FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

# Closed-loop Systems for CMOS-based Neural Probes in Electrophysiology Experiments

**Maria Cunha Ribeiro Rocha**







Master's in Bioengineering

Supervisor at FEUP: Prof. Luís Jacinto

Supervisor at IIT: Prof. Luca Berdondini

October 13, 2026

# Closed-loop Systems for CMOS-based Neural Probes in Electrophysiology Experiments

**Maria Cunha Ribeiro Rocha**

Master's in Bioengineering

October 13, 2026

# Abstract

Closed-loop neural interfaces, systems that can detect neural activity and respond to it in real time, are becoming increasingly important for both neuroscience research and future clinical neurotechnology. However, building such systems in practice requires solving several interconnected problems, including reliable high-density neural recording, probe insertion with minimal penetration trauma, low-latency signal processing and robust extraction of meaningful neural events from noisy data, all under strict real-time constraints. This dissertation presents the design, implementation and validation of a closed-loop electrophysiology platform built around the SiNAPS CMOS neural probe, while utilizing the NeuralGlider ultrasonic insertion system. The NeuralGlider was first evaluated across dura-intact and durotomy conditions, with and without ultrasonic actuation, over in vivo recording sessions, to establish whether this insertion method reliably improves signal quality, a prerequisite for any downstream real-time processing. Building on this, a dual-workstation architecture coordinated by an Opal Kelly FPGA was evaluated for its ability to acquire, visualize and process neural data in parallel, achieving a mean closed-loop latency of 3.04 ms between spike detection and actuation, with a jitter of 0.453 ms. A causal spike detection algorithm utilizing an adaptive noise-tracking threshold with the ASO operator was then implemented, with detection parameters selected through a systematic grid search. This algorithm was validated across three levels of increasing complexity: synthetic recordings with known ground truth, in vitro recordings with optogenetic stimulation requiring artifact rejection and in vivo recordings benchmarked against offline Kilosort4 sorting. As an additional exploratory line of this work, two online spike sorting strategies, a single-channel K-means baseline and a graph-based Louvain community detector, were evaluated across both single-channel and multichannel configurations. While both were fast enough to run in real time, neither provided a complete sorting solution. Overall, these results demonstrate that a stable low-latency closed-loop pipeline is achievable on high-density recordings, while also revealing the practical limits of real-time processing relative to offline sorting, pointing toward richer biomarkers and improved online sorting strategies as key directions for future work.

**Keywords**: Ultrasonic insertion; Closed-loop neural interface; Real-time spike detection; High-density CMOS probe; Spike sorting; Neural signal processing

# Resumo

As interfaces neuronais em ciclo fechado, sistemas capazes de detetar a atividade neuronal e de lhe responder em tempo real, estão a tornar-se cada vez mais importantes, tanto para a investigação em neurociências como para a futura neurotecnologia clínica. No entanto, a construção de tais sistemas exige a resolução de vários problemas, incluindo o registo neuronal de alta densidade, a inserção de sondas com trauma de penetração mínimo e o processamento de sinal e extração robusta de eventos neuronais significativos em tempo real. Esta dissertação apresenta o projeto, a implementação e a validação de uma plataforma de eletrofisiologia em ciclo fechado desenvolvida em torno da sonda neuronal SiNAPS CMOS, utilizando o sistema de inserção ultrassónica NeuralGlider. O NeuralGlider foi primeiramente avaliado em condições de dura-máter intacta e de durotomia, com e sem atuação ultrassónica, ao longo de sessões de registo in vivo, para determinar se este método de inserção melhora de forma fiável a qualidade do sinal, um pré-requisito para qualquer processamento em tempo real. Adicionalmente, foi avaliada uma arquitetura de duas estações de trabalho coordenada por um FPGA Opal Kelly quanto à sua capacidade de adquirir, visualizar e processar dados neuronais em paralelo, alcançando uma latência média em ciclo fechado de 3,04 ms entre a deteção de spikes e a atuação, com um jitter de 0,453 ms. Posteriormente, foi implementado um algoritmo causal de deteção de spikes que utiliza um limiar adaptativo de rastreio de ruído com o operador ASO, tendo os parâmetros de deteção sido selecionados através de uma procura em grelha sistemática. Este algoritmo foi validado em três níveis de complexidade: registos sintéticos com ground truth conhecido, registos in vitro com estimulação optogenética a exigir a rejeição de artefactos e registos in vivo comparados com a deteção offline do Kilosort4. Como uma adição, foram avaliadas duas estratégias de spike sorting em tempo real, uma assente em K-means e um detetor de comunidades de Louvain, em caracteristicas de monocanal e multicanal. Embora ambas tenham sido suficientemente rápidas para execução em tempo real, nenhuma forneceu uma solução completa de spike sorting. De forma geral, estes resultados demonstram que é possível obter um pipeline estável e de baixa latência em ciclo fechado para registos de alta densidade, assim como revelam os limites práticos do processamento em tempo real em comparação com o sorting offline, apontando para biomarcadores mais ricos e estratégias otimizadas de sorting em tempo real como direções fundamentais para trabalhos futuros.

**Keywords**: Inserção ultrassónica; Interface neuronal em ciclo fechado; Deteção de spikes em tempo real; Sonda CMOS de alta densidade; Ordenação de spikes; Processamento de sinal neuronal

# Contents


**1 Introduction** **1**
- 1.1 Context and Motivation . . . 1
- 1.2 Objectives . . . 2
- 1.3 Contributions . . . 3
- 1.4 Sustainability and Global Impact . . . 3
- 1.5 Dissertation outline . . . 4

**2 Background** **5**
- 2.1 Electrophysiology Fundamentals . . . 5
  - 2.1.1 Neuron Structure and Action Potentials . . . 5
  - 2.1.2 Electrophysiology monitoring . . . 9
- 2.2 CMOS-based neural probes . . . 11
  - 2.2.1 The Evolution Towards Active CMOS Neural Probes . . . 11
  - 2.2.2 The CMOS technology . . . 12
  - 2.2.3 Comparative Analysis of State-of-the-art Platforms . . . 13
  - 2.2.4 Architectural Evolution . . . 14
  - 2.2.5 Structural Evolution . . . 14
  - 2.2.6 Material Evolution . . . 15
- 2.3 Probe Insertion Strategies and Biomechanical Integration . . . 16
- 2.4 Closed-loop . . . 17
  - 2.4.1 The Limitations of Open-Loop Systems . . . 18
  - 2.4.2 Biomarkers . . . 19
  - 2.4.3 Feature Extraction . . . 20
  - 2.4.4 Signal Processing and Artifact Mitigation Pipelines . . . 22
  - 2.4.5 Control Theory and Feedback Loops . . . 23
  - 2.4.6 Neuro-Actuation and Stimulation Modalities . . . 25
  - 2.4.7 Integrated Sensing, Stimulation and Closed-Loop CMOS Architectures 26
  - 2.4.8 Development and Applications . . . 26
  - 2.4.9 Current Challenges . . . 29

