to the output of the wired-OR architecture, 1000× compression can be achieved over the baseline recordings.**

***Index Terms*—A/D conversion, analog-to-digital compression, brain-machine interfaces, compression algorithm, neural interfaces.**

Manuscript received 22 February 2023; revised 17 May 2023; accepted 25 June 2023. Date of publication 4 July 2023; date of current version 9 October 2023. The work of Pumiao Yan was supported by a Stanford Bio-X SIGF Fellowship. This work was supported in part by Stanford's Wu Tsai Neurosciences Institute and in part by the Dutch Research Council under Grant 024.005.022. This paper was recommended by Associate Editor V. Chen. *(Corresponding author: Pumiao Yan.)*

Pumiao Yan and Boris Murmann are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: pumiaoy@stanford.edu; murmann@stanford.edu).

Arash Akhoundi and Dante G. Muratore are with the Microelectronics Department, Delft University of Technology, CD 2628 Delft, The Netherlands (e-mail: a.akhoundi@tudelft.nl; d.g.muratore@tudelft.nl).

Nishal P. Shah is with the Department of Neurosurgery, Stanford University, Stanford, CA 94305 USA (e-mail: nishalps@stanford.edu).

Pulkit Tandon is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: tpulkit@stanford.edu).

E. J. Chichilnisky is with the Department of Neurosurgery and Ophthalmology, Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA 94305 USA (e-mail: ej@stanford.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TBCAS.2023.3292058.

Digital Object Identifier 10.1109/TBCAS.2023.3292058

I. Introduction

HIGH-THROUGHPUT and high-density neural interfaces will enable better technology to study, restore and augment functionality in the nervous system [1]. By providing simultaneous cellular-resolution recording of activity in large populations of neurons, these interfaces can shed light on the complex interactions between neurons and their cooperative behavior in a manner that has previously been unattainable [2], [3], [4], [5], [6], [7]. In addition, high bandwidth and single-cell resolution offer promising prospects for treating neurological diseases and expanding human sensory perception in clinical applications [8], [9]. Therefore, there is an increasing need to record more neurons over a longer duration in vivo [1]. The ideal system to fulfill this need must have many simultaneous recording channels and high temporal resolution, and must be fully wireless to provide stable recording for long durations.

In a typical fully implantable neural interface design, action potentials from neurons are captured using microelectrode arrays connected to multi-channel recording electronics (see Fig. 1(a)). Each recording channel undergoes amplification, filtering and digitization before being transmitted outside the implant for further processing. However, previous systems following such design are limited to approximately a thousand channels [10], [11], [12]. This is because as channel density and the number of simultaneous recording channels increase, processing and transmitting the huge amount of generated data given the power and area constraints of a fully implantable device becomes difficult or impossible. Leading designs high channel count devices such as the Neuropixels [13] and Argo [14] systems attempt to work around the issue with a switch matrix scheme [13], [15] or on-chip multiplexing [12], [13]. The Neuropixels system applies a switch matrix scheme to access more electrodes; however, the number of simultaneous recording channels remains the same [13], [15]. On-chip multiplexing increases the density of electrodes; however, the total amount of data to transmit remains unchanged [13], [14].

Take the Argo system, for example, each of its CMOS sensor arrays is able to simultaneously record 2048 channels. However, this is achieved by multiplexing all pixels that are sampled at the 32 kHz Nyquist-rate, to 32 high-speed analog outputs with output buffers driving long transmission lines to external

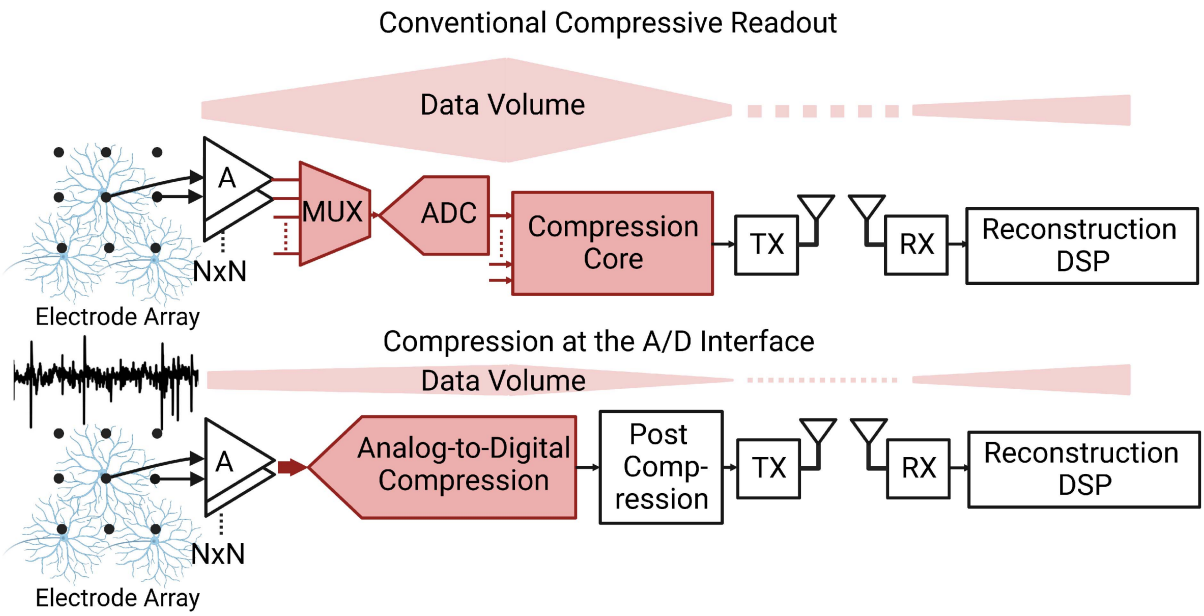


Fig. 1. System overview of single-cell resolution neural interfaces. Top: In conventional fully implantable neural recording systems, a data deluge is created at the A/D interface. Data reduces as the computation increases. Bottom: Compression happens at the A/D interface. This greatly reduces the amount of memory and computation before transmitting the data.

12-bit analog-to-digital converters (ADCs). As a result, such Nyquist-sampling-based systems generate an immense amount of data (Argo: ~ 0.8 GB/s), posing a great throughput challenge to the communication link. The substantial data rate necessitates the system to operate within the limitation imposed by tethering cables for effective data readout, consequently limiting its mobility and suitability for fully implantable applications.

Researchers have also investigated a range of on-chip compression approaches to address the throughput challenge. For cortical applications that only require binary threshold features to estimate low-dimensional manifolds, large power and data reduction is possible through thresholding [16], [17], [18], [19], [20]. However, this precludes off-chip spike sorting that is required to resolve the spikes originating in different cells of different types. Determining the appropriate per-channel threshold also requires substantial computational resources and power. To achieve data reduction without sacrificing such information, on-chip spike sorting [21], [22], [23] and compression [24], [25] have been considered. However, digitizing and moving the immense amount of data for such approaches to work is another major hurdle. Therefore, data compression needs to happen as close as possible to the analog-to-digital interface as shown in Fig. 1(b). In order to avoid a large data rate at any point in the system, part of the analog signal that is not important to the application must be removed using lossy compression [26].

In neural recordings, most of the information corresponding to the extracellular activities that are sensed and captured by the microelectrodes are in the spike waveforms [27]. Given the significance of these waveforms, one idea for data compression is to detect the spike times and only record samples in their vicinity (thus eliminating baseline samples between spikes). From a hardware perspective, a key issue with this approach lies in finding the proper threshold and managing the data movement with limited resources in a dense array. These issues are seen in previous works [18], [28], which use analog memory cells and additional computation to find the thresholds appropriate for spike detection in each recording channel.

Our previous work [29] overcomes this issue using a wired-OR analog-to-digital converter (ADC) array, reviewed in Section II A. In this architecture, samples are discarded based on a wired-OR competition between the pixels and no thresholding is needed. In [29], we demonstrated this technique with a simulation-based study in cell receptive mapping tasks and showed that for the three retina recording datasets, $\sim 40\times$ compression was achieved while missing less than 5% of cells in ex vivo primate retina recordings. Through previous analysis, we observed that wired-OR is capable of almost always recovering spike samples while discarding the baseline samples [29]. In the present study, we build upon our previous discoveries from cell-receptive mapping tasks by evaluating the performance of the wired-OR method in preserving key features for a wide range of neuroscience and neuroengineering applications.

For neural signal compression, it is crucial to distinguish the signal's salient information from unnecessary data samples, to efficiently extract the features used in various applications. In motor brain-computer-interface (BCI) applications, only threshold crossing of spikes is needed to decode very simple tasks [16] (Fig. 2(a)). For scientific studies investigating the behavior of individual neurons, identifying certain neural circuits, or enhancing the efficacy of brain-machine interfaces, it is necessary to distinguish spikes detected from different neurons (Fig. 2(b)). This procedure is commonly known as spike sorting [21], [30]. Finally, cell type identification is heavily dependent on the spike waveform shape over space (shown in Fig. 2(c)), i.e. the average electrical image (EI) of spikes from a cell [31], [32], [33], [34], [35].

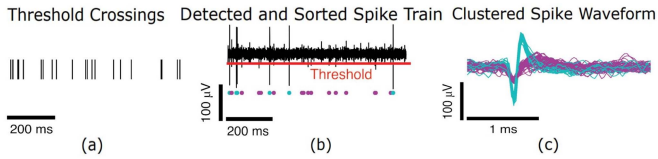


Fig. 2. Spike features for different applications.

