Recording Front-End Electronics for Large-Scale Implantable Brain-Computer Interfaces: A Design Perspective

Xiaohua Huang, Dante G. Muratore

Delft University of Technology, The Netherlands

Abstract – This paper discusses recent advancements in recording front-end electronics for large-scale implantable brain-computer interfaces. Various system architectures and circuit techniques can be leveraged to achieve both area- and power-efficient implementations. Here, we elaborate on the trade-offs between different approaches, with specific examples highlighting details of recently proposed solutions. We also provide several practical design tips and tricks for implementing such front-end electronics in standard CMOS technologies. Finally, we discuss the most interesting future directions for the field.

Introduction

Implantable brain-computer interfaces (BCIs) establish direct communication pathways between the human brain and external devices, holding promise to revolutionize therapies for neurological diseases because they interface with the nervous system with higher spatio-temporal resolution than traditional pharmacological, surgical, or gene-based approaches [1], [2]. These technologies involve the surgical implantation of electrodes either on the brain surface (electrocorticography or ECoG) or inside the brain (penetrating electrodes) [3], [4]. Future implantable BCIs for neuroscience research and clinical applications will greatly benefit from compact, low-power neural recording ICs with high channel counts.

This paper presents a general overview of large-scale implantable BCIs, focusing on addressing key challenges in the design of analog front-end (AFE) electronics across the entire signal chain (Fig. 1). Various system architectures and circuit techniques are discussed, with solutions to address critical application challenges, such as sensor-circuit connectivity, high-impedance input bias network, low noise electrode DC offset (EDO) compensation, area-power-efficient analog-to-digital converters (ADCs), and data compression.

Neural signals and electrode-tissue interface

Typically, two types of brain signals can be captured with implantable electrodes: low-frequency (~[1:300] Hz) local field potentials (LFPs) or electrocorticography (ECoG), which represent the aggregate activity of a group of neurons in a volume of tissue, and high-frequency (~[300:10k] Hz) spikes or action potentials (APs), which reflect single-neuron activity. These signals generally have amplitudes varying from a few μV to several hundreds of μV [5].

When electrodes are in contact with brain tissue, a DC potential known as half-cell potential (V_{HC}), develops at the electrode-tissue interface to facilitate the transduction from ionic to electronic current [6]. The difference in DC potentials between the recording and reference electrodes is commonly known as electrode DC offset (EDO). EDOs can reach values as high as tens of mV, depending on the materials and dimensions of the recording and reference electrodes, and can saturate the low dynamic-range AFE [7].

Very small electrodes (with diameters in the tens of μm or less) are highly desirable in large-scale BCIs to increase spatial resolution. This leads to very high electrode-tissue interface impedances that contribute significant thermal noise. These small electrodes typically have impedances ranging from several hundreds of kΩ to a few MΩ at 1kHz [8], with much higher values at lower frequencies (e.g., hundreds of MΩ at 1Hz).

Key challenges of neural front-end electronics

Developing large-scale implantable BCIs that integrate high-density electrode arrays with CMOS readout ICs (ROICs) for direct interfacing with the brain presents numerous challenges, such as biocompatibility, heat generation, device stability and longevity, low signal amplitudes, large EDOs, high electrode impedance, and efficient wireless data transmission. Therefore, the devices and circuits designed for these interfaces must meet a series of requirements. This section examines these critical requirements, with a special focus on the design of front-end electronics.

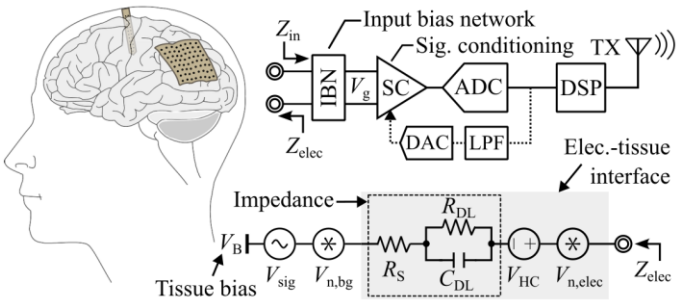


Fig. 1. Block diagram of recording front-end electronics for implantable BCIs, using surface or penetrating electrodes. Dashed line in feedback path indicates mixed-signal DSL in DC-coupled AFE. The bottom inset shows a simplified electrical model of electrode-tissue interface.

Sensor-circuit connectivity: Interfacing with the brain with high channel count is essential for achieving both high spatio-temporal resolution and/or large brain coverage. Research suggests that simultaneous recording of 5k-10k neurons is needed to restore limb movement, while 100k neurons are needed to facilitate full-body movements [9]. In retina implants, simultaneous recording from tens of thousands of electrodes would be necessary to enable closed-loop clinical devices [10]. However, conventional passive electrode arrays paired with CMOS ROICs pose significant challenges for high-channel-count systems, as each electrode must be addressed individually, resulting in a wiring bottleneck at the sensor-circuit interface.

High input impedance: To minimize signal attenuation, the input impedance (Z_{in}) of the front-end recording electronics must be designed to be significantly higher (typically >10x) than the electrode impedance across all frequencies of interest. The actual required input impedance is highly dependent on the electrode size and material, but in general it needs to be designed at least in the tens of MΩ or higher range [11].

Input bias network: When interfacing with very-high-impedance electrodes and recording extremely low-frequency signals (such as those below 1Hz), the design and implementation of an effective input bias network (IBN) presents significant challenges. The primary goal of this bias network is to establish a stable input common-mode voltage for the first-stage signal conditioning circuitry, typically a low-noise amplifier. This common-mode voltage is crucial for ensuring that the front-end electronics operate within their optimal bias range. In addition, the network must maintain a much higher input impedance (Z_{in}) than the electrode impedance (Z_{elec}) to minimize signal attenuation and distortion. This challenge gets further exacerbated by the need to maintain the high input impedance across the entire frequency range of interest, including ultra-low frequencies. Furthermore, the bias network must be designed to introduce negligible noise and offset, as even small amounts can obscure the μV-level neural signals and cause baseline drift, leading to dynamic range reduction, or potentially complete saturation of the front-end electronics.

Small form factor: Minimizing the implant footprint is critical for minimizing tissue damage, improving biocompatibility, and mitigating foreign body response and inflammation, which are key factors for chronic implantation [12]. Ultimately, the goal is to achieve the smallest possible overall implant size, while maintaining its core functionality and performance.

Low power: Maintaining low power consumption is crucial to prevent excessive heat generation, which can lead to tissue damage. While there are currently no established standards in clinical practice, it is generally recommended that the increase in body temperature remains below 1˚C [13], which, for example, corresponds to a power density of ~1mW/mm^2 in retinal implants [14]. With typical reported area per channel of ~0.01mm^2, the maximum allowable power consumption per channel is limited to ~10μW.