**3 Materials and Methods** **31**
- 3.1 AI models . . . 31
- 3.2 SiNAPS probe . . . 32

3.3 NeuralGlider . . . 32
3.4 Electrophysiological Data . . . 33
3.4.1 Synthetic Data . . . 33
3.4.2 In vitro Data . . . 33
3.4.3 In vivo Data . . . 34
3.4.4 Closed-loop simulation data . . . 36
3.5 Software Environment . . . 36
3.5.1 Feature Extraction from In Vivo Data . . . 36
3.5.2 Statistical Analysis . . . 37
3.5.3 Data filtering . . . 38
3.5.4 Spike Detector . . . 39
3.5.5 Real-Time Spike Detection Pipeline . . . 43
3.5.6 Real-Time Spike Sorting and Clustering . . . 45
3.5.7 Validation with Ground Truth . . . 48
3.6 Closed-loop System Architecture . . . 48

**4 Results and Discussion** **53**
4.1 Neural Glider - probe insertion . . . 53
4.2 Synthetic data - spike detection . . . 63
4.3 In vitro data - artifact removal and spike detection . . . 64
4.4 Real time spike detection and sorting . . . 72
4.4.1 In vitro data . . . 73
4.4.2 Synthetic data . . . 78
4.4.3 In vivo data . . . 84
4.5 Closed-loop system . . . 90
4.6 Summary . . . 91

**5 Conclusion** **93**
5.1 Future Work . . . 94

**Referências** **97**

# List of Figures


2.1 Anatomical structure of a biological neuron . . . . . 6
2.2 Stages of the action potential and associated ion channel activity . . . . . 8

3.1 Closed-loop system implementation architecture . . . . . 50
3.2 Python Script implemented for Closed-loop Control . . . . . 51

4.1 Analysis of spike waveforms, template morphology and SNR stability . . . . . 55
4.2 Assessment of global single-unit signal quality across all identified units and raw background noise tracking during a representative NeuralGlider recording session . . . . . 57
4.3 Smoothed spike SNR trajectories calculated across all identified single units on their respective primary channels, categorized by dura status and actuation state during NeuralGlider insertion . . . . . 59
4.4 Boxplot comparisons of signal and unit metrics across recording halves and conditions . . . . . 61
4.5 Comparative evaluation of spike detection and feature extraction algorithms on synthetic extracellular traces . . . . . 64
4.6 Overview of spike detection performance on real-time extracellular data, along with template shape comparisons and feature space distributions for extracted single units . . . . . 66
4.7 Evaluation of Common Average across all channels during optogenetic LED stimulation . . . . . 68
4.8 Algorithmic Parameter Optimization for Post-CAR Spike Detection and Optical Artifact Suppression . . . . . 70
4.9 Single-channel firing rate dynamics . . . . . 72
4.10 Evaluation of real-time spike detection performance and population firing dynamics across an optogenetic recording session . . . . . 75
4.11 Performance comparison between the real-time spike detector and the offline adaptive moving median benchmark . . . . . 77
4.12 Real-time single-feature $k$-means spike sorting performance on synthetic extracellular data . . . . . 79
4.13 Real-time two-feature spike sorting performance on synthetic extracellular data 81
4.14 Real-time three-feature spike sorting performance on synthetic extracellular data 83

4.15 Validation of real-time ASO spike detection against offline Kilosort4 benchmarks on NeuralGlider data . . . . . 85
4.16 Real-time streaming Louvain community clustering vs. Kilosort4 (KS4) ground truth on in-vivo data . . . . . 88
4.17 Closed-loop system implementation focused on the latency definition. Blue arrows represent data flow and green arrows represent closed-loop feedback and actuation signal paths. . . . . 90
4.18 Evaluation of real-time spike detection and stimulation feedback timing during closed-loop operation . . . . . 91

# List of Tables


2.1 Comparative overview of state-of-the-art active CMOS neural probe platforms. 13

3.1 Summary of experimental in vivo recording sessions categorized by NeuralGlider actuation state and dura integrity. . . . . 35

4.1 Single-unit quality metrics and yield across NeuralGlider actuation states and dura conditions . . . . . 54
4.2 Signal-to-Noise Ratio (SNR) and Trend Analysis across Experimental Sessions 60
4.3 Independent-samples $t$-test results comparing single-unit performance metrics across recording halves . . . . . 62
4.4 Independent-samples $t$-test results comparing actuation states in intact dura during the first recording half . . . . . 62
4.5 Spike extraction metrics across signal processing stages, for in vitro data. . . . 71
4.6 Real-time processing latency (average execution time per data chunk), execution budget (chunk time duration) and total spike detection yield across the entire recording for different buffer chunk sizes using in vitro data. . . . . 76
4.7 Spike detection comparison report for in vivo data . . . . . 78
4.8 Spike detection validation metrics and real-time processing latency per channel benchmarked against Kilosort4 (KS4). . . . . 86
4.9 Performance and processing latency of real-time single-channel (Channel 413) vs. multichannel Louvain spike sorting against Kilosort4. . . . . 89

# Acronyms

| | |
|---|---|
| ADC | Analog-to-Digital Conversion |
| ADO | Absolute Differential Operator |
| AI | Artificial Intelligence |
| APS | Active Pixel Sensor |
| ASO | Amplitude Slope Operator |
| ATP | Adenosine Triphosphate |
| CLoSES | Closed-Loop System for Electrical Stimulation |
| CMOS | Complementary Metal-Oxide-Semiconductor |
| CPU | Central Processing Units |
| DC | Direct Current |
| ECAP | Evoked Compound Action Potentials |
| EEG | Electroencephalography |
| FBR | Foreign Body Reaction |
| FDA | United States Food and Drug Administration |
| fMRI | Functional Magnetic Resonance Imaging |
| fNIRS | Functional Near-Infrared Spectroscopy |
| FPGA | Field-Programmable Gate Array |
| LED | Light-emitting diode |
| LFP | Local Field Potential |
| LSTM | Long Short-Term Memory |
| MEA | Multi-Electrode Arrays |
| MUA | Multi-Unit Activity |
| NMOS | N-type Metal-Oxide-Semiconductor |
| PCA | Principal Component Analysis |
| PID | Proportional-Integral-Derivative |
| RMS | Root-mean-square |
| RNN | Recurrent Neural Network |
| SDGs | Sustainable Development Goals |
| SiNAPS | Simultaneous Neural Active Pixel Sensor |
| SUA | Single-Unit Activity |
| TDM | Time-Division Multiplexing |

# Chapter 1

# Introduction

## 1.1 Context and Motivation

The central nervous system processes information through high-dimensional networks of interconnected neurons. Understanding these neural dynamics is essential both for fundamental neuroscience, unravelling how neural circuits encode behaviour, cognition and perception, and for clinical neurology, where mapping network activity is crucial for diagnosing and treating neurological and psychiatric disorders such as epilepsy, Parkinson's disease and stroke [1].