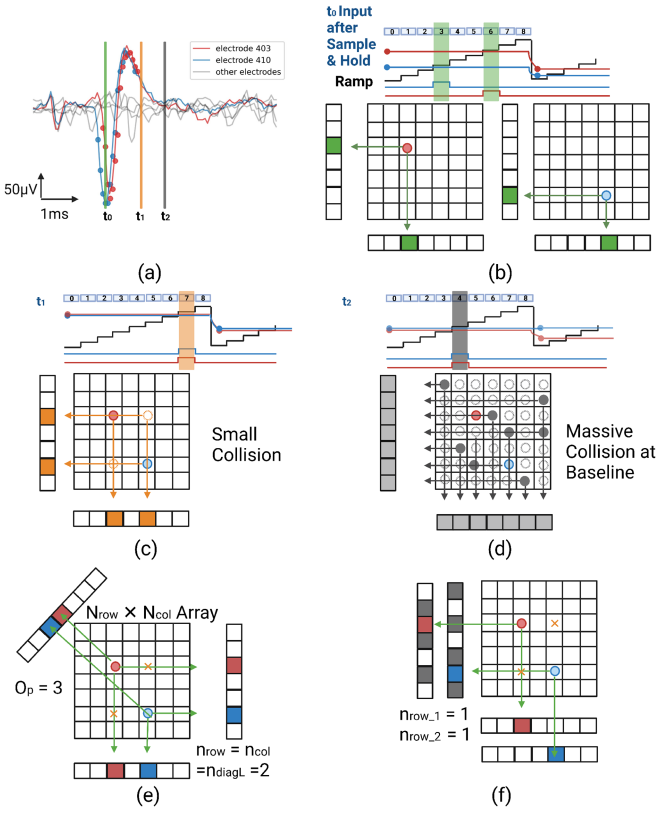


Fig. 3. Wired-OR readout concept. (a) A snippet of action potential waveform seen on different electrodes. (b) Conversion of voltage to pulse position and collision-free readout of one pixel. (c) A collision between two pixels. (d) Massive collision across the array at the baseline level. (e) Diagonal wiring conceptual drawing. (f) Interleaving wiring.

In [36], we explored generalizations and reconfigurability of the wired-OR architecture to assess its suitability for a broad range of applications. By constructing a single-electrode neural signal processing pipeline that incorporated commonly used methods, we sought to comprehend the tradeoff between performance and compression ratio through a simulation study of 3 ex vivo primate retina recordings.

This article is an extension of [36] focusing on the conceptual exploration of the wired-OR architecture. In addition to previous analyses:

- We extend the single-electrode analysis to consider multiple nearby electrodes as a unit patch, which better preserves the spatiotemporal information of spikes.
- We assess various task-specific signal fidelity metrics and investigate the performance variation across retina recordings.
- We explore the potential to compress the output of the wired-OR readout further.
- We assess the scalability of wired-OR to a higher number of channels.

This article presents an analysis of 18 ex vivo primate retina recordings representing a wide range of signal-to-noise ratio (SNR) scenarios. We compare the performance at each stage of processing (spike detection, spike classification and waveform estimation) with an event SNR metric for each electrode, developed in Section III, so that the results can be translated to any neural system provided the corresponding SNR is known. To better comprehend the performance variation as we scale up our analyses to more datasets, we also accounted for the cell density and average firing rate characteristics defined in Section V. These findings contribute to the future development of a data-driven approach for optimizing wired-OR configurations for different applications.

The remainder of this article is organized as follows: Section II presents the wired-OR compressive readout architecture and discusses the configuration space for different applications. Section III describes the signal fidelity simulation and analysis methods and introduces a range of quantitative metrics of interest to different neural interface applications. Section IV presents the simulation and analysis results, using ex vivo primate retina recordings. The results are further discussed in Section V. The wired-OR topology captures at least 80% of the spike waveforms with at least $50\times$ data compression across all 18 datasets.

II. Wired-OR Compressive Readout Architecture

A. Wired-OR Readout Concept

The wired-OR readout architecture simultaneously achieves analog-to-digital compression and channel multiplexing for neural recordings by exploiting the sparsity and diversity of neural signals (Fig. 3(a)). In the wired-OR readout architecture [29], each pixel conditions and samples the input as commonly done in neural interfaces. The sampled voltage is then converted into a pulse position, which is achieved by comparing it to a globally-distributed ramp step signal (see Fig. 3(b)). In the most basic implementation, the pulses from pixels in the same row or column are combined into single wires using wired-OR circuitry. In essence, signal compression occurs by having the pixels compete for these limited wire resources. If only a single pixel produces a pulse at a given time step (i.e., it is the only channel with a quantized voltage corresponding to the time step, see Fig. 3(b)), then the pixel location and its A/D conversion result (ramp counter state) can be uniquely recovered. On the other hand, if multiple pulses from different pixels occur at the same time step (i.e., the quantized voltages on two or more channels are equal) multiple rows and/or columns are activated (collision case in Fig. 3(c)) and the conversion results cannot be recovered (samples are discarded, which leads to the desired compression). As discussed in [29], this compression approach is effective for neural signals due to their long-tailed probability distribution. Voltage samples associated with spikes tend to be

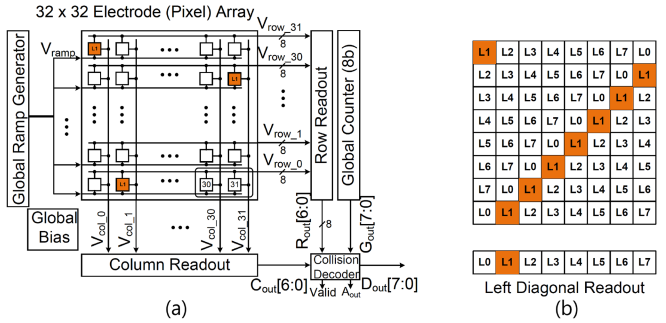


Fig. 4. Wired-OR wiring implementations. (a) Two-projection example. (b) Left diagonal wiring.

unique and are typically retained, while baseline samples falling within a certain voltage range tend to be discarded (see Fig. 3(d)).

B. Generalization of wired-OR Configurations

Wired-OR compressive readout architecture achieves data compression by discarding collision samples, which are most likely to occur near the baseline level of neural signals [29]. While massive collisions (Fig. 3(d)) occur for ramp values around zero, which do not correspond to useful information (no spike activity), small collisions (Fig. 3(c)) may still contain useful spike information. Different wiring configurations can be used to resolve small collisions and retain more spike information. Diagonal wiring [36] (see Fig. 3(e)) in addition to interleaving wiring [29] (see Fig. 3(f)) are assessed for their merits to our readout scheme.

Consider a collision case in which two pixels record the same voltage levels simultaneously (see t_1 in Fig. 3(a)). Through the wired-OR logic, the pixels are projected onto the row and column index of the data matrix (see Fig. 3(c)). In this case, the address of the channels is not uniquely decodable, resulting in a small collision. Different wiring configurations can be abstracted as different projections. One way to decode this small collision is by adding a right diagonal projection through extra wiring (conceptually shown in Fig. 3(e)). In a $N_{\mathrm{col}} \times N_{\mathrm{row}}$ array (32×16 in this work), the number of elements activated in each projection p are denoted in n_{p}. When $\mathrm{Max}(n_{\mathrm{p}}) \leq 2$, the triggered channels are uniquely decodable with one set of added diagonal wiring. A left diagonal projection can be added as another projection (the number of projections/wiring is denoted by O_{p}). For $O_{\mathrm{p}} = 4$, when $Max(n_{\mathrm{p}}) \leq 4$, the triggered channels are uniquely decodable. While there are more uniquely decodable cases when $\mathrm{Max}(n_{\mathrm{p}}) > 4$, for hardware simplicity, the decoding criteria of wired-OR is set only to consider cases when no more than four wires in each projection are activated. In its physical implementation, each pixel has O_{p} static connections to the periphery (four for $O_{\mathrm{p}} = 4$ with both left and right diagonal wiring): the first two connect the pixel by row and column, realizing x and y projections like in a traditional projection scheme as shown in Fig. 4(a). The diagonal projection mapping illustrated in Fig. 4 resembles the diagonal connection mapping for pixels. Similar to [37], [38], the relationship used to reconstruct the original pixel position from the data stream given each projected output $P_i (i \leq 4)$ is shown as follows:

$$\begin{cases} P_1(x,y) = x, & \text{column readout} \\ P_2(x,y) = y, & \text{row readout} \\ P_3(x,y) = (x+y)\ mod\ \mathrm{N}, & \text{left diagonal} \\ P_4(x,y) = (x+y\times(N-1))\ mod\ \mathrm{N}, & \text{right diagonal} \end{cases} \tag{1}$$

This solution effectively disentangles two or more channels recording the same spike voltage levels.

When multiple channels are recorded simultaneously through the wired-OR, a prefix code is needed to record the number of collision-free channels. Here, we use Huffman coding for the prefix. Effectively, at each ramp step, when collecting k uniquely decodable channels, the number of bits transmitted is:

$$S_{\mathrm{p}} = \text{Huffman Code} + k \times \sum_{O_{\mathrm{p}}} log_2(N_{\mathrm{p}}) \tag{2}$$

where N_{p} denotes the dimension of the corresponding projection. In simple scenarios where there is only one channel active, although there are more projections available, only sending two of the projection readout outputs can reduce the number of bits to be transmitted. The resulting data rate depends on the average number of bits transmitted, denoted as $\bar{S}_{\mathrm{p}}$. For a sampling frequency of f_{s}, the data rate is:

$$R_{\mathrm{p}} = \bar{S}_{\mathrm{p}} \times f_{\mathrm{s}} \tag{3}$$

In the previously proposed interleaving wiring [29], small collisions are solved by effectively separating the array into smaller sub-arrays that still use the row and column projections - see Fig. 3(f) for $\mathrm{W} = 2$ number of interleaving wires. The resulting data rate also depends on the average rate of decodable channels per sample, denoted as $\alpha_{\mathrm{d,p}}$:

$$R_{\mathrm{W}} = [log_2(N_{\mathrm{row}}/W) + log_2(N_{\mathrm{col}})]\alpha_{\mathrm{d,W}} f_{\mathrm{s}} \tag{4}$$

For B-bit ADC resolution, the corresponding compression ratio in both strategies is:

$$CR = \frac{N_{\mathrm{col}} \times N_{\mathrm{row}} \times B \times f_{\mathrm{s}}}{R} \tag{5}$$

Knowing the estimated average rate of decodable channels per sample in a neural dataset, one can calculate the output data rate of wired-OR with (3) and (4) for diagonal and interleaved wire configurations, and its corresponding compression ratio with (5). An optimal design should maximize collision events for baseline samples (maximum compression) and collision-free events for spike samples (maximum performance) [29].