2025 IEEE Custom Integrated Circuits Conference (CICC) | 979-8-3315-1745-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/CICC63670.2025.10983696

979-8-3315-1745-8/25/$31.00 ©2025 IEEE

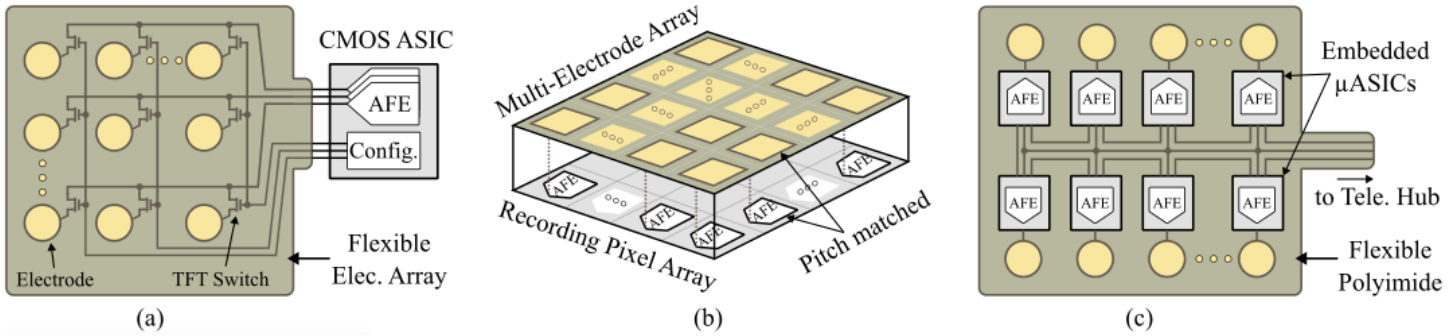


Fig. 2. Scalable architectures for large-scale BCIs overcoming the wiring bottleneck: (a) hybrid integration, (b) monolithic integration, and (c) NeuroBus.

Large EDO tolerance: A major challenge in developing large-scale BCIs is the accurate acquisition of weak neural signals in the presence of large EDOs, with minimal area, power, and noise. This is because EDOs tend to drift over time, and can be up to 2 to 3 orders of magnitude larger than the signals of interest. In addition, the very-low-frequency nature of neural signals (near-DC to several kHz) makes it particularly challenging to filter out large unwanted EDOs efficiently [15].

Low noise: The front-end electronics must be designed to achieve low noise (typically <10µV_{rms}), lower than the background noise $V_{n,bg}$, such that their contribution would not degrade the signal-to-noise ratio (SNR) of the recorded neural signals. In general, two main noise sources are present in front-end electronics: the intrinsic noise generated in the MOS transistors (including thermal and flicker noise) [16], as well as the extrinsic noise from quantization [17]. For neural recordings with moderate resolution requirements, the quantization noise can be designed negligible with relatively low area and power overhead, making it less of a concern in most implementations. Hence, the intrinsic flicker and thermal noise represent the major limitations in the noise performance of most front-end electronics. Finally, note that electrode noise ($V_{n,elec}$) is typically negligible within the signal band, but may become significant for very small, high-impedance electrodes [18].

Data compression: Over the past few decades, the number of simultaneously recorded neurons has followed a trajectory similar to Moore's law, though at a slower pace, doubling approximately every 6 years [19]. Currently, state-of-the-art high-density neural recording systems can record data from up to several thousands of neurons. However, as the number of channels continues to increase, the resulting raw data throughput can quickly become unmanageable (e.g., 10,000 channels digitized at 20kS/s and 10-bit resolution generate 2Gb/s), and wireless transmission of these data would consume a prohibitive amount of power. Thus, compressing the raw data on-chip before the wireless transmission is highly desirable to ensure system scalability and performance.

Architectures and circuit techniques for large-scale BCIs

Various system architectures and circuit techniques have been proposed in literature to address the challenges mentioned above. Their characteristics, advantages, and disadvantages will be discussed in this section.

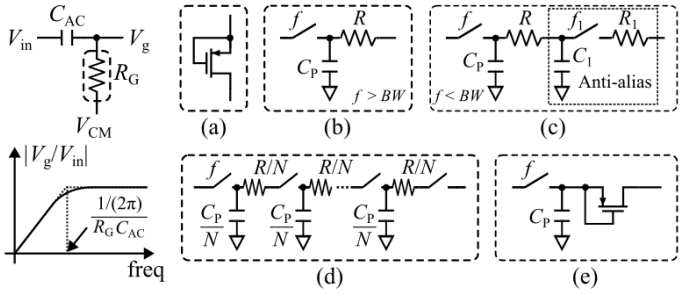
A. Scalable architectures: overcoming the wiring bottleneck

A critical challenge in developing large-scale BCIs lies in the wiring bottleneck at the sensor-circuit interface. Three key strategies have emerged to tackle this issue (see Fig. 2).

Active multiplexing in hybrid integration: Embedding active components, such as transistors, directly into the electrode arrays, introduces the concept of active arrays (Fig. 2(a)). These arrays allow for the multiplexing of multiple electrodes onto few readout data lines [20], [21], [22]. For an active array structured in a *N* x *M* matrix, only *N* digital multiplexing lines and *M* analog readout lines are required to interface with the recording electronics, compared to *N* x *M* connections required in traditional configurations of passive arrays paired with CMOS ROICs [23]. For instance, [20] implements 16:1 time-division multiplexing (TDM), enabling simultaneous recording from 256 electrodes while using only 16 multiplexed channels (termed as super-channels) and 16 digital multiplexing lines. Recent advancements have demonstrated the capability to record from up to 4096 (64 x 64) electrodes using only ~130 connections between the active array and CMOS ROIC [22]. Typically, these active arrays can be fabricated on thin and conformal polymeric substrates using flexible electronics, allowing them to closely adhere to the curved surface of the cerebral cortex [21]. This enhances the stability of electrical and mechanical contacts while minimizing tissue damage.