Historically, electrophysiological studies relied on passive low-density microelectrodes with limited spatial coverage, offering only a sparse view of neural dynamics [2]. To overcome this limitation, the research paradigm has driven recent technological advancements toward active Complementary Metal-Oxide-Semiconductor (CMOS) interfaces. By integrating microelectronic signal-conditioning circuitry directly onto the probe substrate, CMOS technology enables thousands of densely packed recording sites without requiring a dedicated output wire for every channel [3]. As a key example of CMOS-based probes, the SiNAPS platform [4] leverages this monolithic architecture to achieve simultaneous large-scale monitoring of neural ensembles across cortical and subcortical structures with single-neuron resolution [1, 4, 5].

Parallel to these hardware advancements, neuromodulation paradigms have been shifting from traditional open-loop approaches, where electrical stimulation is delivered on a fixed schedule, toward adaptive closed-loop architectures. Open-loop approaches often suffer from limited efficacy, accelerated battery depletion and side effects resulting from unnecessary tissue stimulation. Closed-loop systems address these limitations by evaluating real-time neural brain activity biomarkers to deliver targeted stimulation only when pathologically necessary [6, 2]. While clinically focused systems often rely on electrical microstimulation, optogenetics has emerged as a

powerful modality in preclinical research, using genetically encoded light-sensitive ion channels to enable cell-type-specific neural modulation with minimal electrical artifact contamination [7].

However, closed-loop systems can only act on the neural activity they can reliably measure and process in real time. Realizing high-performance closed-loop neuromodulation in practice requires bridging the gap between high-density neural recording capacities and the physical and computational realities of real-time hardware-software architectures. At the physical level, as with all implantable neural probes, inserting a rigid silicon substrate into delicate brain parenchyma risks mechanical trauma, tissue dimpling and acute neuroinflammation that compromise initial signal quality [8]. Addressing these insertion mechanics using specialized tools, such as the vibration-assisted NeuralGlider system [9], is crucial for preserving tissue integrity and ensuring reliable signal acquisition. Simultaneously, at the processing level, high-density sensors generate massive high-bandwidth data streams. Performing real-time signal processing, specifically the detection and extraction of neuronal spikes across hundreds to thousands of channels, presents a major computational bottleneck [10, 11]. Without an optimized low-latency processing pipeline, the system cannot respond on the millisecond timescale required for effective neural control [12, 13]. Overcoming these biomechanical and computational constraints demands a platform capable of integrating both tissue-probe mechanics with ultra-low-latency real-time processing.

## 1.2 Objectives

The primary objective of this dissertation work was to design, implement and validate a closed-loop platform for high-density neural recording based on the SiNAPS CMOS probe [4], tackling the data acquisition and real-time processing challenges involved, while utilizing the NeuralGlider insertion system [9] to address insertion mechanics and tissue trauma.

This work first evaluated the NeuralGlider's capacity to insert the SiNAPS probe with minimal tissue trauma, tracking signal quality and single-unit yield during in vivo recording sessions. In parallel, an FPGA-centered acquisition architecture was developed to achieve low-latency online signal processing and effector control at high spatiotemporal resolution.

To process the high-bandwidth data stream generated by this FPGA architecture and provide the real-time feedback signal required for closed-loop control, a major focus was the design of a causal spike detection algorithm. This pipeline was developed to identify spiking activity under strict latency constraints, supported by a systematic parameter optimization procedure to maximize detection performance. The detector was validated across three distinct experimental conditions: synthetic recordings, in-vitro recordings conducted during optogenetic stimulation

(which served to evaluate the pipeline's ability to reject optical artifacts) and in vivo recordings benchmarked against offline Kilosort4 spike sorting.

Building on this detection pipeline, this dissertation evaluated and compared several real-time spike sorting strategies operating within the same latency budget. Resolving individual single-unit identities in real time is useful for closed-loop platforms, as it allows feedback interventions to target specific neural sub-populations rather than unselective population activity. This included comparing a simple fixed-cluster approach with a graph-based method capable of inferring the number of units automatically, as well as evaluating single-channel versus multi-channel feature sets. Together, these contributions provided a practical characterization of the latency and accuracy trade-offs involved in deploying high-density CMOS probes within a real-time closed-loop system.

## 1.3 Contributions

This dissertation provides a validated groundwork for a closed-loop platform for high-density electrophysiology, integrating acquisition hardware and real-time signal processing, complemented by an evaluation of microvibration-assisted probe insertion mechanics. Concretely, this includes an evaluation of the NeuralGlider insertion system's effect on recording stability; an FPGA-based dual-workstation acquisition architecture achieving sub-4-millisecond sensing-to-actuation latency with a jitter of 0.453 ms; a causal parameter-optimized spike detection algorithm validated across synthetic, in vitro and in vivo datasets, the last one against an offline reference sorter; and a comparative framework evaluating the feasibility and limitations of real-time spike sorting across various feature extraction and clustering methods. Together, these contributions provide a characterized reproducible foundation for building closed-loop experiments on high-density CMOS probes, along with a realistic account of the trade-offs involved in doing so in real time.

## 1.4 Sustainability and Global Impact

The development of advanced neural interfaces, such as the SiNAPS platform, is aligned with the United Nations Sustainable Development Goals (SDGs) [14], particularly regarding global health and sustainable innovation. This research contributes to SDG 3 (Good Health and Well-being) [14] by establishing technical foundations for more effective and personalized treatments for neurological disorders, a leading cause of disability worldwide. Closed-loop neuromodulation, delivering intervention only when needed, represents a shift toward more efficient

and personalized healthcare technologies that can reduce side effects and improve long-term quality of life.

From a technical perspective, this work supports SDG 9 (Industry, Innovation and Infrastructure) [14] by leveraging CMOS-integrated architectures and low-overhead real-time processing to reduce the computational and data-storage demands typically associated with high-density neural recording. The system's sensing-to-stimulation latency and its ability to run within strict real-time budgets on standard hardware serve as concrete indicators of this efficiency and support future integration into more resource-constrained implantable devices. Furthermore, these technical advancements provide essential tools for neuroscience research, enabling deeper insights into fundamental brain mechanisms.

## 1.5 Dissertation outline

The rest of this dissertation is organized as follows. Chapter 2 provides the necessary background, covering the biophysical fundamentals of neural signaling and activity acquisition methods, CMOS-based neural probe technology, probe insertion strategies, closed-loop control strategies, stimulation artifact mitigation and an overview of current applications and open challenges in the field. Chapter 3 describes the methodology developed in this work, including the SiNAPS probe and NeuralGlider insertion system, the signal preprocessing and real-time spike detection pipeline, the parameter optimization procedure, the spike sorting strategies compared across single- and multichannel configurations and the FPGA-based closed-loop architecture and its software implementation. Chapter 4 presents the experimental validation of this pipeline, progressing from NeuralGlider insertion performance, through detection and sorting results on synthetic and in vitro data, to in vivo validation against offline Kilosort4 sorting, to closed-loop latency. Finally, Chapter 5 summarizes the main findings of this work and outlines directions for future research.