III. Action Potential Signal Processing Methods

To understand how wired-OR compression affects the signal fidelity of action potential signal recordings in a broader range of applications, we establish a neural signal analysis pipeline that extracts various features of interest to different applications, as shown in Fig. 5. The recorded raw data is re-processed in software to simulate the wired-OR readout scheme. In addition to the traditional reconstruction and spike-sorting approach discussed in [29], a tailored neural signal processing pipeline is built to handle the compressed outputs of wired-OR better, also

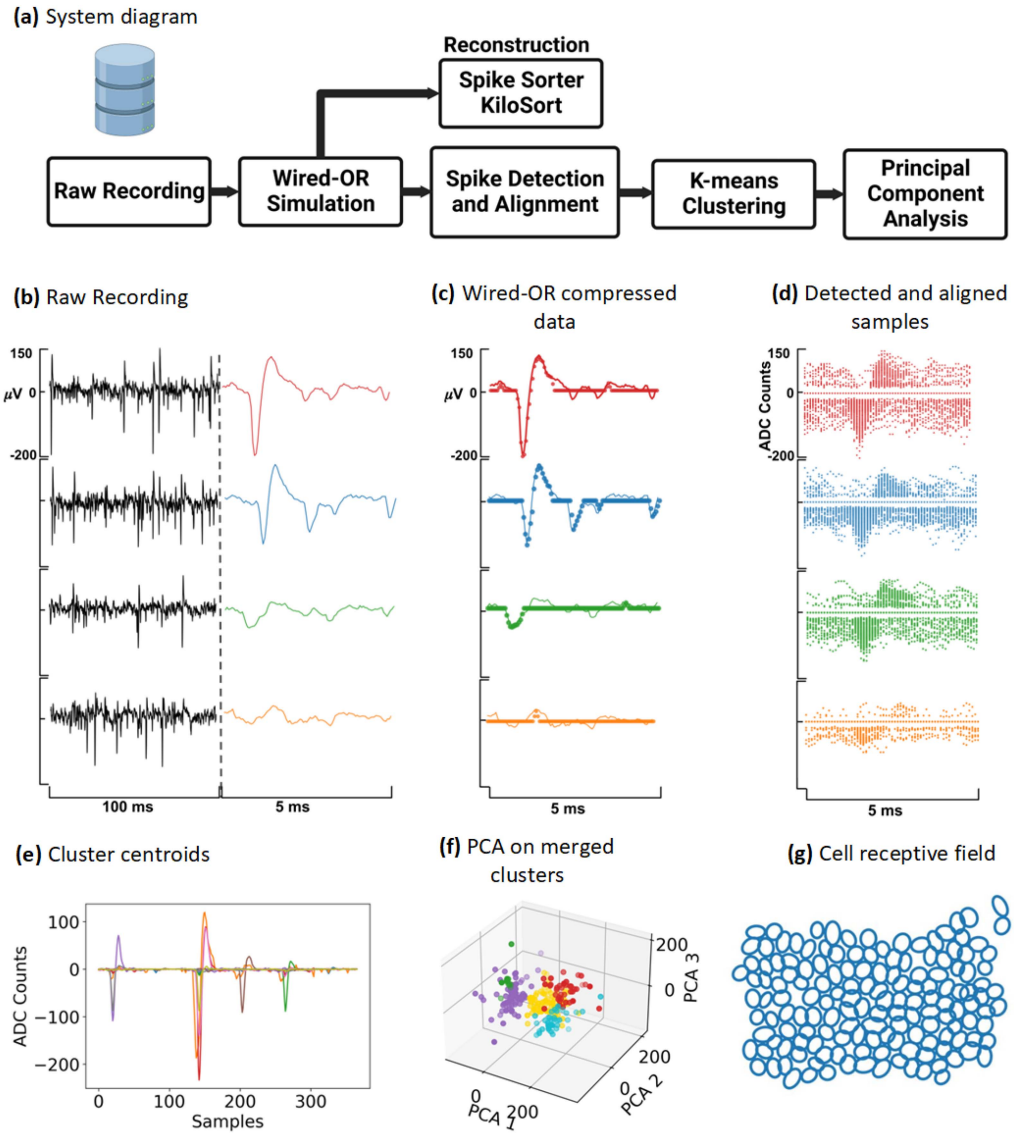


Fig. 5. Neural signal analysis pipeline for our simulation study. (a) Analysis diagram (b) Example of original action potential recording from retina datasets. (c) wired-OR encoded outputs. (d) Detected and aligned spike samples (e) Cluster centroids from multi-electrode K-means clustering. (f) Visualization of merged clusters with principal component analysis. (g) Cell receptive field mapping analyzed with Kilosort after reconstructing wired-OR encoded outputs.

shown in Fig. 5(a). The pipeline consists of spike detection, spike alignment, clustering, and dimensionality reduction analysis using principal component analysis (PCA).

A. *Spike Detection and Alignment*

The most often used spike detection method for implantable neural signal processors is thresholding. Conversely, the wired-OR scheme by construction discards baseline samples and almost always recovers spike samples. Hence, the wired-OR recorded samples can be directly analyzed for spike detection. We adopt the simplest spike detection and alignment method based on the wired-OR readout strategy by assuming all collision-free samples belong to a spike. For any given channel, to preserve the spatiotemporal features for further analysis, we select the surrounding channels of the given channel to form a patch (see Fig. 6). When uniquely decodable samples are recorded, we search for the minimum-value sample in the window of the next 30 samples (this number is empirically chosen) to align the spikes by their peak. This spike detection and alignment approach is hardware friendly and also achieves spike detection in the compressed domain.

B. *Spike Classification–Clustering*

Clustering is commonly used in neural signal processing for spike classification tasks to distinguish spikes from different measured neurons and extract neural spike trains. Clustering is an essential step in spike sorting, which is one of the most important data analysis problems in neurophysiology. The precision in spike sorting critically affects the accuracy of all subsequent analyses [39]. Traditional spike sorting and other

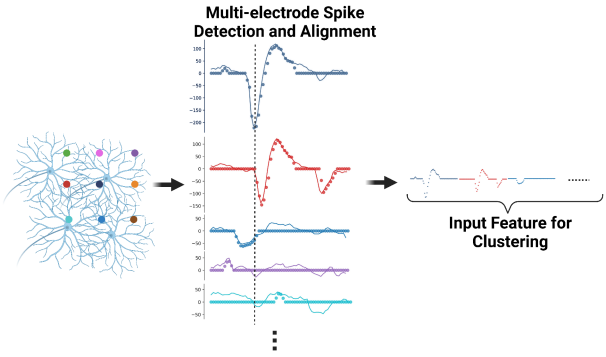


Fig. 6. Multi-electrode detected and aligned signal as input feature for K-means clustering.

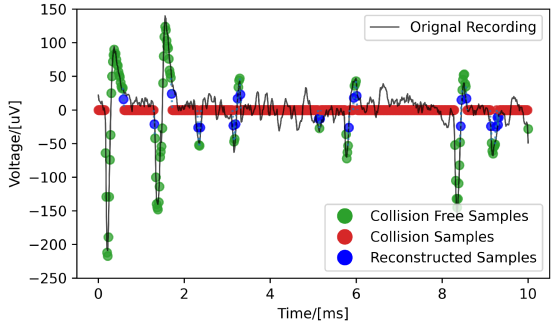


Fig. 7. Wired-OR recorded and reconstructed samples with zero-padding and linear interpolation.

neural signal processing methods require dimensionality reduction to facilitate the clustering process and reduce the computation complexity. This is typically obtained by principal component analysis (PCA). Recent algorithms [40], [41], [42], [43] can perform clustering without the need for feature extraction. In such algorithms, the spike waveform itself can be considered a feature of interest. However, it is not the focus of this work to build a brand new spike sorting algorithm for wired-OR readouts. In this analysis, we implemented a basic clustering algorithm, K-means clustering, to demonstrate the effect of separability among recorded neurons after compression by comparing the pairwise cluster distance in the original recording and the simulated wired-OR outputs. Therefore, the recorded spike samples from the target and surrounding electrodes are detected and aligned by the minimum value, also referred to as the global peak, shown in Fig. 6. We then applied K-means clustering on the reshaped signal array. While multi-electrode clustering preserves the local spatiotemporal features, duplicates of clusters are seen on nearby electrodes. After the full array is analyzed, the clusters are compared to one another to verify that their pairwise cluster distance is greater than a merging threshold T_m. This threshold T_m is approximated from the standard deviation of baseline signal recorded on each electrode when no action potential is detected (denoted as σ), as presented in [41]:

$$T = k(\sigma^2) \tag{6}$$

$$T_M = \sqrt{T} \tag{7}$$

Here, k denotes the number of samples accounted for as input features to the K-means clustering. If two clusters from nearby electrode patches are too close in distance, under the threshold distance, then the clusters are considered to be duplicates and merged.

C. Principal Component Analysis

PCA (Principal Component Analysis) is a statistical method that aims to condense a large set of highly correlated variables into a smaller number of essential variables known as "principal components" while still preserving the significant variations present in the data [44]. We applied PCA to our processing pipeline because it plays an important role in neural signal processing as well as helps us visualize the dimensionality-reduced clusters. An example of the PCA of merged clusters is shown in Fig. 5(f).

D. Cell Classification

Accurate identification of distinct cell types in complex tissue samples is a critical prerequisite for elucidating the roles of cell populations in various biological processes [45]. In the retina particularly, neighboring retina ganglion cells may contain opposite information, some respond to the increase of light intensity (ON response), and others respond to a decrease in light intensity (OFF response). The neural response properties of these different types of cells must be recorded to allow such cells to be differentiated and treated separately to enable an effective prosthesis. To perform cell classification, we use the spike sorting algorithm kiloSort [30]. To interface with this state-of-the-art spike sorting algorithm, we need to reconstruct the original signal. Missing samples due to collisions are initially set to zero and subsequently reconstructed using a 3-tap non-causal finite impulse response (FIR) filter with coefficients $b_{-1} = 0.5$, $b_0 = 0$, $b_{+1} = 0.5$ [29]. This filter operates in the time domain and makes the assumption that the missing sample can be approximated as the average of the previous and the next samples. An example neural signal waveform after reconstruction is shown in Fig. 7. After KiloSort identifies different cell units, cell type classification is performed manually using their measured receptive field properties.