TDM can address or mitigate the wiring bottleneck, but it suffers from noise folding due to the required higher acquisition bandwidth. The recording circuits can be designed with lower noise to maintain an acceptable noise level after folding (note: power efficiency remains uncompromised when amortized across multiplexed channels). However, the noise folding from the electrodes and switches in the active array can only be reduced by lowering the electrode impedance and switch ON-resistance. For this reason, the multiplexing ratio is kept low [24] when the electrode impedance and/or the recording bandwidth are high, as in the case of small electrodes for single-unit recordings. For the relatively large surface electrodes used in (µ)ECoG, direct TDM at the electrodes can be applied without incurring significant noise degradation due to the lower electrode impedance and narrower signal bandwidth (500Hz). Notably, in active electrode arrays, the noise folding from the switch ON-resistance can dominate, due to the lower electron mobility of transistors fabricated in flexible electronics [25]. Hence, the multiplexing switches need to be sized reasonably large to achieve sufficiently low ON-resistance [20].

Monolithic integration: Employed in both penetrating probes and surface electrode arrays [26], [27], [28], [29], this approach seamlessly integrates sensing electrodes onto the same CMOS substrate as the recording channels (known as active pixels), effectively enhancing the scalability of one-to-one connections between the electrode and recording circuit (Fig. 2(b)). A key advantage of monolithic integration is the ability to maintain high signal integrity due to the close proximity of the sensing electrodes and recording circuitry, leading to less parasitic effects and reduced sensitivity to crosstalk and electromagnetic interference (EMI) [28]. Furthermore, this approach allows for higher-level channel multiplexing within the circuit hierarchy (e.g., several dedicated front-end amplifiers sharing a single ADC stage), thereby enabling very high integration densities and channel counts. For instance, implantable neural probes have emerged as the most widely used tools to monitor electrical activity at single-cell resolution, and they are established tools in fundamental neuroscience [30], [31]. Recent efforts have led to a highly integrated design that features 1536 channels and 5120 TiN electrodes distributed across four 10mm penetrating shanks [32]. However, the pixel size is heavily constrained by the electrode pitch, which usually leads to worse power efficiency. In addition, due to the large mechanical mismatch between silicon-based probes/arrays and brain tissue, monolithic integration exhibits a significant limitation in terms of flexibility when distributing many recording channels across different regions of interest. As a result, this approach cannot record activity from large brain regions, which is important to map connectivity or for functional brain analysis [33].



	(a)	(b)	(c)	(d)	(e)
$R_{G,max}$	Large	Small	Medium	Large	Medium
Variation	Large	Small	Small	Small	Medium
Area	Small	Large	Large	Large	Small

Fig. 3. Implementations of large resistance for high-pass filtering in AC-coupled neural AFEs: (a) pseudo-resistor, (b) duty-cycled resistor, (c) multi-rate duty-cycled resistor, (d) segmented duty-cycled resistor, and (e) duty-cycled pseudo-resistor.

NeuroBus: Finally, it is important to highlight a recently proposed solution called NeuroBus (Fig. 2(c)), which utilizes tiny CMOS readout ASICs distributed across a thin, ultra-flexible polyimide (PI) substrate to strike a balance between hybrid and monolithic integration [34]. These ASICs are placed in close proximity to the sensing electrodes and arranged in a bus-like configuration, which simplifies the wiring and allows for short connections between the electrodes and the CMOS ASICs. Furthermore, this approach supports the customizable placement of each ASIC during CMOS post-processing, enabling the versatile manufacturing of brain implants that can be tailored to specific applications.

B. Input bias network

AC-coupling is the simplest and yet highly effective method to deal with EDOs, since it can provide rail-to-rail EDO rejection without requiring complex cancellation loops. This approach employs passive RC-based high-pass filtering to block unwanted DC components at the electrode-tissue interface while allowing the desired signals to pass through. A key advantage of this approach is that it allows the front-end's input common-mode level to be set independently from the electrode's DC potential.

However, in practical on-chip implementations that utilize pF-level coupling capacitors, the required resistor values typically reach TΩ range for ultra-low-frequency LFP recordings (<1Hz). For this reason, transistors operating in cutoff region, known as pseudo-resistors (Fig. 3(a)), have been proposed [15]. These elements can provide the required ultra-high resistances while being compact and power-efficient, making them particularly well-suited for recording low-frequency signals. Nonetheless, at these high-resistance values, pseudo-resistors are extremely sensitive to process variation and light, nonlinear, and very prone to offsets caused by leakage currents [35]. Duty-cycled resistors (Fig. 3(b) and (c)) with single- and multi-rate control have been introduced to mitigate some of these issues [36], [37]. However, their effectiveness is limited by parasitic capacitance, which limits the maximum achievable resistance and large AC-coupling capacitors are needed to achieve a high-pass corner frequency below 1Hz. The segmented duty-cycled resistor, which divides a resistor into multiple segments, mitigates the issue of parasitic capacitance (Fig. 3(d)). This approach effectively achieves TΩ resistance with a switching frequency above the signal bandwidth and improves robustness against process variations [38]. Nevertheless, (segmented) duty-cycled resistors still need to be sized reasonable large (usually hundreds of kΩ or more), resulting in reduced area efficiency, despite offering improved robustness. Hence, a hybrid approach using a duty-cycled pseudo-resistor (Fig. 3(e)) can be area efficient and more robust to leakage currents compared to standalone pseudo-resistors, since the required resistance is much lower in the duty-cycled structure [39]. Finally, all the aforementioned methods suffer from long settling times in the presence of large artefacts (such as those due to motion or stimulation). To address this, a simple solution is to use a switch that remains OFF during normal recording for high resistance, and is switched ON to enable rapid recovery from artifacts or leakage induced saturation [35], [40]. This method is highly power-area-efficient, making it particularly attractive for applications where precise control of the high-pass corner frequency is not critical. The switch's OFF resistance can be designed to be sufficiently large to maintain the required high-pass corner under worst-case conditions, such as the fast-fast (or FF) process corner.

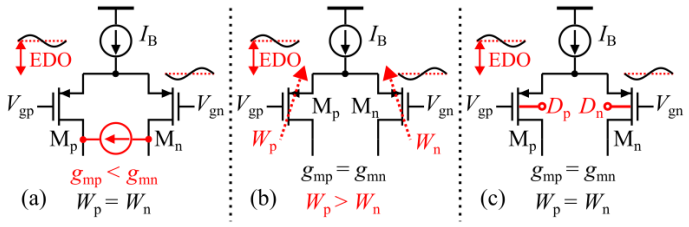


Fig. 4. EDO compensation DACs in DC-coupled neural AFEs: (a) IDAC, (b) Merged-amplifier DAC, and (c) Bulk-DAC.