# Chapter 2

# Background

This chapter presents an overview of the theoretical foundations, architectural principles and state-of-the-art strategies governing closed-loop CMOS neural probes, tracing the complete pipeline from hardware design and signal conditioning to real-time control and translational deployment.

## 2.1 Electrophysiology Fundamentals

Electrophysiology forms the foundational bridge between biological neural networks and microelectronic instrumentation [15]. This section details the fundamental principles of neuronal excitability and action potential dynamics, addresses how these biological dynamics convert into readable extracellular electrical signals and surveys the primary hardware modalities used for neural recording.

### 2.1.1 Neuron Structure and Action Potentials

The nervous system is a specialized cellular network dedicated to the generation, transmission and processing of information. Functionally, it acts as a communication apparatus that enables an organism to sense internal or external states and coordinate corresponding responses across other physiological systems. To convey these signals, the system utilizes a sophisticated combination of electrical and chemical messengers [16, 17].

The fundamental unit of this system is the neuron, an excitable cell specialized for processing and transmitting information throughout the nervous system using rapid and transient electrical signals known as action potentials. The anatomical structure of a classical neuron is organized

into functional zones optimized for receiving, integrating and propagating electrical events [16, 17], as illustrated in Figure 2.1. These include:

**Dendritic Tree**: Highly branched projections specialized for receiving incoming synaptic inputs.

**Soma (Cell Body)**: The metabolic center that integrates graded potentials received across the dendritic arbor.

**Nucleus**: The central organelle housing genetic material and directing cellular metabolism.

**Axon Hillock**: The specialized junction connecting the soma to the axon where action potentials are initiated.

**Axon**: A single, elongated projection responsible for propagating the action potential away from the cell body toward target cells.

**Myelin Sheath & Nodes of Ranvier**: In myelinated neurons, insulating layers of myelin are interrupted by uninsulated gaps (nodes of Ranvier) that house dense clusters of voltage-gated channels, enabling rapid signal propagation via saltatory conduction.

**Schwann Cell**: A glial cell in the peripheral nervous system responsible for producing the myelin sheath.

**Axon Terminals**: Specialized distal endings where the action potential completes its path down the axon to carry out synaptic transmission to target cells.

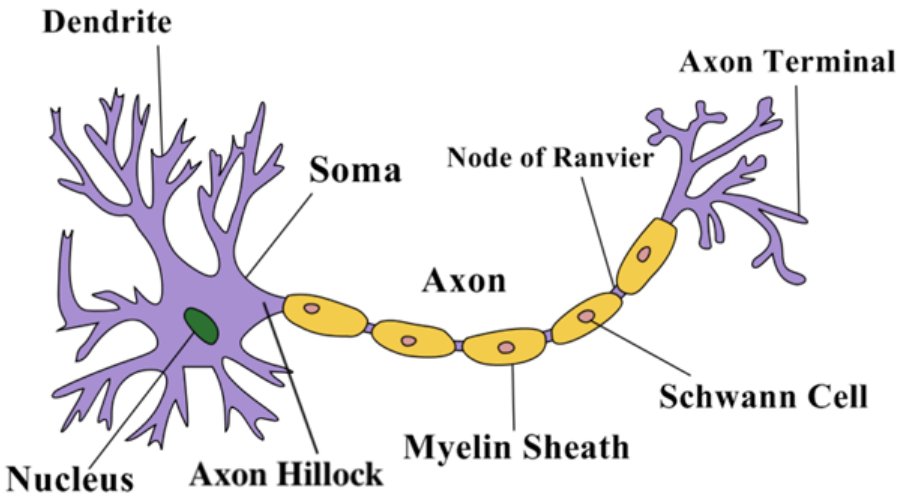



Figure 2.1: Anatomical structure of a biological neuron. Adapted from [17].

Under resting conditions, a neuron maintains a resting membrane potential typically ranging from $-60$ to $-70$ mV (measured inside relative to the outside). This potential is established and maintained by an unequal distribution of ions, primarily sodium ($Na^+$) and potassium ($K^+$),

which are actively regulated by an energy-dependent transporter known as the $Na^+/K^+$-ATPase (or sodium-potassium pump). This pump concentrates $Na^+$ in the extracellular space and $K^+$ within the cytosol by exchanging three internal $Na^+$ ions for two external $K^+$ ions per ATP molecule hydrolyzed [18]. However, the primary reason the inside of the cell remains negative relative to the outside is the high resting membrane permeability to potassium. The membrane contains abundant and continuously open $K^+$ leak channels. Driven by the concentration gradient established by the pump, positively charged $K^+$ ions continuously diffuse out of the cell, leaving behind impermeable negatively charged intracellular proteins and organic anions. This drives the negative resting potential close to the equilibrium potential for potassium [18, 17].

Information processing in the nervous system relies on neurotransmission, predominantly mediated by fast electrical signals (action potentials) within a neuron and chemical signaling (neurotransmitters) or direct electrical coupling (gap junctions) between neurons [18, 17].

**Phase 1: Action Potential Generation**

An action potential is a rapid, temporary flip in a neuron's electrical voltage. Initiation begins with local depolarizing events, such as sensory inputs at peripheral receptors or excitatory postsynaptic potentials triggered by neurotransmitter binding across the dendritic tree. These signals travel to the cell body and axon hillock (the axon's starting point). If the total voltage reaches a critical threshold (around $-45$ mV), an "all-or-nothing" action potential fires across three steps [18, 17]:

1. **Depolarization (Rising Phase):** Reaching the threshold opens voltage-gated $Na^+$ channels. Positively charged $Na^+$ ions rush into the cell, driving the internal voltage up from its negative baseline to about $+40$ mV.

2. **Repolarization (Falling Phase):** Within 1 ms, $Na^+$ channels automatically close (inactivate), stopping $Na^+$ entry. Simultaneously, voltage-gated potassium ($K^+$) channels open, letting positive $K^+$ ions rush out of the cell. This pulls the voltage back down toward negative levels.

3. **After-Hyperpolarization (Undershoot):** Because $K^+$ channels close slowly, $K^+$ continues leaving briefly, dropping the voltage slightly below its normal resting level before returning to baseline.

Throughout this cycle, the sodium-potassium pump ($Na^+/K^+$-ATPase) continuously works in the background to restore baseline ion balances [18, 17].