IV. Signal Fidelity Analysis Results

To evaluate the performance of wired-OR architecture, we use 512-channel ex vivo primate retina data recordings and process them in software according to the scheme described in Section II. Different wiring schemes, as described in Section II-B, and ramp resolutions (6-10 bits) define the design space for us to analyze the tradeoff between the signal fidelity of the compressed recording and the overall compression ratio. To quantify the performance of wired-OR in each step of the neural signal processing pipeline shown in Fig. 5(a), different metrics such as spike detection accuracy, average waveform normalized

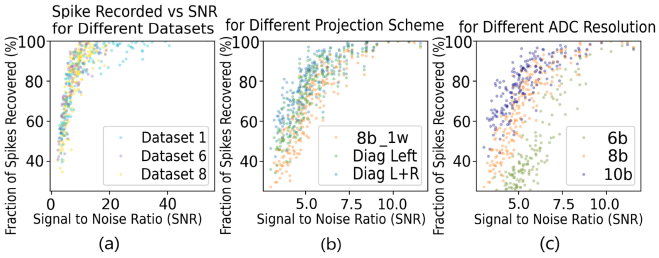


Fig. 8. Spike capturing performance for (a) datasets 1,6 and 8 in 8-bit 1 wire configuration (b) dataset 17 with different numbers of diagonal wires (c) dataset 17 with different ADC solutions.

mean squared error (NMSE), and pairwise cluster distance are analyzed to study their trade-off with compression ratio.

A. *Spike Detection and Alignment*

To quantify the spike detection performance, we analyze the percentage of spikes captured by the output of wired-OR and compare it to the original full-bandwidth offline dataset. We use the spike times detected by KiloSort as ground truth. The percentage of spikes captured for each identified neuron in KiloSort as a function of the event signal-to-noise ratio (SNR) is shown in Fig. 8. The event SNR is approximated by [47]:

$$SNR = \frac{V_{\ spike\ peak\ amplitude}}{V_{\sigma,channel}} \tag{8}$$

The amplitude of the spike peak is determined by identifying the electrode with the most significant negative peak. The noise level is calculated as the median absolute deviation when no action potential is detected on the electrode. The correlation between the percentage of spikes captured and the event signal-to-noise ratio (SNR) is consistent across various datasets (Fig. 8(a)). This correlation allows the performance metrics to be transferable, given the SNR.

A survey of the state-of-the-art neural interfaces shows an SNR range of 7-10 (see Fig. 9). Dataset 17 in our analysis shows similar SNR values (with a mean of 8.32 and a standard deviation of 2.49), which suggests it has a comparable level of SNR to other designs in the survey. The performance of different wiring configurations simulated with dataset 17 is demonstrated in Fig. 8(b), with "soma electrode" referring to the wired-OR system with additional diagonal wiring, as illustrated in Fig. 4(b). "Diag L+R" extends this to two directions of diagonal wiring. As explained in Section II, incorporating diagonal wiring enhances the wired-OR architecture's ability to decode situations where multiple channels are activated simultaneously, resulting in a higher percentage of spikes detected. The percentage of spikes captured for an SNR range of 3-12 given different ramp signal resolutions is shown in Fig. 8(c). Increasing ramp signal resolution is shown to improve performance because the finer the voltage levels, the less chance of multiple channels falling into the same quantization voltage levels. For signal SNR over 7, at least 80% of the spikes are captured for all configurations of wired-OR. For a typical neural recording system design such as the Neuropixels probe in mice, where SNR is around 8 [47],

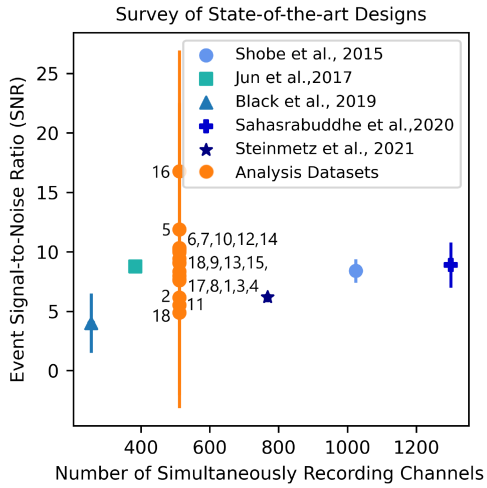


Fig. 9. Recorded signal-to-noise ratio for state-of-the-art neural interface technologies and their number of simultaneous recording channels. The datasets used in this analysis recorded with the system described in [35] has a wide range of signal SNR compared to the design from Shobe et al. [46], NeuroPixel 1.0 [47], BlackRock Utah arrays in [48], [49], NeuroPixel 2.0 [13] and Argo [14].

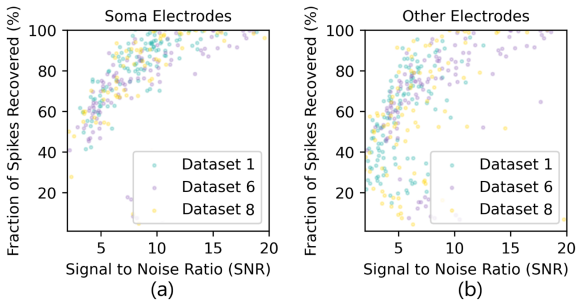


Fig. 10. Spike detection accuracy (a) soma electrodes (b) axon electrodes.

over 90% of the spikes are predicted to be captured by adding diagonal wiring.

Although multiple nearby electrodes see the extracellular signal from one neuron, it only takes one electrode (typically, with the highest SNR) to extract the spike timing information. Given the density of our microelectrode array, the action potential signals seen on this electrode are always generated by the cell body, also known as soma. Here, we refer to this main electrode that records the highest SNR signal of the neuron of study as the "soma electrode." The extracellular voltage signal recorded by other electrodes contains potentially useful information such as axon conduction velocity, axon and dendritic location, etc. Therefore, we further compared the spike-capturing performance of "soma electrodes" and "other electrodes" for each identified neuron in the dataset, showing that soma electrodes capture a higher percentage of spikes (see Fig. 10). For other electrodes along the axon direction, due to the spatial symmetry, they are more likely to record similar amplitude and result in collisions.

As one would expect, the performance improvement from extra wiring and higher ADC resolution comes at the cost of compression ratio. For both diagonal and interleaving wiring,

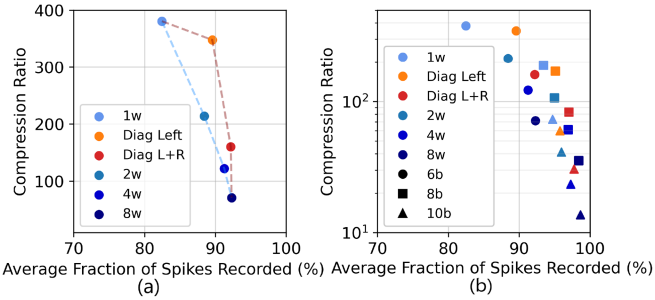


Fig. 11. Pareto frontier of wired-OR average spike capturing performance (8 b 1 wire configuration) for (a) different wiring schemes. (b) All configurations assessed.

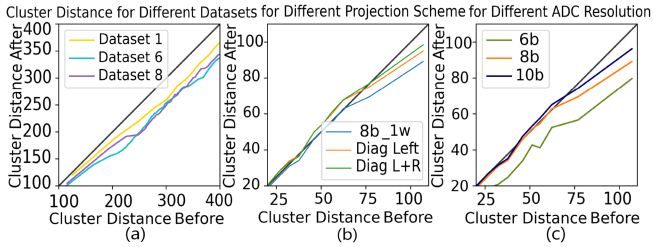


Fig. 12. Average K-means cluster distance before and after wired-OR compression for (a) datasets 1,6 and 8 in 8-bit 1 wire configuration (b) dataset 17 with different numbers of diagonal wires (c) dataset 17 with different ADC solutions.

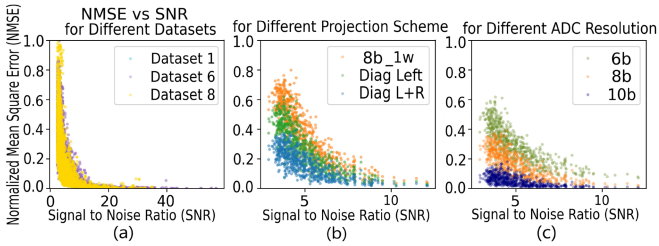


Fig. 13. Waveform recording performance for (a) datasets 1,6 and 8 in 8-bit 1 wire configuration(b) dataset 17 with different numbers of diagonal wires (c) dataset 17 with different ADC solutions.

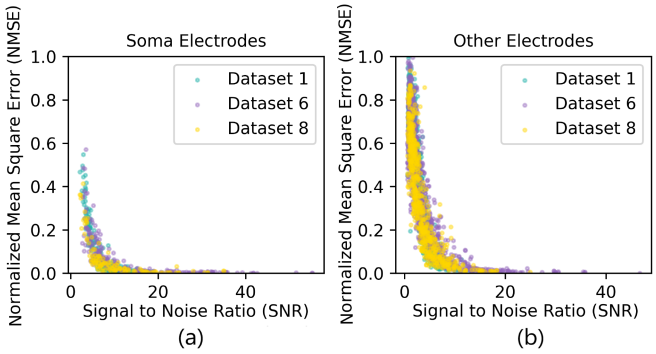


Fig. 14. Waveform recording performance for (a) soma electrodes (b) axon electrodes.

increasing the number of wires boosts the spike detection performance but also decreases the achievable compression ratio (see Fig. 11(a)). Previously proposed interleaving wiring scheme results are demonstrated in blue, and diagonal wiring results are shown in warm colors. Diagonal wiring results are shown to surpass the previous Pareto frontier. This result is also confirmed in different ADC resolution configurations, as shown in Fig. 11(b). Comparing the configuration of diagonal wiring in both directions (4 projections) to 4 interleaved wires, fewer wires and higher compression is possible while further improving the spike capturing performance.