Alternatively, DC coupling can be employed with DC-servo loops (DSLs) to compensate for the large EDOs. A DSL consists of a low-bandwidth low-pass filter in the feedback path, implemented in either analog or mixed-signal domain [41], [42]. The high loop gain at low frequencies effectively attenuates low-frequency signals at the output, thus realizing the required high-pass filtering. Note that large analog time constants are still needed in analog DSLs, which makes them area inefficient [41], unless the DSL is shared among nearby channels [43]. In contrast, DC-coupling with mixed-signal DSLs offers the advantage of improved area efficiency by eliminating the requirement for large analog time constants. However, it also comes with considerable challenges of input bias, particularly when interfacing with small, high-impedance electrodes. Even very low leakage currents at these high-impedance nodes can shift the bias voltage outside the front-end's input common-mode range. Therefore, a carefully designed input bias network, tailored to the electrode characteristics, is essential to maintain a proper input bias of the recording front-end. An example of such co-design strategy is presented in [44].

C. Low-noise EDO compensation in DC-coupled AFE

While DC-coupling with mixed-signal DSLs offers potentially improved area efficiency by eliminating the requirement for large analog time constant, it presents significant challenges in avoiding noise degradation in the presence of large EDOs. Typically, the tight constraint on leakage current flowing toward the electrode prevents the use of direct EDO compensation at the input of the recording front-end through Gm-based servo loops [44]. However, compensating the EDO after the input can introduce significant noise degradation. Figure 4(a) illustrates how compensating for large EDOs using a conventional current DAC (IDAC) at the drains of the differential pair (i.e., after the input) substantially increases the noise, mainly due to unbalanced transconductances, IDAC noise, and asymmetric propagation of tail-current noise [28]. Note that noise from bias circuitry, while not shown in the figure, also contributes to the overall degradation.

To mitigate noise degradation in conventional IDAC-based EDO cancellation, a merged amplifier-DAC structure has been introduced, where the EDO is cancelled by varying the transistor width of the input differential pair biased in weak inversion (Fig. 4(b)) [44]. Essentially, by changing the width ratio of the input pair, an offset equal in magnitude but opposite in polarity to the EDO is generated, effectively performing EDO cancellation. This approach ensures that the Gm symmetry of the input differential pair is well maintained after EDO compensation, thereby, does not incur penalty in thermal noise degradation (to first order approximation). However, flicker noise still gets modulated due to the varying device width and inversion level of the input transistors. This leads to a substantial increase in the flicker noise corner frequency when compensating for large EDOs. For instance, for a maximum EDO compensation of 50 mV (i.e., 100 mV_{pp}), the flicker noise corner frequency increases by a factor of 2.5x in [44].

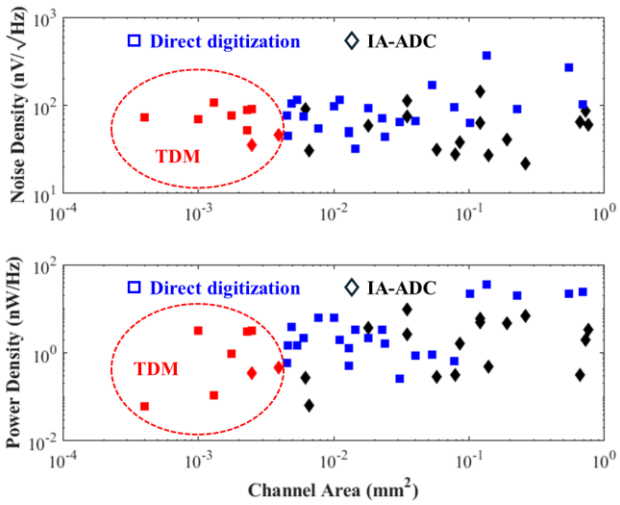


Fig. 5. Performance comparison of state-of-the-art neural AFEs: noise density (top) and power density (bottom) plotted versus channel area, with red markers indicate designs employing TDM.

To overcome the limitations of the merged amplifier-DAC structure and prevent noise degradation in the presence of large EDOs, a novel EDO compensation technique based on a bulk-DAC (BDAC) has been implemented (Fig. 4(c)) [45]. This approach leverages the bulk transconductance of the input transistors, where feedback digital signals are applied directly to the bulk terminals of the input transistors. While bulk modulation in a regular CMOS technology is rather limited and prone to latch-up, a fully depleted silicon on insulator (FDSOI) technology allows for a wide range of body biasing (e.g., -2 to 2 V in 22-nm FDSOI) to modulate the transistor threshold voltage [46]. Since a well-balanced input-pair transconductance and constant device width are maintained after EDO compensation, both thermal and flicker noise remain uncompromised across the entire EDO compensation range ($120mV_{pp}$) [45].

D. Compact and energy-efficient neural AFEs

The conventional IA-ADC architecture typically consists of an AC-coupled instrumentation amplifier (IA) followed by a successive-approximation-register (SAR) ADC. This approach heavily relies on analog-intensive techniques to implement the front-end IA and, in some cases, the bandpass filter to separate LFP and AP bands, which makes their scalability with technology difficult [47].

To improve area and energy efficiency, a paradigm shift towards direct digitization has emerged. This approach utilizes moderate-resolution ADCs (~8-11 bits) to directly digitize raw neural signals (i.e., without front-end high-gain amplification). The large EDOs can be either compensated through mixed-signal DSLs [48] or filtered by conventional AC-coupling [35], [49]. For direct digitization, two important ADC architectures have gained increasing attention: incremental ADCs and single-slope (SS) ADCs.

Incremental ADCs operate on the principle of periodically resetting a sigma-delta modulator, effectively resulting in a Nyquist-rate ADC with the benefit of inherent anti-aliasing, oversampling and low latency [50]. When implemented with a 1-bit internal quantizer, a corresponding 1-bit DAC is required in the feedback and the decimation of the bitstream can be achieved using a simple ripple counter, allowing for highly efficient on-chip implementations [27], [28], [35]. On the other hand, single-slope ADCs represent another promising candidate. The simplicity of this architecture, consisting of a boxcar sampler, comparator, digital counter, and shared ramp generator, contributes to its efficiency in both power consumption and area utilization [29]. Furthermore, both incremental and SS ADCs are inherently compatible with channel multiplexing, making them particularly well-suited for large-scale neural recordings with multiple input channels. This feature enables significant area savings via hardware reuse, as demonstrated in [20], [29], [45].

Figure 5 compares the state-of-the-art neural AFEs in terms of area, power, and noise performance. In general, time multiplexing offers a significant reduction in channel area, while maintaining competitive noise and power efficiency. Notably, when combined with other circuit techniques, the smallest per-channel area, lowest power density, and comparable noise performance are achieved in [45].