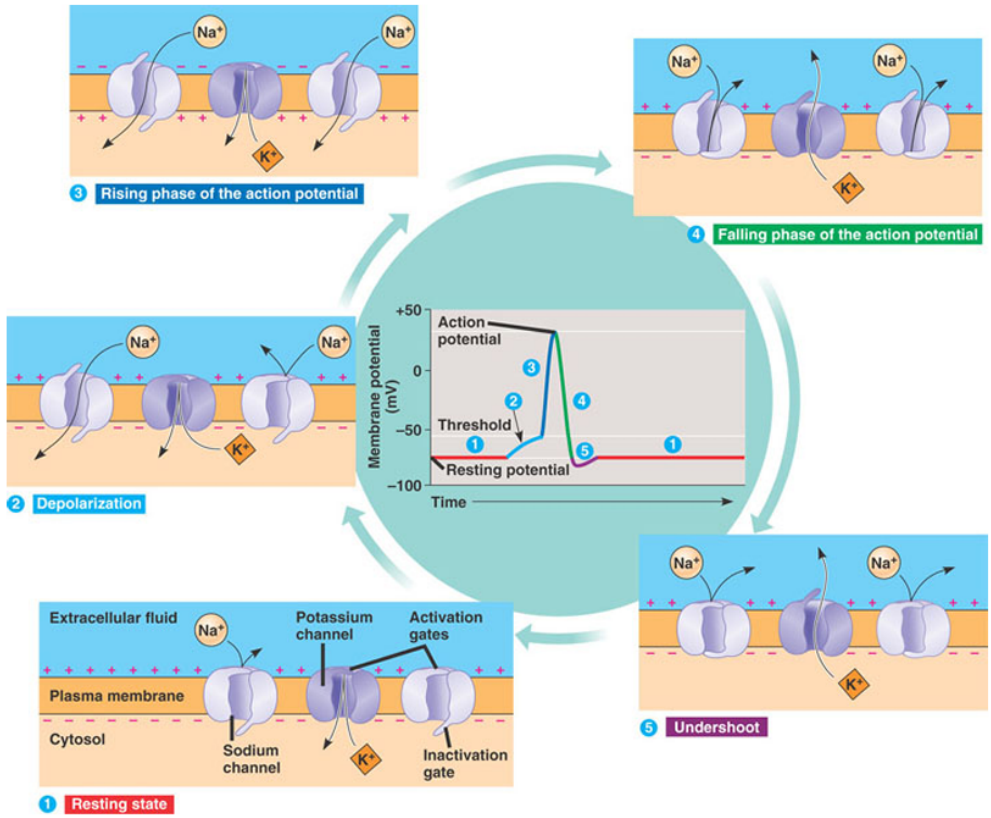



Figure 2.2: Stages of the action potential and associated ion channel activity. Adapted from [17]. **(1) Resting state:** Voltage-gated $Na^+$ and $K^+$ channels are closed, maintaining the resting membrane potential. **(2) Depolarization:** Stimuli cause initial $Na^+$ influx; if depolarization reaches the threshold, an action potential is triggered. **(3) Rising phase:** Rapid opening of voltage-gated $Na^+$ channels causes a massive $Na^+$ influx, driving the membrane potential to its peak. **(4) Falling phase:** $Na^+$ channels become inactivated while voltage-gated $K^+$ channels open, releasing $K^+$ to repolarize the cell. **(5) Undershoot:** Slow closure of $K^+$ channels causes a brief hyperpolarization below baseline before returning to the resting state (1).

### Phase 2: Propagation Down the Axon

Action potentials travel down the axon toward the terminal in one direction only. This strict unidirectional flow is governed by refractory periods (unresponsive recovery phases of the membrane). During the absolute refractory period, inactivated voltage-gated $Na^+$ channels physically cannot reopen, preventing the signal from traveling backward. This period is approximately 2 ms in typical mammalian neurons [19]. During the subsequent relative refractory period, as $Na^+$

channels reset, a new action potential can be triggered, but only by an unusually strong input that overcomes ongoing hyperpolarization [18, 17].

**Phase 3: Synaptic Transmission**

While fast electrical coupling occurs in some cells, communication across the junction is predominantly chemical. In chemical transmissions, when the action potential arrives at the axon terminal, the voltage change opens calcium ($Ca^{2+}$) channels. The entry of $Ca^{2+}$ triggers the release of neurotransmitters across the synaptic gap to neighbouring cells [18, 17].

In the nervous system, information is not coded by the size of the action potential, but by its frequency and pattern. These firing rates vary significantly by cell type. While sensory neurons may generate 1-20 action potential pulses per second, certain hippocampal neurons can reach up to 300 pulses per second [20]. Maintaining the stability of these electrophysiological pathways is vital, as their impairment is a primary driver in the development of various severe neurological pathologies. For instance, epilepsy is often linked to mutations in ion channels causing hyperexcitability, multiple sclerosis is caused by impaired signal propagation due to the disruption of the myelin sheath and chronic pain results from altered function in sensory neuron ion channels [18, 17].

### 2.1.2 Electrophysiology monitoring

Electrophysiology measures cellular electrical activity to investigate fundamental nervous system function and advance medical technologies. When ions flow across neuronal membranes during synaptic activity and action potentials, they generate transient electrical currents in the surrounding extracellular space. An electrophysiological recording system captures these macroscopic ionic events across different spatial scales, using frequency filtering to separate the raw voltage data into two distinct signal streams: action potentials (AP) and local field potentials (LFPs) [1, 15].

**Spikes:** High-pass filtering the broadband signal (typically above 300 Hz) isolates sharp action potentials from neurons adjacent to the electrode site, defined by parameters such as amplitude, duration and firing rate. Depending on electrode proximity and waveform sorting, these high-frequency transients represent either single-unit activity (SUA) from an isolated individual neuron or multi-unit activity (MUA) from a small local cluster of nearby cells [1].

**Local Field Potentials (LFPs):** Low-pass filtering below 300 Hz isolates the aggregate synchronized electrophysiological dynamics of local neural populations. Dominated by dendritic synaptic inputs, LFPs also integrate intrinsic membrane oscillations, hyperpolarization currents and nearby action potentials to provide critical insights into network-level behaviors [1].

### Electrophysiological Recording Modalities

To study brain activity, neuroscientists utilize a wide range of non-invasive, semi-invasive and optical recording techniques characterized by varying levels of invasiveness, spatial resolution and temporal resolution. Among non-invasive techniques, electroencephalography (EEG) records synchronous electrical signals produced by millions of neurons via scalp electrodes; magnetoencephalography uses highly sensitive magnetometers in shielded facilities to detect weak magnetic fields from cortical currents, offering better spatial and temporal resolution than EEG; functional near-infrared spectroscopy (fNIRS) detects hemodynamic changes associated with cerebral blood flow using near-infrared light; and functional magnetic resonance imaging (fMRI) measures metabolic brain activity by tracking blood oxygen level-dependent signals across the whole brain [21, 22].

In a semi-invasive approach, electrocorticography requires a craniotomy to place electrode grids directly on the cortical surface (epidurally or subdurally), providing superior spatial and temporal resolution compared to EEG, though it cannot reliably isolate individual single-neuron spikes. Alternatively, endovascular electrodes (Stentrodes) are threaded through the vascular system into cortical blood vessels, capturing regional neural signals without penetrating brain tissue or breaching the blood-brain barrier. Finally, optical recording represents a minimally invasive approach that monitors fluorescent indicators within neurons using high-resolution laser-scanning microscopy [21, 22].