B. Clustering

In order to assess the impact of loss incurred through wired-OR compression to distinguish between cells, two metrics are evaluated: cluster distance and average waveform distortion.

1) Pairwise Cluster Distance: To determine the effect of loss introduced by wired-OR compression on the separability of clusters, we evaluate the pairwise cluster distance between identified units. We match the clusters identified by our neural signal processing pipeline from wired-OR compressed data to that from the original full-bandwidth data. The matching is based on clusters that have the most similar average spike waveform, which is also the centroid of sorted clusters.

The pairwise cluster Euclidean distance for signals seen on each electrode is calculated between each cluster in the multi-electrode patch.

In order to assess the variation in wired-OR performance across different datasets, we compared the average pairwise cluster distance extracted from original data and after wired-OR compression. By analyzing the pairwise cluster distance before and after wired-OR compression in various datasets, a consistent correlation was observed across these datasets (see Fig. 12(a)). Cells with smaller spike event SNR also have smaller cluster distance to other cells. Such clusters are also harder to accurately define after wired-OR compression. The performance of different configurations are demonstrated in Fig. 12(b)–(c). As we expected, diagonal wiring, which resolves more cases where multiple channels are activated, reduces the cluster distance differences comparing wired-OR compressed data to the original full-bandwidth recording. Increasing the quantization resolution of the wired-OR architecture also improves cluster separability.

2) Waveform Estimation: The waveform shape of the action potential spikes are shown to contain valuable information enabling cell-type classification [31] as well as diagnosis and treatment of neurological disorders [50]. We analyze the waveform distortion of the average action potential waveform by studying the normalized mean square error in the spike waveforms in each cell-electrode pair compared to that from the uncompressed dataset. The wired-OR performance shows a strong correlation to the event SNR (Fig. 13). As expected, additional wiring and higher ADC resolution reduce the NMSE (Fig. 13(b)–(c)).

We further examine the waveform estimation performance differentiating soma and axon electrodes. Under a given configuration, the NMSE of the average spike waveform is not affected by whether the signal is recorded on the soma or the axon electrodes and instead depends on the signal SNR (see Fig. 14). Although wired-OR compression is lossy, the spike

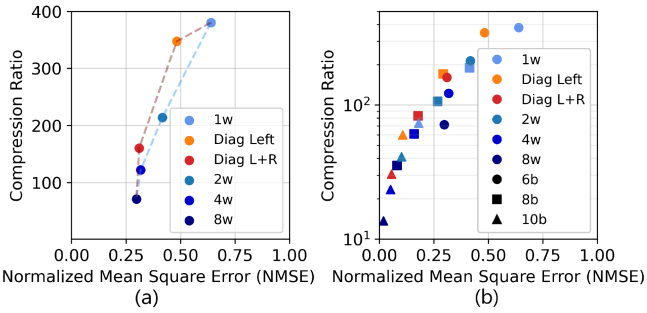


Fig. 15. Pareto frontier of wired-OR waveform recording performance for (a) different wiring schemes. (b) All configurations assessed.

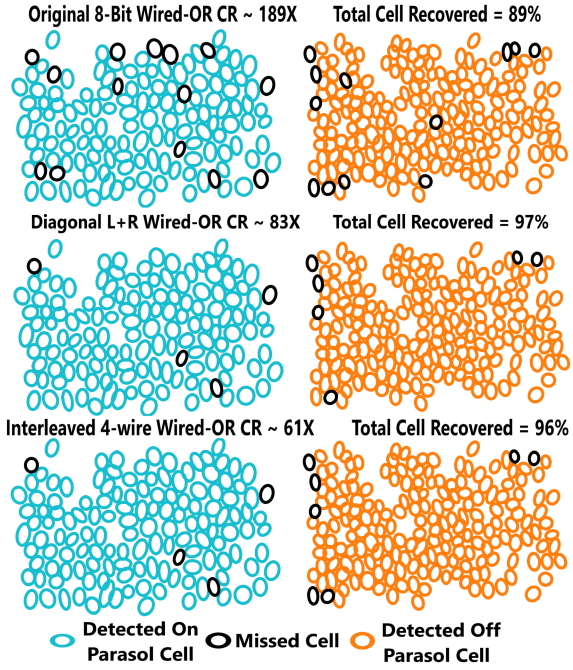


Fig. 16. Receptive field mosaic for ON and OFF parasol cells from several wired-OR configurations.

waveform information is well preserved by averaging tens or hundreds of spikes, which diminishes the loss due to multiple channels activated at the same time, making wired-OR efficient and robust in recording action potential waveforms.

The increase in waveform recording performance resulting from additional wiring or higher ADC resolution is at the penalty of compression ratio, as shown in Fig. 15(a). Similarly, diagonal wiring results exceed the previous Pareto frontier proposed in [29]. The trade-off between the average signal NMSE and compression ratio for all studied configurations is summarized in Fig. 15(b).

C. Cell Classification

To also assess the performance of diagonal wiring to the end result for cell-receptive-field mapping application in retinal prostheses, we also passed the compressed data through state-of-the-art spike sorting and cell-type classification using KiloSort. The recovered receptive field mosaic of the collection of ON and OFF parasol cells is illustrated in Fig. 16. For the same percentage of cells recovered, diagonal wiring achieves higher compression than interleaved wiring.

V. Discussion

A. Performance Across Datasets

The wired-OR readout architecture achieves analog-to-digital compression for neural recordings by exploiting neural signal spatiotemporal sparsity and diversity. Previously, we have analyzed the performance of wired-OR through a "simulation-driven" approach, which requires us to re-run the simulation every time the design parameters or constraints change (e.g., the gain and noise of the recording electronics). Such extensive Nyquist-rate recordings may not be available when we apply wired-OR to different applications. Therefore, it is important to address the dataset variations to build toward a data-driven approach for optimizing wired-OR configurations for different applications. In this article, we extend our analysis to 18 primate datasets collected over the past eight years from different retina samples, front-end settings, and various tissue health conditions.

We characterize the datasets by the total number of cells in each dataset and its average firing rate to study the correlation of performance variation to such characteristics. The number of cells refers to the total number of retinal ganglion cells recorded (2 mm x 1 mm array) and classified manually by domain experts (ground truth) based on their temporal response properties and receptive fields. This metric can translate to cell density as the number of cells per unit size array. The average firing rate describes the average number of spike events per second for all classified neurons.

We then study the performance variation and how much compression wired-OR achieves in different datasets as these biological metrics (event SNR, cell density and average firing rate) change. This comparison is done using the single wire configuration and 10-bit resolution. The average performance for each of the 18 datasets such as spike detection accuracy, waveform NMSE, and compression ratio are summarized in Fig. 17(a)-(c). Examining the performance metrics averaged across each dataset for all 18 datasets, we find the dominating factor for performance degradation in each case. Fig. 17(a) shows that the higher the average firing rate and cell density, the lower the spike detection accuracy after wired-OR compression. However, a higher signal SNR could greatly compensate for the performance degradation due to firing rate and cell density variations. The distortion of the average spike waveform is barely affected by the cell density or firing rate, and depends more on the signal SNR (Fig. 17(g)-(i)), as shown in Fig. 17(b). The number of cells collected in each dataset contributes most to the overall compression ratio, shown in Fig. 17(c). Datasets with higher signal SNR show to achieve lower compression ratio through wired-OR.