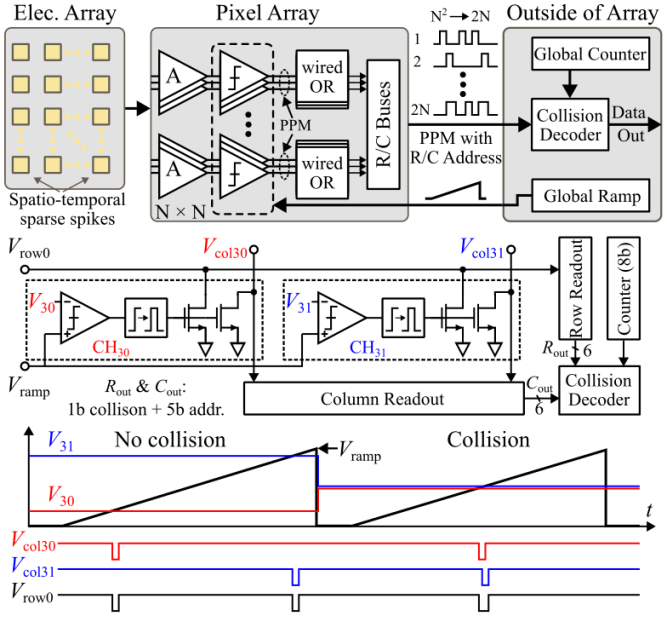


Fig. 6. Block diagram of wired-OR compressive readout with the collision principle illustrated with a two-channel example.

E. Data compression

To achieve data reduction without losing important information, on-chip spike detection, sorting algorithms, and data compression have been implemented [51], [52], [53], [54], [55], [56], effectively reducing the data volume that needs to be stored and/or transmitted. However, these solutions compress the data only after quantization, leading to large power consumption in the AFE and cache memory. To address this data deluge issue, compressed sensing has been employed to compress the input neural signals prior to digitization, at the cost of utilizing complex circuitry to perform analog product-sum operations [57], [58]. In [59], a compelling approach was proposed that combines both multiplexing and data compression using analog channel superposition. The primary disadvantage lies in the noise accumulation from the superimposed channels, which imposes restricting limitations on scalability.

To overcome these limitations, the wired-OR compressive readout has emerged as a promising alternative, providing both data compression and channel multiplexing in mixed-signal domain for action potential recordings [60]. As shown in Fig. 6, this method takes advantage of the spatio-temporal sparsity of neural spikes [29]. First, each input from the electrode array (N x N) is conditioned by an amplifier tuned to the band of interest. A continuous-time (CT) comparator in each pixel applies pulse position modulation (PPM) to the amplifier's output using a globally distributed ramp signal. Then, the PPM output is routed to the row and column addresses of each pixel through wired-OR logic. In this way, the N x N array is simultaneously readout with a reduced number of wires (from N^2 to $2N$ when compared with conventional pixels). Outside the array, a collision decoder reads the wired-OR PPM outputs and assigns corresponding digital values based on a global counter synchronized with the ramp generator. When multiple pixels access the row or column buses during the same ramp step, decoding is not possible. These events (*collisions*) are discarded (i.e., not stored) by the decoder. Consequently, only data from pixels that provide a unique digital value within a single ramp period are stored (see [60], [61] for extensive validation of the wired-OR compression algorithm). Most of the time, electrodes record noise around the baseline (not unique signals) and only the few electrodes recording a spike will have a unique signal. Notably, this architecture can achieve ~100x compression of APs with <500nW/ch power consumption. Furthermore, the event-driven output of the wired-OR readout can enable low-power spike sorting on-chip at <75nW/ch power consumption and >1000x overall compression [62].

Tips and tricks

The design of front-end electronics for high-density neural recording in standard CMOS technologies has evolved over decades of research and innovation. However, failing to follow certain guidelines can be frustrating for inexperienced designers who have encountered large mismatches between silicon measurements and SPICE simulations. Hence, this section primarily delves into several common mistakes (often made by the authors of this paper), offering practical design tips and tricks. By sharing valuable experiences and insights from previous designs, we aim to raise awareness of these aspects, helping new designers minimize risks and improve their designs.

Minimum conductance (g_{min}) in simulation: In CMOS design, when dealing with fA-level currents and/or TΩ-range pseudo-resistors, setting a proper g_{min} value is critical for ensuring simulation accuracy. The default value (typically 1e-12) is often too large and inadequate for simulating circuits with very high impedance nodes. Therefore, designers should carefully adapt g_{min} to a sufficiently small value, such that the simulation reflects the actual behavior of these components while still ensuring simulation convergence does not get significantly affected.

Modeling of gate leakage: In front-end signal conditioning circuits, thick-oxide transistors are commonly chosen for input transistors and pseudo-resistors, due to their significantly lower gate leakage compared to thin-oxide transistors. However, in many standard CMOS technologies, the gate leakage of these devices is either inaccurately modeled, or completely omitted in simulation models. Designers should be aware of these potential limitations and may need to take additional steps to better evaluate the gate leakage and its potential impact on their designs.

Parasitic diode extraction: In post-layout simulations, it is important to enable the extraction of parasitic diodes. Designers should verify that this option is activated in the rule file as it may not be enabled by default. Failing to include these diodes can lead to inaccurate simulations, since even tiny leakage currents from these parasitic diodes can potentially degrade circuit performance or cause malfunctioning.

Gamma noise factor in weak inversion: Often, the gamma noise factor in weak inversion is overestimated in simulations. Most foundries focus on good digital models, and weak inversion is seen as “leakage” by digital designers. Since there is no conducting channel in weak inversion, there is no thermal noise, and the dominant noise mechanism is shot noise. Hence, the gamma noise factor can be accurately approximated by $n/2$, where n is the subthreshold slope factor.

Guard ring protection for pseudo-resistors: In the layout of pseudo-resistors, fully enclosing these components with guard rings is highly recommended. This helps isolate the highly sensitive pseudo-resistors from external interference and parasitic leakage paths (e.g., lateral BJT transistors formed with nearby wells) that may not be captured in simulations. Based on the author’s experience, using well-taps instead of full guard ring enclosures (to save area) can lead to significant mismatches between the measured resistance values of pseudo-resistors and those predicted by SPICE simulations.

Capacitor leakage: When implementing AC-coupling using high-density MOM capacitors with narrow finger spacings in scaled technology nodes (e.g., 22nm), the leakage of these coupling capacitors becomes increasingly significant, as illustrated in [35]. However, in certain cases, the observed MOM capacitor leakage (via indirect measurements) can be substantially lower than what SPICE simulations predict [35], [49], suggesting that leakage behavior of MOM capacitors may not be accurately modeled, similar to the gate leakage modeling in thick-oxide transistors.