To capture electrical activity closest to its physical source, intracranial recording methods insert microelectrodes directly into brain tissue to record extracellular action potentials (spikes) from individual neurons or LFPs from localized populations. Passive intracranial architectures lack integrated electronic signal-conditioning circuitry on the probe substrate itself, merely routing raw unamplified analog signals directly to external amplification and digitizing hardware via physical wires [21, 22]. Representative passive architectures include:

**Microwire Arrays and Recording Cubes:** Consist of a grid of guiding tubes containing fine metal microwires lowered into tissue. They offer chronic longevity (over 7 years in primates) and high yields (up to 1000 sensors per volumetric cube), but require major neu-

rosurgery, slow multi-day insertion to prevent tissue trauma and suffer gradual signal deterioration from glial encapsulation [21].

**Utah Array [23]:** A rigid silicon arrangement of 100 microelectrode needles inserted via a high-speed pneumatic impact gun. It is currently approved by the FDA for human clinical trials, but micromovements induce local lesions and chronic inflammation. Glial encapsulation typically limits single-unit spike isolation to weeks or months and it is anatomically restricted to flat cortical surfaces up to 1.5 mm deep [23].

**Michigan Probe (NeuroNexus Array):** Features a planar silicon architecture with multiple active recording sites positioned sequentially along a single shank, enabling simultaneous recording across cortical or hippocampal layers. However, signal quality typically deteriorates rapidly after the initial post-operative days [21].

**Neurotrophic Electrode:** Consists of a hollow glass cone containing insulated gold wires preloaded with localized nerve growth factors to encourage neurites to grow directly into the cone. It offers exceptional biocompatibility and chronic stability (over 4 years in humans), but suffers from low channel counts, low yield and difficulty verifying the exact physical nature of recorded potentials [24].

## 2.2 CMOS-based neural probes

Integrating active complementary metal-oxide-semiconductor (CMOS) microelectronics directly onto implantable substrates overcomes the physical routing bottlenecks of passive neural probes [25, 26]. This section explores the manufacturing workflows, circuit architectures, structural scaling limits and biocompatible material innovations that define modern high-density CMOS neural recording platforms.

### 2.2.1 The Evolution Towards Active CMOS Neural Probes

Resolving single-unit spikes alongside local field potentials with high spatiotemporal resolution requires physical sensors inserted directly into the cerebral cortex. Scaling channel density is crucial because high-density recordings allow researchers to track thousands of individual neurons simultaneously across multiple cortical layers, providing the data required to decode complex neural circuits and advance high-performance brain-computer interfaces [27, 26]. As introduced in Section 2.1.2, early intracortical probes route each recording site's signal off-probe via a dedicated metallic trace to external amplification and digitization hardware. This passive approach hits a physical "routing bottleneck": because every additional recording site requires

its own dedicated trace, increasing channel count drastically expands the physical width and thickness of the probe shank. This footprint expansion causes severe mechanical damage to brain tissue, strictly capping the maximum number of sensors that can safely fit on a single implant [25, 26].

The concept of embedding active signal-conditioning circuitry directly onto the probe shank dates back to the mid-1980s and 1990s, when Kensall Wise's group at the University of Michigan first demonstrated silicon probes with on-chip amplification and multiplexing [28], later extending this to fully integrated CMOS multiplexed recording interfaces [29, 30]. This active paradigm was subsequently validated on planar in vitro multi-electrode arrays (MEAs), proving that dense 2D microchip grids could locally amplify, record and stimulate cultured neural populations [26].

Translating these active microelectronics to penetrating in vivo shanks marked a decisive technological inflection point. What distinguishes modern high-density platforms, such as SiNAPS [4] and Neuropixels [31], from earlier active probes is not the presence of on-chip circuitry itself, but the massive scale at which it is implemented. Advances in standard complementary metal-oxide-semiconductor (CMOS) foundry processes have driven exponential increases in transistor density, enabling thousands of independent amplification, multiplexing and digitization channels to be integrated within sub-millimeter shank footprints, a scale unachievable with earlier microfabrication techniques [27, 26, 4, 32]. By processing signals locally at the recording site, these modern active probes bypass traditional cabling bottlenecks, eliminate noise accumulation and unlock high-channel-count volumetric recording in living brain tissue [26].

### 2.2.2 The CMOS technology

Complementary Metal-Oxide-Semiconductor (CMOS) technology is the foundation for manufacturing modern microchips by pairing n-type (negative charge carriers) and p-type (positive charge carriers) transistors on a single silicon chip to control current flow efficiently. To keep up with growing power and size demands, CMOS design has evolved from flat 2D transistors to 3D nano-scale structures and specialized channel materials. These advanced designs suppress unwanted current leakage and reduce power draw. Meanwhile, sub-10 nm nodes utilize Extreme Ultraviolet (EUV) light, an ultra-short wavelength light source that acts like a microscopic stencil to print extremely fine circuit patterns, along with atomic-layer techniques to build precise nanoscale features without damaging the silicon lattice [26].

Manufacturing active CMOS neural probes for brain insertion requires a two-stage workflow combining standard commercial foundry chip fabrication with biomedical post-processing. In the first phase, foundries build the probe's underlying circuitry and signal processing systems

using established chip nodes, placing circuitry either directly along the shank beneath recording sites or concentrating it outside neural tissue via shank-embedded switch arrays [25, 3]. In the second phase, wafer-scale post-processing coats toxic exposed metals with inert noble materials (like platinum or titanium nitride) and seals them with protective moisture barriers to improve signal quality and prevent corrosion [25, 3]. Finally, the bulk silicon substrate is ground down to a 15 - 30 $\mu$m thickness and precision-etched into slender shanks engineered to minimize tissue damage upon chronic brain insertion [25, 3].

### 2.2.3 Comparative Analysis of State-of-the-art Platforms

To contextualize advances in active architectures, Table 2.1 summarizes the key structural, manufacturing and functional parameters of representative state-of-the-art CMOS neural probes.

The choice of CMOS technology node directly correlates with shank width across these selected platforms. The 0.13 $\mu$m nodes used by Neuropixels and NeuroSeeker enable the narrowest shafts (70 - 100 $\mu$m), while the SiNAPS and Lopez et al. platforms rely on the coarser 0.18 $\mu$m node, and the earliest design, the EDC probe, required the widest shaft (160 $\mu$m) at 0.6 $\mu$m [31, 33, 4, 32, 34, 3, 35]. Within the SiNAPS family specifically, the transition from the first-generation design to the 8-shank architecture achieved a narrower shank (88 $\mu$m vs. 120 $\mu$m) despite an increase in thickness (30 $\mu$m to 50 $\mu$m) [4].