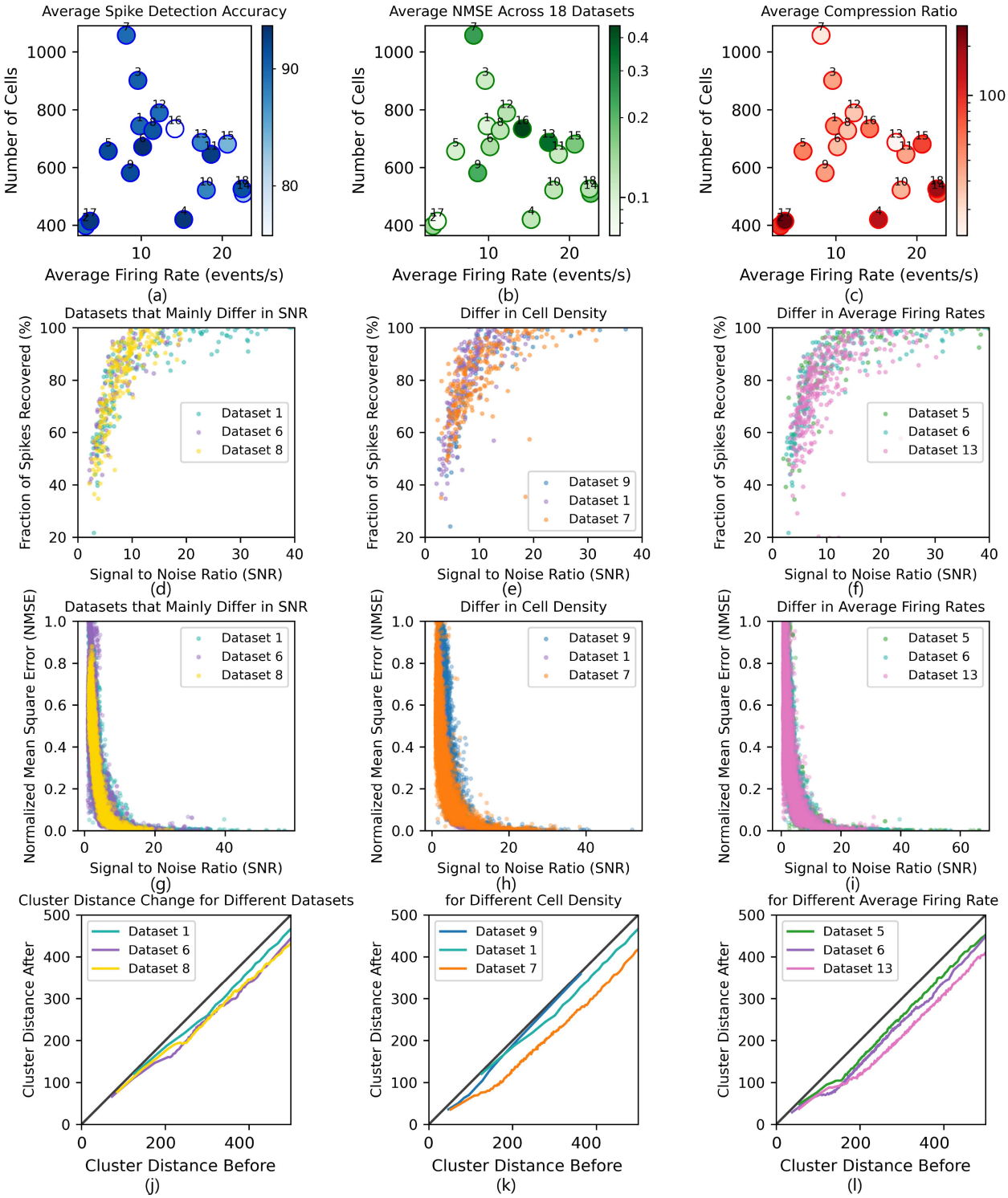


Fig. 17. Wired-OR performance across 18 datasets over 8 years that differs in SNR, average firing rates and number of recorded cells in the array. (a) Average spike detection accuracy. (b) Average Waveform recording performance. (c) Average compression ratio for different datasets. (d) Apike detection accuracy performance variation as the datasets vary in SNR but share similar average firing rate, and number of cells. (e) Spike detection accuracy performance in dataset 9, dataset 1 and dataset8 that are in ascending order of the number of cells recorded in each dataset. (f) Spike detection accuracy performance in datasets that has an increasing overall average firing rate in each dataset. (g) Spike waveform recording performance variation as the datasets vary in SNR but share similar average firing rate, and number of cells. (h) Spike waveform recording performance variation for datasets that vary in the number of cells recorded in each dataset. (i) Spike waveform recording performance variation for datasets that vary in the overall average firing rate in each dataset. (j) Cluster distance variation in multi-electrode K-means clustering analyses as the datasets vary in SNR but share similar average firing rate, and number of cells. (k)Cluster distance variation in datasets that vary in the number of cells recorded in each dataset. (l) Cluster distance variation in datasets that vary in the overall average firing rate in each dataset.

To further decouple the signal SNR variation factor, we studied each dataset more closely to make the following observations:

- Datasets 1, 6 and 8 share similar cell density and average firing rates, and only the event SNR varies. Both the spike-capturing performance and waveform distortion follow a consistent correlation to the event SNR in all three datasets (Fig. 17(d) and Fig. 17(g)). Although errors are introduced by the naive K-means clustering spike classification analysis, the pairwise cluster distance reduction is comparable across all three datasets (Fig. 17(j)).
- Datasets 9, 1 and 7 share similar event SNR and average firing rates and have an ascending cell density (dataset 9 $>$ dataset 1 $>$ dataset 7). Higher cell density results in a wider spread of spike-capturing performance (Fig. 17(e)). This is because higher cell density means each electrode records the extracellular voltage signal from more cell units, causing a higher chance of having multiple rows/columns activated. The more collision cases, the more missing samples from the spikes, which make the clusters less separable, resulting in a reduction of pairwise cluster distance (Fig. 17(k)). Waveform distortion is not impacted by the cell density (Fig. 17(h)).
- Datasets 5, 6 and 13 share similar event SNR and cell density and have an increasing average firing rate (dataset 13 $>$ dataset 6 $>$ dataset 5). Although the three datasets have a different range of event SNR, higher firing rates affects spike-capturing performance as there are more spikes, and a higher chance of more than one row/column getting activated(Fig. 17(f)). Similarly, a higher firing rate also causes degradation in cluster separability (Fig. 17(l)). Waveform distortion is not impacted by the average firing rate (Fig. 17(i)).

B. Post Compression

Traditional full-bandwidth, Nyquist-sampled neural recordings carry a high percentage of random noise with sparse spike activity. From information theory, random noise has very high information content [51]. Therefore, any lossless compression that attempts to do entropy coding offers very little benefit. This is confirmed when we apply Zip, which uses an LZ77 compressor [52], to 5-second original full-bandwidth recording segments, and only achieve $1.53 \sim 1.76\times$ compression ratio.

However, the wired-OR architecture, by design, discards the baseline samples that comprise most of the noise in the recording that cannot be compressed, leaving only spike samples. If we compress the wired-OR encoded bitstream (5-second segments of recordings encoded by an 8-bit, four-interleaving wiring configuration) with Zip, we achieve another $4.3 \sim 4.6\times$ compression of all 5-second segments in one 30-minute retina dataset. Zip performs lossless compression, which is confirmed by comparing the extracted and decoded segments to the wired-OR encoded bitstream and the original dataset, a snippet of which is shown in Fig. 18(b). Interestingly, the more interleaving wiring is used in wired-OR configuration, the more the outputs

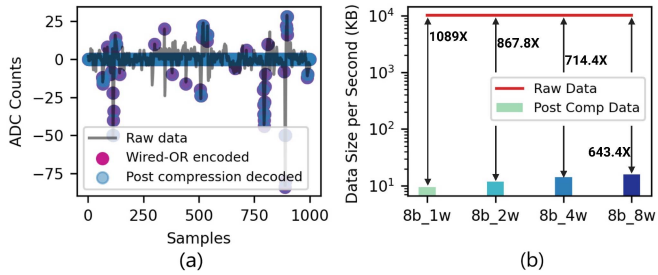


Fig. 18. (a) Example of wired-OR encoded action potential recording and decoded output after zip reconstruction. (b) Lossless compression performance across different wired-OR configurations.

TABLE I
COMPRESSION RATE COMPARISON FOR DIFFERENT ARRAY SIZES AND DIFFERENT WIRED-OR CONFIGURATIONS

Configuration/Array Size	16X32	32X32	64X32
1-Wire 8-Bit	134.7x	217.3x	320.1x
1-Wire 9-Bit	85.3x	137.5x	205.2x
2-Wire 8-Bit	78.3x	130.6x	197.4x
1-Wire 10Bit	73.6x	82.7x	130.8x
4-Wire 8-Bit	41.2x	73.7x	121.0x

can be compressed by Zip (Fig. 18(c)). Overall, lossless compression by Zip of wired-OR output could achieve an overall compression rate of $\sim 1000\times$. These promising results justify exploring future hardware-friendly lossless post-compression of the wired-OR output, as an alternative to the Zip compressor.

C. Scalability

The scalability performance of the wired-OR architecture was evaluated by applying the wired-OR algorithm to arrays of different sizes and assessing the results in terms of spike detection accuracy and average waveform NMSE. We selected four 512-channel datasets with similar cell density and average firing rates and combined to form arrays with sizes 16×32, 32×32, and 64×32. For these artificial datasets, the percentage of spikes captured and the average waveform NMSE consistently correlate with the event SNR across different array sizes (Fig. 19(a), (b) - 8-bit, 1-wire). However, both metrics decline as the array size increases. To mitigate the performance drop, two potential solutions are utilizing a configuration with more wires and increasing the ramp resolution. Increasing the number of wires to 2 and enhancing the ramp resolution to 9 bits for 32×32 or increasing the number of wires to 4 and enhancing the ramp resolution to 10 bits for 64×32 has been shown to lead to an improvement in performance and in some cases, resulting in superior performance outcomes (Fig. 19(b)–(c)), ((e)–(f)). Table I provides the compression rate for various array sizes and different wired-OR configurations. Notably, the maximum compression rate for different array sizes is similar, even though achieved with different configurations. Hence, the wired-OR scheme can be easily scaled to larger arrays by exploiting different wire configurations and bit resolution. The 32 x 32 8-bit configuration wired-OR data-compressive neural recording IC was fabricated and tested in [53]. The IC is tested to accurately

TABLE II
COMPARISON WITH PREVIOUS WORK

Compression Approach	Thresholding Spike-detection [19]–[21], [55]	Compressive Sensing [56], [57]	ML-based Autoencoder [25]	On-chip Spike Sorting [23], [24]	**Wired-OR (This Work)**
Waveform Preservation	x	✓	✓	x	✓
Minimal Computing Overhead	x	x/✓	x	x	✓
Memory Access Free	x	x/✓	x	x	✓
Training Free	x	x/✓	x	x/✓	✓
No. of Channels	64-256	16-32	256	96-384	512-4096
Compression Ratio	10-116×	8-30×	20-500×	240-39272×	50-320×

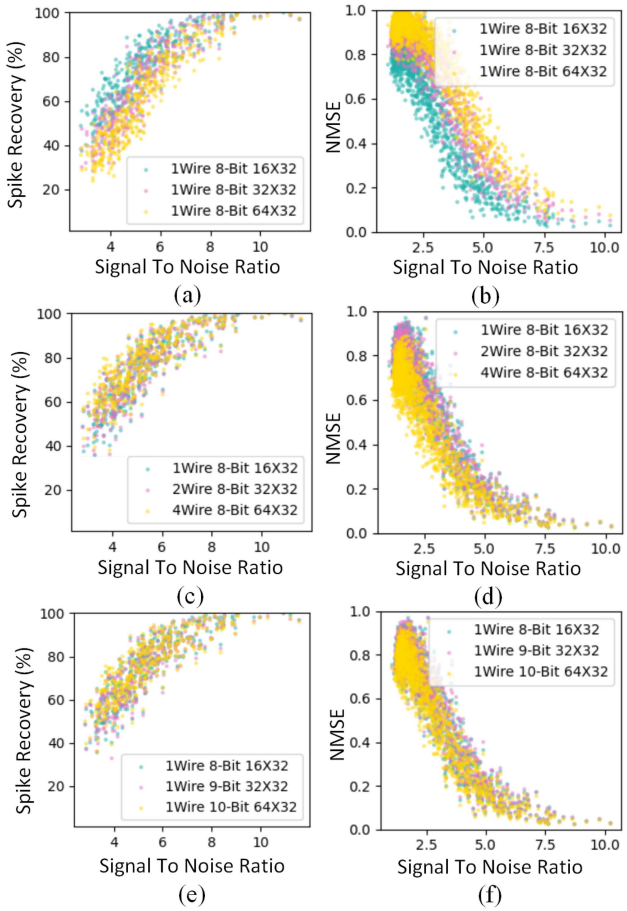


Fig. 19. Evaluation of wired-OR scalability. (a) Spike detection accuracy and (b) spike waveform recording accuracy for different array sizes. Increasing the number of wires to mitigate (c) spike detection accuracy drop and (d) spike waveform recording performance drop. Increasing the ramp resolution to mitigate (e) spike detection accuracy drop and (f) spike waveform recording performance drop.

record neural spikes with 36 μm pixel pitch and consumes only 268 nW per pixel from a single 1 V supply.