Power-efficient standard-cell libraries: In low-frequency neural recordings, typically leakage power dominates the overall power consumption in digital circuits. Therefore, digital standard-cell libraries implemented with high-threshold-voltage (HVT) or ultra-high-threshold-voltage (UHVT) transistors are commonly used to reduce leakage currents and improve power efficiency. These libraries can offer significant power savings while maintaining the required performance.

Light sensitivity: AC-coupling with pseudo-resistor biasing is sensitive to light due to photo-induced leakage currents, which typically cannot be simulated. Designers, especially for those working on emerging optoelectrical neural applications, should be mindful of this issue. The light sensitivity can be mitigated through leakage compensation, which involves the use of deep N-wells and careful biasing of the OTA input at half the supply voltage [63]. Metal shielding is also effective at reducing the amount of light that can reach the photosensitive junctions, but it requires multiple metal layers to be effective.

Noise degradation in multi-channel recording systems: In large-scale neural recording systems employing digital-intensive architectures, the very close integration of digital and sensitive analog circuitry presents challenges in maintaining low noise performance. Digital switching in one channel can couple into adjacent analog channels, degrading overall noise performance. This issue becomes increasingly pronounced in densely packed designs, where device proximity and shared analog bias lines increase the likelihood of interferences. Hence, careful considerations must be paid to layout, isolation techniques, and bias distribution in order to minimize inter- and cross-channel interference, preserving signal integrity. In addition, attention must be paid to voltage drops when distributing long shared analog bias lines.

Tissue bias in DC-coupled AFE: When the AFE is DC-coupled to the electrode, the tissue bias (V_B) set through a reference electrode plays a critical role due to its direct impact on the AFE’s input common-mode voltage. This voltage is determined by the tissue bias and the DC voltage drop, which is set by the half-cell potential and the leakage current flowing through the total series resistance (R_{DL} and R_s) at the electrode-tissue interface. However, leakage current, primarily originating from ESD protection diodes and AFE input gate leakage, is often not well-controlled. For this reason, designers may need to adjust the tissue bias in order to ensure the AFE input remains within the allowed input common-mode range.

Conclusion and future perspectives

The development of large-scale implantable brain-computer interfaces (BCIs) requires the design of power- and area-efficient front-end electronics that can effectively address critical application challenges, such as low noise, large EDO tolerance, high input impedance, and efficient data compression. This paper reviews recent advancements in front-end electronics reported over the past years, categorizing the various architectures and circuit techniques, and examining their respective characteristics, advantages, and limitations.

Looking forward, while it remains quite challenging to predict how exactly BCIs will evolve, a broader perspective reveals several primary avenues that hold promise for future explorations.

Performance improvements: Increasing the number of parallel readout channels is still highly desirable, as the current number of simultaneously recorded neurons is far from sufficient to understand the whole brain of a mouse, monkey, or human [64]. To achieve this, further enhancing recording channel performance in terms of power and area remains critical. Additionally, extending the EDO compensation range in DC-coupled front-ends is desirable for reliable operation across a wide range of conditions (such as those involving very small electrodes with significant EDOs).

Integration of new functionalities: System-level integration of new functionalities, such as stimulation, is essential in establishing a bidirectional neural interface that enables closed-loop neuromodulation [65]. Additionally, wireless power and data telemetry are also highly desirable to overcoming the limitations of conventional wired systems, which limits the natural movements of the subject under study and increases the risk of infection [66]. On-chip signal processing is also becoming increasingly important for reducing the amount of neural data to be transmitted, while machine-learning based biomarker extraction and classification improve detection and/or prediction accuracy for various neurological diseases [67].

Application-tailored optimization: It is important to note that most current recording systems have been developed primarily for fundamental neuroscience research, focusing on full-bandwidth, low-noise recordings for flexibility but at the cost of high power

consumption. While these wide-band and low-noise neural signals might be of high interest for neuroscientists, they are not necessarily essential for certain clinical purposes, such as deducing brain state and intent [68]. In fact, several recent studies have shown that for clinical BCIs, certain design specifications (e.g., bandwidth and resolution) can be relaxed without significantly sacrificing performance [69], [70], [71], [72] . This opens a new avenue for exploring different recording strategies tailored to various applications, thereby enabling more efficient and targeted solutions that maintain essential functionality and performance while minimizing overall system power and footprint.

Overall, as an emerging research field that is evolving rapidly, the design of large-scale BCIs involves a wide spectrum of disciplines, such as micro-fabrication, microelectronics, bioelectronics, communication, micro-system integration, neurobiology, neuroscience, signal processing, artificial intelligence, etc. It is expected that with further development in these disciplines, future BCIs will be smaller and more energy-efficient, fully wireless, with higher level of integration, and more intelligent to cover a broader range of applications. Importantly, the recent industrial interest in BCIs for clinical applications will provide further innovation and facilitate translational efforts from research into products [73].

References (with comments on those of special interest)

[1] M. A. Lebedev et al., "Brain-machine interfaces: From basic science to neuroprostheses and neurorehabilitation," *Physiolog. Rev.,* 2017.

[2] M. W. Slutzky, "Brain-machine interfaces: Powerful tools for clinical treatment and neuroscientific investigations," *Neuroscientist,* 2019.

[3] K. J. Miller et al., "The current state of electrocorticography-based brain-computer interfaces," *Neurosurg. focus,* 2020.

[4] L. Luan et al., "Emerging penetrating neural electrodes: in pursuit of large scale and longevity," *Annu Rev Biomed Eng.,* 2023.

[5] G. Buzsáki et al., "The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes," *Nat. Rev. Neurosci.,* 2012.

[6] L. A. Geddes et al., "Principles of applied biomedical instrumentation," *J. Clin. Eng.,* 1977.

[7] M. Sharma et al., "Acquisition of neural action potentials using rapid multiplexing directly at the electrodes," *Micromachines,* 2018.

[8] G. Hong et al., "Novel electrode technologies for neural recordings," *Nat. Rev. Neurosci.,* 2019.

[9] D. A. Schwarz et al., "Chronic, wireless recording of large-scale brain activity in freely moving rhesus monkeys," *Nat. Methods,* 2014.

[10] D. G. Muratore et al., "Artificial Retina: A future cellular-resolution brain-machine interface," *NANO-CHIPS 2030: On-Chip AI for an Efficient Data-Driven World,* 2020.

[11] S. Wang et al., "A compact chopper stabilized Δ-ΔΣ neural readout IC with input impedance boosting," *OJSSCS,* 2021.

[12] S. I. Ryu et al., "Human cortical prostheses: Lost in translation?," *Neurosurg Focus,* 2009.

[13] S. Kim et al., "Thermal impact of an active 3-D microelectrode array implanted in the brain," *IEEE Trans. Neural. Syst. Rehabil. ,* 2007.