Table 2.1: Comparative overview of state-of-the-art active CMOS neural probe platforms.

| **Probe / Platform** | **Arch.** | **Shaft Dim. (L $\times$ W $\times$ T) (mm $\times \mu$m $\times \mu$m)** | **Tech Node / Electrode** | **Channels (Tot. / Sim.)** | **RMS Noise (AP / LFP)** | **Power / Pixel** |
|---|---|---|---|---|---|---|
| SiNAPS 1st Gen (2019) [32] | APS | $6.5 \times 120 \times 30$ mm $\times \mu$m $\times \mu$m | 0.18 $\mu$m Pt | 512/512 | 7.5 $\mu$V 24.5 $\mu$V | $< 6$ $\mu$W |
| SiNAPS 8-Shank (2025) [4] | APS | $5.07 \times 88 \times 50$ mm $\times \mu$m $\times \mu$m | 0.18 $\mu$m Pt | 1024/1024 | 6.67 $\mu$V 16.45 $\mu$V | 6 $\mu$W |
| Neuropixels (2019) [31] | ASM | $10 \times 70 \times 20$ mm $\times \mu$m $\times \mu$m | 0.13 $\mu$m TiN | 960/384 | 5 $\mu$V 9 $\mu$V | $< 50$ $\mu$W |
| NeuroSeeker (2018) [33] | APS | $8.15 \times 100 \times 50$ mm $\times \mu$m $\times \mu$m | 0.13 $\mu$m TiN | 1344/1344 | 31 $\mu$V 125 $\mu$V | 63 $\mu$W |
| Lopez et al. (2013) [34] | ASM | $10 \times 100 \times 50$ mm $\times \mu$m $\times \mu$m | 0.18 $\mu$m Pt/Au/IrOx | 455/52 | 3.2 $\mu$V (all band) | 18.2 $\mu$W |
| Torfs et al. (EDC) (2009) [35] | ASM | $8 \times 160 \times 50$–$100$ mm $\times \mu$m $\times \mu$m | 0.6 $\mu$m Pt | 1028/8 (per shaft) | N/A | $< 7.2$ $\mu$W |

### 2.2.4 Architectural Evolution

To overcome the routing constraints of passive arrays, modern active CMOS probes utilize two primary structural paradigms to select and process large neural data arrays: Active Switching Matrix (ASM) architectures (exemplified by the Neuropixels platform [31]) and Active Pixel Sensor (APS) architectures (exemplified by the SiNAPS platform [4]) [26].

In ASM designs like Neuropixels, an active switch array beneath the electrodes dynamically connects a configurable subset of dense sensing sites (e.g., 384 active channels selected out of 960 total sites) to base-level amplifiers [26]. While this approach minimizes the circuit footprint required directly on the shank, placing amplification at the base limits the number of channels that can record simultaneously and increases the overall base area.

In contrast, the APS approach integrates a miniaturized front-end circuit directly beneath every individual electrode site on the shank [32, 27]. Each pixel contains a dedicated amplifier, a low-pass filter and an active DC feedback loop that cancels electrode-tissue offsets in real time [32]. Unlike traditional architectures (such as Neuropixels) that rely on bulky AC-coupling capacitors to block DC offsets, this fully DC-coupled design eliminates integrated capacitors entirely, permitting continuous, unconstrained full-array recording across thousands of active sites simultaneously within a minimal shank footprint.

To route these high-density data streams off-chip, both ASM and APS topologies employ time-division multiplexing. While APS probes typically multiplex analog signals on-chip before off-chip digitization, architectures like Neuropixels integrate on-chip analog-to-digital converters directly on the probe base, digitizing signals locally before streaming serialized digital data off-chip [36, 25, 3].

A major milestone in the evolution of both ASM and APS platforms is their integration into real-time electrophysiology systems, interfacing active CMOS front-ends directly with System-on-Chip back-ends—a crucial step toward fully autonomous brain-machine interfaces [27].

### 2.2.5 Structural Evolution

The structural evolution of active CMOS neural probes is defined by a dual-track progression: optimizing the physical cross-section of the implantable shank to limit tissue displacement, while leveraging advanced 3D semiconductor scaling to pack more control electronics into that constrained footprint [3].

Unlike ultrathin polymer or passive thread probes, active CMOS probes are inherently constrained in their minimum size because they must house integrated circuitry on-chip. Typical active silicon shanks are engineered with high-aspect-ratio geometries, typically around 70 - 120 $\mu$m in width and several millimeters in length [32, 3]. To mitigate acute tissue trauma

and chronic Foreign Body Reaction (FBR) caused by inserting these rigid structures into soft brain tissue, modern active probes undergo post-fabrication thinning via mechanical grinding to shank thicknesses down to 25 - 30 $\mu$m [4], or silicon-on-insulator microfabrication to reach thicknesses as low as 15 $\mu$m [31]. While the silicon shank remains mechanically rigid compared to neural tissue, reducing its cross-sectional area minimizes tissue volume displacement and vascular damage during insertion.

Internally, advancing to smaller CMOS technology nodes improves circuit integration density and reduces power consumption, enabling more complex front-end processing and multiplexing logic within a compact shank footprint. Managing power dissipation is critical, as active circuitry operating inside brain parenchyma must keep local power consumption low enough to maintain tissue temperature elevation well below strict physiological limits (typically $< 1\,^{\circ}$C) to avoid cell damage and hyperthermia [26, 27]. However, scaling electrode density without increasing shank width is constrained not by lithographic feature size, but by electrode-tissue impedance, which demands a minimum electrode surface area (typically $\geq 10\,\mu\text{m} \times 10\,\mu\text{m}$) to prevent excessive thermal noise. Because silicon active shanks cannot shrink beyond these limits, alternative engineering strategies are required to minimize tissue trauma and mitigate the chronic FBR [27].

### 2.2.6 Material Evolution

Material innovation is essential to bridge the gap between industrial semiconductor manufacturing and the biological environment of the brain. Standard CMOS foundries utilize aluminum-copper (Al-Cu) alloys for top-metal interconnect layers, which are unsuitable for direct neural contact due to cytotoxicity and electrochemical corrosion. Specialized post-processing workflows instead cap the exposed aluminum pads with noble metals or inert metal nitrides, primarily Platinum (Pt), Gold (Au), Titanium Nitride (TiN) or Iridium Oxide ($IrO_x$) [27]. Pt electrodeposition, for instance, creates a porous high-surface-area micro-texture that lowers electrochemical impedance, suppressing thermal noise to capture microvolt-scale single-unit spikes and local field potentials [32]. Beyond noble metals, nanomaterials such as Carbon Nanotubes (CNTs) and conductive polymers like PEDOT:PSS are applied during post-processing to further enhance signal-to-noise ratio and long-term signal stability [37].

Maintaining long-term electrical coupling and low baseline noise within miniaturized active pixels also requires optimized dielectric encapsulation. Silicon dioxide ($SiO_2$) and silicon nitride ($Si_3N_4$), deposited during foundry processing, are often augmented with post-fabrication thin films, such as atomic layer deposition of Hafnium Oxide ($HfO_2$) or Aluminum Oxide ($Al_2O_3$),

to prevent ion penetration and moisture-induced leakage current over chronic implantation windows [26, 27].