D. Comparison With Prior Work

Table II compares this work to previous action potential compression approaches for neural interface designs. This work achieves compression at the analog-to-digital interface, which avoids a large data rate at any point in the system, reducing substantial computational resources and power during digitization, data movement and downstream processing. Compared to other approaches such as on-chip spike detection [18], [19], [20], [54], on-chip spike sorting [22], [23], compressive sensing [55], [56], and machine learning (ML) based autoencoder [24], wired-OR is scalable to higher number of recording channels. And no additional computational resources or memory accesses are needed during compression. Wired-OR also achieves 3x higher compression compared to thresholding-based spike detection, and 10x higher compression compared to compressive sensing. Wired-OR attains similar data-rate reduction compared to ML-based autoencoder while no training is required. Wired-OR attains similar data-rate reduction compared to ML-based autoencoder while no training is required. Compared to on-chip spike sorting, wired-OR preserves the waveform information and compresses action potential signal with minimal computing overhead.

VI. CONCLUSION

We presented a simulation study of the wired-OR compressive readout architecture using a range of multi-electrode neural signal processing methods. We demonstrated that diagonal wiring is a more effective configuration compared to interleaved wiring and showed that the wired-OR readout can effectively capture spikes with high compression. For event SNR of 7-10, which is typical in most neuroscience recordings, the wired-OR readout captures at least 80% of the spikes with at least 50× compression, while maintaining sufficient waveform fidelity for spike sorting. We also showed that wired-OR can be scaled to larger arrays by exploiting different wiring configurations and bit resolution. Additionally, our findings regarding the biological metrics that impact performance variations between datasets will help practitioners estimate the utility of wired-OR across different neural systems. Along with lossless post-compression, Wired-OR can potentially give more than 1000× overall compression.

REFERENCES

[1] Y. Wang, X. Yang, X. Zhang, Y. Wang, and W. Pei, "Implantable intracortical microelectrodes: Reviewing the present with a focus on the future," *Microsyst. Nanoeng.*, vol. 9, 2023, Art. no. 7.
[2] T. Matsuo et al., "Simultaneous recording of single-neuron activities and broad-area intracranial electroencephalography: Electrode design and implantation procedure," *Neurosurgery*, vol. 73, no. 2, pp. ons146–ons54, Dec. 2013.
[3] A. L. Juavinett, G. Bekheet, and A. K. Churchland, "Chronically implanted neuropixels probes enable high-yield recordings in freely moving mice," *ELife*, vol. 8, 2019, Art. no. e47188, doi: 10.7554/elife.47188.
[4] J. Putzeys et al., "Neuropixels data-acquisition system: A scalable platform for parallel recording of 10 000+ electrophysiological signals," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1635–1644, Dec. 2019.

[5] T. Z. Luo, A. G. Bondy, D. Gupta, V. A. Elliott, C. D. Kopec, and C. D. Brody, "An approach for long-term, multi-probe neuropixels recordings in unrestrained rats," *Elife*, vol. 9, 2020, Art. no. e59716.

[6] A. C. Paulk et al., "Large-scale neural recordings with single neuron resolution using neuropixels probes in human cortex," *Nat. Neurosci.*, vol. 25, no. 2, pp. 252–263, Feb. 2022.

[7] E. M. Trautmann et al., "Large-scale brain-wide neural recording in nonhuman primates," *bioRxiv*, 2023, doi: 10.1101/2023.02.01.526664.

[8] F. Willett et al., "A high-performance speech neuroprosthesis," *bioRxiv*, 2023, doi: 10.1101/2023.01.21.524489.

[9] P. D. Ganzer et al., "Restoring the sense of touch using a sensorimotor demultiplexing neural interface: 'Disentangling' sensorimotor events during brain-computer interface control," in *Springer-Briefs in Electrical and Computer Engineering*. Cham: Springer, 2021, pp. 75–85.

[10] D. A. Borton, M. Yin, J. Aceros, and A. Nurmikko, "An implantable wireless neural interface for recording cortical circuit dynamics in moving primates," *J. Neural Eng.*, vol. 10, no. 2, Apr. 2013, Art. no. 026010.

[11] A. M. Sodagar, K. D. Wise, and K. Najafi, "A wireless implantable microsystem for multichannel neural recording," *IEEE Trans. Microw. Theory Tech.*, vol. 57, no. 10, pp. 2565–2573, Oct. 2009.

[12] E. Musk and Neuralink, "An integrated brain-machine interface platform with thousands of channels," *J. Med. Internet Res.*, vol. 21, no. 10, 2019, Art. no. e16194.

[13] N. A. Steinmetz et al., "Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings," *Science*, vol. 372, no. 6539, 2021, Art. no. eabf4588.

[14] K. Sahasrabuddhe et al., "The argo: A high channel count recording system for neural recording in vivo," *J. Neural Eng.*, vol. 18, no. 1, 2021, Art. no. 015002.

[15] X. Yuan, A. Hierlemann, and U. Frey, "Extracellular recording of entire neural networks using a dual-mode microelectrode array with 19 584 electrodes and high SNR," *IEEE J. Solid-State Circuits*, vol. 56, no. 8, pp. 2466–2475, Aug. 2021.

[16] E. M. Trautmann et al., "Accurate estimation of neural population dynamics without spike sorting," *Neuron*, vol. 103, no. 2, pp. 292–308, Jul. 2019.

[17] N. Even-Chen et al., "Power-saving design opportunities for wireless intracortical brain–computer interfaces," *Nature Biomed. Eng.*, vol. 4, no. 10, pp. 984–996, Aug. 2020.

[18] S.-Y. Park, J. Cho, K. Lee, and E. Yoon, "Dynamic power reduction in scalable neural recording interface using spatiotemporal correlation and temporal sparsity of neural signals," *IEEE J. Solid-state Circuits*, vol. 53, no. 4, pp. 1102–1114, Apr. 2018, doi: 10.1109/JSSC.2017.2787749.

[19] X. Guo, M. Shaeri, and M. Shoaran, "An accurate and hardware-efficient dual spike detector for implantable neural interfaces," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, 2022, pp. 70–74, doi: 10.1109/BioCAS54905.2022.9948602.

[20] M. A. Shaeri and A. M. Sodagar, "A Method for Compression of Intra-Cortically-Recorded Neural Signals Dedicated to Implantable Brain– Machine Interfaces," *EEE Trans. Neural Syst. Rehabil. Eng.*, vol. 23, no. 3, pp. 485–497, May 2015, doi: 10.1109/TNSRE.2014.2355139.

[21] D. Valencia and A. Alimohammad, "A real-time spike sorting system using parallel OSort clustering," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1700–1713, Dec. 2019.

[22] Y. Chen et al., "A 384-Channel online-spike-sorting IC using unsupervised Geo-OSort clustering and achieving 0.0013mm2/Ch and 1.78μW/ch," in *Proc. IEEE Int. Solid- State Circuits Conf.*, 2023, pp. 486–488, doi: 10.1109/ISSCC42615.2023.10067264.

[23] J. Li et al., "A 0.78-μW 96-Ch deep sub-vt neural spike processor integrated with a nanowatt power management unit," in *Proc. IEEE 44th Eur. Solid State Circuits Conf.*, 2018, pp. 154–157, doi: 10.1109/ESSCIRC.2018.8494273.

[24] T. Wu, W. Zhao, E. Keefer, and Z. Yang, "Deep compressive autoencoder for action potential compression in large-scale neural recording," 2018, *arXiv:abs/1809.05522*.

[25] M. Pagin and M. Ortmanns, "A neural data lossless compression scheme based on spatial and temporal prediction," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, 2017, pp. 1–4.

[26] A. Kipnis, Y. C. Eldar, and A. J. Goldsmith, "Analog-to-digital compression: A new paradigm for converting signals to bits," *IEEE Signal Process. Mag.*, vol. 35, no. 3, pp. 16–39, May 2018.

[27] B. Gosselin, "Recent advances in neural recording microsystems," *Sensors (Basel)*, vol. 11, no. 5, pp. 4572–4597, Apr. 2011.

[28] J. Wang, Y. Hua, and Z. Zhu, "A 10-bit reconfigurable ADC with SAR/SS mode for neural recording," *Analog Integr. Circuits Signal Process.*, vol. 101, no. 2, pp. 297–305, Nov. 2019.

[29] D. G. Muratore, P. Tandon, M. Wootters, E. J. Chichilnisky, S. Mitra, and B. Murmann, "A data-compressive Wired-OR readout for massively parallel neural recording," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1128–1140, Dec. 2019.

[30] M. Pachitariu, N. Steinmetz, S. Kadir, M. Carandini, and H. D. Kenneth, "Kilosort: Realtime spike-sorting for extracellular electrophysiology with hundreds of channels," Jun. 2016, *BioRxiv:061481*.

[31] L. J. Sukman and E. Stark, "Cortical pyramidal and parvalbumin cells exhibit distinct spatiotemporal extracellular electric potentials," *eNeuro*, vol. 9, no. 6, Nov. 2022.

[32] E. K. Lee et al., "Non-linear dimensionality reduction on extracellular waveforms reveals cell type diversity in premotor cortex," *Elife*, vol. 29, 2021, Art. no. e67490.

[33] D. G. Muratore and E. J. Chichilnisky, "Artificial retina: A future cellular-resolution brain-machine interface," in *NANO-CHIPS 2030: On-Chip AI for an Efficient Data-Driven World*, B. Murmann and B. Hoefflinger, Eds. Cham: Springer, 2020, pp. 443–465.

[34] E. S. Frechette, A. Sher, M. I. Grivich, D. Petrusca, A. M. Litke, and E. J. Chichilnisky, "Fidelity of the ensemble code for visual motion in primate retina," *J. Neurophysiol.*, vol. 94, no. 1, pp. 119–135, Jul. 2005.