[14] C. Y. Liu et al., "In vivo thermal evaluation of a subretinal prosthesis using an integrated resistance temperature detector," *J. Micro/Nanolith. MEMS MOEMS,* 2014.

[15] R. R. Harrison et al., "A low-power low-noise CMOS amplifier for neural recording applications," *IEEE JSSC,* 2003.

[16] Y. Tsividis, Operation and Modeling of the MOS Transistors, Oxford University Press, 1999.

[17] F. Maloberti, Data Converters, Springer, 2007.

[18] P. R. F. Rocha et al., "Electrochemical noise and impedance of Au electrode/electrolyte interfaces enabling extracellular detection of glioma cell populations," *Sci. Rep.,* 2016.

[19] "Tracking Advances in Neural Recording," [Online]. Available: https://stevenson.lab.uconn.edu/scaling. [Accessed 17 Oct. 2024].

[20] *X. Huang et al., "Actively multiplexed μEcoG brain implant system with incremental-ΔΣ ADCs employing bulk-DACs," *IEEE JSSC,* 2022.

*This paper presents a scalable neural recording architecture that combines a flexible active electrode array with a multiplexed CMOS readout IC, overcoming the wiring bottleneck in conventional passive arrays. Signals from 256 (16 x 16) electrodes are recorded with 16 analog signal lines and 16 digital multiplexing lines. Multiplexing ratio and ON resistance of the TFT switches are optimized to avoid significant noise folding. In addition, a novel bulk-DAC is proposed to improve power-area-efficiency of feedback DACs in direct-digitization neural readouts. Multiplexed EDOs are compensated efficiently without requiring high-resolution DACs typically seen in traditional mixed-signal DSLs.

[21] H. Londoño-Ramírez et al., "Multiplexed surface electrode arrays based on metal oxide thin-film electronics for high-resolution cortical mapping," *Adv. Sci.,* 2024.

[22] X. Sheng et al., "High-resolution spatial mapping of electrocorticographic activities with a 4096-channel, multiplexed thin-film transistor array," *bioRxiv,* 2024.

[23] Y. Tchoe et al., "Human brain mapping with multithousand-channel PtNRGrids resolves spatiotemporal dynamics," *Sci. Transl. Med.,* 2022.

[24] B. C. Raducanu et al., "Time multiplexed active neural probe with 678 parallel recording sites," in *IEEE ESSDERC*, 2016.

[25] K. Myny et al., "The development of flexible integrated circuits based on thin-film transistors," *Nat. Electron,* 2018.

[26] C. Mora Lopez et al., "A neural probe with up to 966 electrodes and up to 384 configurable channels in 0.13 μm SOI CMOS," *IEEE TBCAS,* 2017.

[27] D. De Dorigo et al., "Fully immersible subcortical neural probes with modular architecture and a delta-sigma ADC integrated under each electrode for parallel readout of 144 recording sites," *IEEE JSSC,* 2018.

[28] *D. Wendler et al., "A 0.0046-mm2 two-step incremental delta-sigma analog-to-digital converter neuronal recording front end with 120-mVpp offset compensation," *IEEE JSSC,* 2023.

*This paper presents a compact, low-power nerual readout IC that integrates DC-coupled neural AFEs with two-step incremental conversion and IDAC-based EDO compensation. It offers detailed discussions on the extended input range and the noise degradation caused by the IDAC-based EDO compensation.

[29] *M. Jang et al., "A 1024-channel 268-nW/pixel 36 ×36 μm2/channel data-compressive neural recording IC for high-bandwidth brain-computer interfaces," *IEEE JSSC,* 2023.

*This paper describes a lossy compressive readout for action potentials and its hardware implementation for massively parallel BCIs with active digital pixels. A distributed single-slope ADC performs pulse-position modulation on each pixel, and the OR combination of the outputs across rows and columns in the array is read by a decoder to perform compression and event-driven output. Reconstructed signals are sufficient to perform spike sorting with high accuracy, and the implementation is low-power and low-area.

[30] J. J. Jun et al., "Fully integrated silicon probes for high-density recording of neural activity," *Nature,* 2017.

[31] N. A. Steinmetz et al., "Neuropixels 2.0: A miniaturized high-density probe for stable long-term brain recording," *Science,* 2021.

[32] X. Yang et al., "A highly-integrated 1536-channel quad-shank monolithic neural probe in 55nm CMOS for full-band raw-signal recording," in *IEEE VLSI*, 2024.

[33] E. F. Chang, "Towards large-scale human-based mesoscopic neurotechnologies," *Neuron,* 2015.

[34] M. Sporer et al., "NeuroBus - Architecture for an ultra-flexible neural interface," *IEEE TBCAS,* 2024.

[35] X. Huang et al., "A compact, low-power analog front-end with event-driven input biasing for high-density neural recording in 22-nm FDSOI," *IEEE TCSII,* 2022.

[36] W. Jiang et al., "A ±50-mV linear-input-rage VCO-based neural-recording front-end with digital nonlinearity correction," *IEEE JSSC,* 2017.

[37] *H. Chandrakumar et al., "An 80-mVpp linear-input range, 1.6-GΩ input impedance, low-power chopper amplifier for closed-loop neural recording that is tolerant to 650-mVpp common-mode interference," *IEEE JSSC,* 2017.

*This paper presents a DC-coupled chopper amplifier for closed-loop neuromodulation, featuring high input impedance and large tolerance to common-mode interference. Large electrode DC offsets are cancelled through an analog DSL, where multi-rate duty-cycled resistors are proposed to increase the effective resistance compared to single-rate duty-cycled resistors. The key idea is to reduce the switching frequency (lower than signal bandwidth), with in-band switching artefact being effectively suppressed by anti-alias filtering.

[38] *C. Livanelioglu et al., "A 0.0014 mm2, 1.18 TΩ segmented duty-cycled resistor replacing pseudo-resistor for neural recording interface circuits," in *IEEE VLSI*, 2022.

*This paper introduces a segmented duty-cycled resistor (SDR) for neural recording amplifiers. The resistor segmentation reduces parasitic charge transfer compared to conventional duty-cycled resistors, achieving effective resistance in the TΩ range without lowering the switching frequency into the signal band. In addition, the SDR's PVT robustness allows for stable cut-off frequencies across various conditions.

[39] M. Jang et al., "A 1024-channel 268 nW/pixel 36x36 μm2/ch data-compressive neural recording IC for high-bandwidth brain-computer interfaces," in *IEEE VLSI*, 2023.

[40] H. Gao et al., "HermesE: A 96-channel full date rate direct neural interface in 0.13 μm CMOS," *IEEE JSSC,* 2012.

[41] H. Chandrakumar et al., "A 15.2-ENOB 5-kHz BW 4.5-μW chopped CT ΔΣ-ADC for artifact-tolerant neural recording front ends," *IEEE JSSC,* 2018.

[42] J. Huang et al., "A 0.01-mm2 mostly digital capacitor-less AFE for distributed autonomous neural sensor nodes," *IEEE LSSC,* 2018.

[43] B. Abdelgaliel et al., "A 2 × 2 neural amplifier macro-pixel with shared DC servo loop for high-density brain-computer interfaces," in *IEEE BioCAS*, 2024.

[44] *R. Muller et al., "A 0.013 mm2, 5 μW, DC-coupled neural signal acquisition IC with 0.5 V supply," *IEEE JSSC,* 2012.

*This paper presents a compact and power-efficient DC-coupled neural readout IC with mixed-signal DSLs. Merged-amplifier DAC is proposed to compensate for large EDOs, preventing thermal noise degradation typically seen with traditional IDACs. Additionally, an input bias network is carefully designed to be compatible with electrode characteristics, ensuring the DC-coupled AFE maintains the desired input common-mode range.

[45] *X. Huang et al., "A 3072-channel neural readout IC with multiplexed two-step incremental-SAR conversion and bulk-DAC-based EDO compensation in 22nm FDSOI," in *IEEE VLSI*, 2024.

*This paper presents a high-channel-count neural readout IC featuring multiplexed, DC-coupled AFE with bulk-DAC-based EDO compensation. The bulk-DAC compensates for large EDOs without degrading both thermal and flicker noise compared to traditional IDAC and merged amplifier-DAC. Two-step incremental-SAR conversion significantly reduces oversampling frequency of the ADC, improving power efficiency for simultaneous local field potential and action potential recording.

[46] X. Huang et al., "A 256-channel actively-multiplexed μECoG implant with column-parallel incremental ΔΣ ADCs employing bulk-DACs in 22-nm FDSOI technology," in *IEEE ISSCC*, 2022.

[47] C. M. Lopez et al., "An implantable 455-active-electrode 52-channel CMOS neural probe," *IEEE JSSC,* 2014.

[48] E. Greenwald et al., "A bidirectional neural interface IC with chopper stabilized BioADC array and charge balanced stimulator," *IEEE TBCAS,* 2016.

[49] X. Yang et al., "An AC-coupled 1st-order Δ-ΔΣ readout IC for area-efficient neural signal acquisition," *IEEE JSSC,* 2023.

[50] J. Steensgaard et al., "Noise–power optimization of incremental data converters," *IEEE TCSI,* 2008.

[51] S.-Y. Park et al., "Dynamic power reduction in scalable neural recording interface using spatiotemporal correlation and temporal sparsity of neural signals," *IEEE JSSC,* 2018.

[52] M. A. Shaeri et al., "A method for compression of intra-cortically-recorded neural signals dedicated to implantable brain–machine interfaces," *IEEE TNSRE,* 2014.

[53] S. M. A. Zeinolabedin et al., "A 16-channel fully configurable neural SoC with 1.52 μW/Ch signal acquisition, 2.79 μW/Ch real-time spike classifier, and 1.79 TOPS/W deep neural network accelerator in 22 nm FDSOI," *IEEE TBCAS,* 2022.

[54] Y. Chen et al., "An online-spike-sorting IC using unsupervised geometry-aware OSort clustering for efficient embedded neural-signal processing," *IEEE JSSC,* 2023.

[55] C. Aprile et al., "Adaptive learning-based compressive sampling for low-power wireless implants," *IEEE TCSI,* 2018.

[56] T. Wu et al., "Deep compressive autoencoder for action potential compression in large-scale neural recording," *J. Neural Eng.,* 2018.

[57] M. Shoaran et al., "A low-power area-efficient compressive sensing approach for multi-channel neural recording," in *IEEE ISCAS*, 2013.

[58] T. Okazawa et al., "A time-domain analog spatial compressed sensing encoder for multi-channel neural recording," *Sensors,* 2018.

[59] J. D. Rieseler et al., "A superposition-based analog data compression scheme for massively-parallel neural recordings," in *IEEE BioCAS*, 2017.

[60] D. G. Muratore et al., "A data-compressive wired-OR readout for massively parallel neural recording," *IEEE TBCAS,* 2019.

[61] P. Yan et al., "Data compression versus signal fidelity tradeoff in wired-OR analog-to-digital compressive arrays for neural recording," *IEEE TBCAS,* 2023.

[62] A. Akhoundi et al., "A 1024-channel 0.00029mm²/ch 74nW/ch online spatial spike sorting chip with event-driven spike detection and self-organizing map clustering," in *IEEE ISSCC*, 2025.

[63] S. Wang et al., "Leakage compensation scheme for ultra-high-resistance pseudo-resistors in neural amplifiers," *Electron. Lett.,* 2019.

[64] A. E. Urai et al., "Large-scale neural recordings call for new insights to link brain and behavior," *Nat. Neurosci.,* 2022.

[65] A. E. Mendrela et al., "A bidirectional neural interface circuit with active stimulation artifact cancellation and cross-channel common-mode noise suppression," *IEEE JSSC,* 2016.

[66] A. B. Islam et al., "Wireless power transfer, recovery, and data telemetry for biomedical applications," in *Handbook of Biochips: Integrated Circuits and Systems for Biology and Medicine*, Springer, 2022.

[67] U. Shin et al., "NeuralTree: A 256-channel 0.227-μJ/class versatile neural activity classification and closed-loop neuromodulation SoC," *IEEE JSSC,* 2022.

[68] E. M. Trautmann et al., "Accurate estimation of neural population dynamics without spike sorting," *Neuron,* 2019.

[69] *****N. Even-Chen et al., "Power-saving design opportunities for wireless intracortical brain-computer interfaces," *Nat. Biomed. Eng.,* 2020.

*****This paper studies the requirements for the recording channels in intracortical motor BCIs based on the output of the motor decoder. The work is validated with in-vivo recordings from primate and human participants and shows that an application-specific design can enable more than an order of magnitude power savings.

[70] 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.,* 2020.

[71] J. Lim et al., "A light-tolerant wireless neural recording IC for motor prediction with near-infrared-based power and data telemetry," *IEEE JSSC,* 2022.

[72] G. Atzeni et al., "An impedance-boosted transformer-first discrete-time analog front-end achieving 0.34 NEF and 389-MΩ input impedance," *IEEE JSSC,* 2024.

[73] L. Drew et al., "Decoding the business of brain-computer interfaces," *Nat. Electron.,* 2023.