[35] A. M. Litke et al., "What does the eye tell the brain?: Development of a system for the large-scale recording of retinal output activity," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 4, pp. 1434–1440, Aug. 2004.

[36] P. Yan, N. P. Shah, D. G. Muratore, P. Tandon, E. J. Chichilnisky, and B. Murmann, "Data compression versus signal fidelity trade-off in Wired-OR ADC arrays for neural recording," in *Proc. IEEE Biomed. Circuits Syst. Conf.*, 2022, pp. 80–84.

[37] P. Giubilato et al., "Low power, high resolution MAPS for particle tracking and imaging," *J. Instrum.*, vol. 10, no. 05, 2015, Art. no. C05004.

[38] P. Giubilato and W. Snoeys, "OrthoPix: A novel compressing architecture for pixel detectors," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. Rec.*, 2012, pp. 1735–1741.

[39] C. R. Caro-Martín, J. M. Delgado-García, A. Gruart, and R. Sánchez-Campusano, "Spike sorting based on shape, phase, and distribution features, and K-TOPS clustering with validity and error indices," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 17796.

[40] M. Pachitariu, N. A. Steinmetz, S. N. Kadir, M. Carandini, and K. D. Harris, "Fast and accurate spike sorting of high-channel count probes with KiloSort," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016.

[41] U. Rutishauser, E. M. Schuman, and A. N. Mamelak, "Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo," *J. Neurosci. Methods*, vol. 154, no. 1/2, pp. 204–224, Jun. 2006.

[42] M. Saif-ur Rehman et al., "SpikeDeep-classifier: A deep-learning based fully automatic offline spike sorting algorithm," *J. Neural Eng.*, vol. 18, no. 1, 2021, Art. no. 016009.

[43] Z. Li, Y. Wang, N. Zhang, and X. Li, "An accurate and robust method for spike sorting based on convolutional neural networks," *Brain Sci*, vol. 10, no. 11, Nov. 2020, Art. no. 835.

[44] J. J. Gerbrands, "On the relationships between SVD, KLT and PCA," *Pattern Recognit.*, vol. 14, no. 1, pp. 375–381, Jan. 1981.

[45] A. Ianevski, A. K. Giri, and T. Aittokallio, "Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data," *Nat. Commun.*, vol. 13, no. 1, 2022, Art. no. 1246.

[46] J. L. Shobe, L. D. Claar, S. Parhami, K. I. Bakhurin, and S. C. Masmanidis, "Brain activity mapping at multiple scales with silicon microprobes containing 1,024 electrodes," *J. Neurophysiol.*, vol. 114, no. 3, pp. 2043–2052, Sep. 2015.

[47] J. J. Jun et al., "Fully integrated silicon probes for high-density recording of neural activity," *Nature*, vol. 551, no. 7679, pp. 232–236, Nov. 2017.

[48] B. J. Black et al., "Chronic recording and electrochemical performance of utah microelectrode arrays implanted in rat motor cortex," *J. Neurophysiol.*, vol. 120, no. 4, pp. 2083–2090, Oct. 2018.

[49] R. Bartolo, R. C. Saunders, A. R. Mitz, and B. B. Averbeck, "Dimensionality, information and learning in prefrontal cortex," *PLoS Comput. Biol.*, vol. 16, no. 4, 2020, Art. no. e1007514.

[50] M. E. J. Obien, K. Deligkaris, T. Bullmann, D. J. Bakkum, and U. Frey, "Revealing neuronal function through microelectrode array recordings," *Front. Neurosci.*, vol. 8, 2014, Art. no. 423.

[51] M. Zbili and S. Rama, "A quick and easy way to estimate entropy and mutual information for neuroscience," *Front. Neuroinform.*, vol. 15, 2021, Art. no. 596443.

[52] R. Waghulde, H. Gurjar, V. Dholakia, and G. P. Bhole, "New data compression algorithm and its comparative study with existing techniques," *Int. J. Comput. Appl. Technol.*, vol. 102, no. 7, pp. 35–38, 2014.
[53] M. Jang et al., "A 1024-Channel 268 nW/pixel 36x36 μm^2/ch data-compressive neural recording IC for high-bandwidth brain-computer interfaces," in *Proc. IEEE Symp. VLSI Technol. Circuits*, 2023.
[54] A. M. Sodagar, K. D. Wise, and K. Najafi, "A fully integrated mixed-signal neural processor for implantable multichannel cortical recording," *IEEE Trans. Bio- Med. Eng.*, vol. 54, no. 6, pp. 1075–1088, Jun. 2007, doi: 10.1109/TBME.2007.894986.
[55] C. Aprile et al., "Adaptive learning-based compressive sampling for low-power wireless implants," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 65, no. 11, pp. 3929–3941, Nov. 2018, doi: 10.1109/TCSI.2018.2853983.
[56] M. Shoaran, M. H. Kamal, C. Pollo, P. Vandergheynst, and A. Schmid, "Compact low-power cortical recording architecture for compressive multichannel data acquisition," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 6, pp. 857–870, Dec. 2014, doi: 0.1109/TBCAS.2014.2304582.

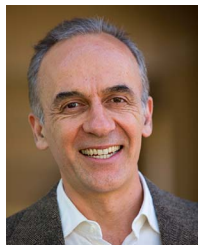
Pumiao Yan (Graduate Student Member, IEEE) received the B.Sc. degree in electrical and computer engineering from Cornell University, Ithaca, NY, USA, in 2018 and the M.S. degree in electrical engineering in 2020 from Stanford University, Stanford, CA, USA, where she is currently working toward the Ph.D. degree. She is Seth A. Ritch Bio-X Graduate Student Fellow with Stanford University. Her research interests include algorithm-hardware co-design and signal processing for analog-to-digital compression hardware architectures for neural interfaces.

Arash Akhoundi (Student Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Tehran, Tehran, Iran, and the M.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, in 2015 and 2018 respectively. He is currently working toward the Ph.D. degree with the Section Bioelectronics of the Microelectronics Department of the Delft University of Technology, Delft, The Netherlands. From 2018 to 2022, he was working as a Digital Signal Processing and FPGA Engineer. His research interests include massive parallel neural interfaces data compression and signal processing.

Nishal P. Shah (Member, IEEE) received the B.Tech and M.Tech degrees in electrical engineering from the Indian Institute of Technology Delhi, New Delhi, India, in 2013 and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2020. He is currently the Milton Safenowitz Postdoctoral Scholar with the Neural Prosthetics Translational Lab, Stanford University and a Member of the BrainGate2 consortium. His research interests include the intersection of neuroscience and neuro-engineering, working on computational methods for reading-out intended movements from the brain as well as writing-in sensory information to the brain.

Pulkit Tandon received the B.tech degree in electrical engineering from the Indian Institute of Technology Bombay, Mumbai, India, in 2016, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2018 and 2022, respectively. He was a Rambus Stanford Graduate Fellow during his PhD. His research interests include the intersection of compression, neuro-engineering, perceptual engineering and brain-machine-interfaces, with components of algorithms, hardware and human perception.

Dante G. Muratore (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the Politecnico of Turin, Turin, Italy, in 2012 and 2013, respectively, and the Ph.D. degree in microelectronics from the Integrated Microsystems Lab, The University of Pavia, Pavia, Italy, in 2017 . He is currently an Assistant Professor with the Microelectronics Department, Delft University of Technology, Delft, The Netherlands. From 2015 to 2016, he was a Visiting Scholar with Microsystems Technology labs, Massachusetts Institute of Technology, Cambridge, MA, USA. From 2016 to 2020, he was a Postdoctoral Fellow with Stanford University. His research focuses on hardware and system solutions for high-bandwidth brain-machine interfaces that can interact with the nervous system at natural resolution. He was the recipient of the Wu Tsai Neurosciences Institute Interdisciplinary Scholar Award. He is an Associate Editor at IEEE TCAS-II.



E. J. Chichilnisky received the B.A. degree in mathematics from Princeton University, Princeton, NJ, USA, and the M.S. degree in mathematics and the Ph.D. degree in neuroscience from Stanford University, Stanford, CA, USA. He is currently the John R. Adler Professor of Neurosurgery, and Professor of Ophthalmology, with Stanford University, where he has been working since 2013. Previously, he was with the Salk Institute for Biological Studies for 15 years. His research interests include understanding the spatiotemporal patterns of electrical activity in the retina that convey visual information to the brain, and their origins in retinal circuitry, using large-scale multi-electrode recordings. His ongoing work now focuses on using basic science knowledge along with electrical stimulation to develop a novel high-fidelity artificial retina for treating incurable blindness. He was the recipient of an Alfred P. Sloan Research Fellowship, McKnight Scholar Award, McKnight Technological Innovation in Neuroscience Award, and Research to Prevent Blindness Stein Innovation Award.

Boris Murmann (Fellow, IEEE) received the Dipl.-Ing. (FH) degree in communications engineering from Fachhochschule Dieburg, Dieburg, Germany, in 1994, the M.S. degree in electrical engineering from Santa Clara University, Santa Clara, CA, USA, in 1999, and the Ph.D. degree in electrical engineering from the University of California at Berkeley, Berkeley, CA, USA, in 2003. From 1994 to 1997, he was with Neutron Mikrolektronik GmbH, Hanau, Germany, where he was involved in the development of low-power and smart-power application-specified integrated circuits (ASICs) in automotive CMOS technology. Since 2004, he has been with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA, where he is currently a Full Professor. His research interests include the area of mixed-signal integrated circuit design, with special emphasis on data converters, sensor interfaces, and circuits for embedded machine learning. Dr. Murmann was the Data Converter Subcommittee Chair and 2017 Program Chair of IEEE International Solid-State Circuits Conference. He was a co-recipient of the Best Student Paper Award at the Very Large-Scale Integration Circuits Symposium in 2008 and 2021, Best Invited Paper Award at the IEEE Custom Integrated Circuits Conference (CICC) in 2008, Agilent Early Career Professor Award in 2009, Friedrich Wilhelm Bessel Research Award in 2012 and SIA-SRC University Researcher Award for lifetime research contributions to the U.S. semiconductor industry in 2021.