Finally, to mitigate the Foreign Body Reaction (FBR) triggered by inserting rigid silicon probes into soft brain tissue, state-of-the-art CMOS arrays incorporate specialized surface functionalization, such as immune-stealthy bio-inert polymer coatings (polyethylene glycol, PEG) or Parylene-C encapsulation, reducing non-specific protein adsorption and physical insertion friction to help preserve long-term tissue viability [27].

## 2.3 Probe Insertion Strategies and Biomechanical Integration

Beyond the material coatings (Section 2.2.6) and shank miniaturization strategies (Section 2.2.5) discussed above, the method used to physically insert a probe into brain tissue represents a further, equally important strategy for mitigating the Foreign Body Reaction (FBR). Far beyond the localized trauma of the surgical procedure itself, the interaction between a rigid device and the highly compliant neural matrix initiates mechanical, biological and spatial challenges that dictate both acute signal quality and long-term interface stability [8].

A variety of insertion strategies and complementary probe designs have been proposed to address this. These range from passive approaches, such as temporary stiffening coatings or dissolvable shuttles that ease the delivery of mechanically compliant probes, to guide-tube-assisted stereotaxic placement, robotic or motorized insertion systems that improve reproducibility and, more recently, active vibration-assisted insertion. Since most of these devices remain tethered to the cranium, chronic implants also experience continuous micromotion as the encapsulated brain tissue shifts with respiratory and cardiovascular cycles, motivating a parallel line of untethered wireless microdevices designed to move in tandem with the surrounding tissue instead [25, 9].

Among tethered approaches, the immediate mechanical bottleneck during insertion is cortical dimpling, where the highly elastic brain tissue compresses and deforms under the probe tip before puncture. The severity of this deformation depends heavily on the cross-sectional geometry of the insertion tool, with larger dimensions driving more severe dimpling and structural damage to the surrounding cortical architecture [8]. Conventional manual insertion compounds this problem, since it lacks the consistent velocity and force needed to navigate varying tissue resistance, introducing erratic speeds and depths that can further exacerbate tissue displacement [9, 8].

To address these limitations, recent methodological paradigms integrate high-frequency or ultrasonic microvibrations, as exemplified by the NeuralGlider Inserter [9]. Rather than relying

on static insertion forces that cause significant tissue dimpling and compressive strain, ultrasonic microvibrations substantially lower peak penetration resistance and surface friction, enabling smooth probe delivery at a microscopic scale. This has been shown to reduce measurable tissue damage by approximately 7.4% compared to standard insertion [9]. When paired with automated velocity control, this approach also eliminates the manual handling errors described above, yielding highly repeatable insertion procedures well suited to delivering thin, flexible or ultrasmall electrodes that would otherwise buckle under static puncture loads [9, 8].

Because severe acute trauma triggers vascular leakage and local edema, the resulting microenvironment directly impacts early electrophysiological recordings. Initial tissue displacement and localized trauma can cause affected neurons to fire with reduced action potential amplitudes, while acute bleeding and edema physically increase the distance between electrode sites and target cell bodies [9]. Conversely, as tissue relaxes and micro-hemorrhages clear, cell bodies settle into closer proximity with the sensor surface. A reduction in insertion-induced mechanical trauma is therefore expected to translate into measurable improvements in acute signal stability [8, 38]. Specifically, insertions using vibration-assisted actuation are expected to yield higher signal-to-noise ratios (SNR) and a greater yield of detected single-unit action potentials compared to non-actuated insertions [8]. For comparison, previous work by Fiáth et al. [38] achieved an average SNR of 9 dB.

Ultimately, pairing optimized insertion mechanics with modern CMOS probe architectures provides a comprehensive pathway toward minimizing the foreign body reaction and securing high-fidelity electrophysiological recordings over extended chronic timelines.

## 2.4 Closed-loop

Beyond physical probe deployment, transitioning from static open-loop setups to adaptive closed-loop architectures offers distinct operational advantages for CMOS neural probes.

In neuroscience and bioengineering, a closed-loop system is a dynamic control process in which a signal is continuously monitored and used to adjust the delivered stimulus, creating a self-regulating loop [6, 2]. Indeed, closed-loop architectures measure the feedback signal, extract relevant features and automatically adapt intervention based on such features ([6, 2]) through a continuous cycle of signal acquisition, decoding, modulation and parameter adjustment [39]. To this end, neuromodulation is the targeted alteration or fine-tuning of physiological neural activity and is increasingly deployed in dynamic closed-loop configurations to achieve adaptive therapy [40].

### 2.4.1 The Limitations of Open-Loop Systems

The fields of electrophysiology and neuromodulation are indispensable for decoding complex neural mechanisms and pathologies [40]. While current neurostimulation devices can manage severe neurological conditions when non-invasive options fail, traditional approaches remain limited by their reliance on open-loop paradigms [40]. These open-loop setups deliver electrical stimulation according to a fixed preset schedule, operating in a "blind" manner without measuring real-time neural responses or dynamically adapting output based on objective physiological feedback [41, 42, 2]. Because spatial, temporal and intensity parameters are set a priori, they remain fixed regardless of the brain's evolving neurophysiological state, intrinsic population dynamics or cell-specific responses during experimental protocols [2]. This lack of real-time adaptation introduces major confounding factors in basic neuroscience. Delivering uncoordinated stimulation independently of underlying brain states yields low specificity when probing neural rhythms, cognitive processes or network dynamics [42]. Consequently, developing sophisticated closed-loop algorithms offers a pivotal opportunity to precisely monitor feedback, target specific neural signatures and better understand the underlying drivers of brain function and network disorders [40].

A major physical and biological challenge comes from ongoing biophysical variability and micro-movements, such as cardio-respiratory cycles or postural shifts, which continuously alter the physical interface between electrodes and target neural tissue [41]. Because current density varies inversely with the square of the distance ($1/d^2$), subtle positional shifts induce large fluctuations in charge delivery [41]. In open-loop setups, this produces unstable activation levels, unrecorded off-target effects and unintended activation of neighboring pathways, which corrupts trial-level neurophysiological recordings and compromises experimental control [41, 43]. Furthermore, static stimulation fails to account for inter-subject variability or baseline neuroplastic drifts over time, often resulting in local tissue reactions (such as glial scarring) that alter impedance and require higher currents to recruit the same neural populations [44].

From an experimental standpoint, static paradigms make it difficult to disentangle evoked neural responses from the intrinsic dynamic fluctuations of the circuit [2]. Researchers are forced to rely on manual empirical parameter adjustments or offline trial-averaging, which can obscure sub-second non-linear circuit interactions [43, 42, 45]. Additionally, continuous pulse delivery and high-bandwidth wireless data streaming drastically increase power consumption and thermal load [44, 46]. In chronic experimental models, rapid battery depletion severely limits the duration of continuous monitoring, increases hardware form-factors and degrades system longevity, ultimately constraining the scope of long-term neuroscientific investigation [47, 46].

### 2.4.2 Biomarkers

The successful implementation of closed-loop neuromodulation for neuroscientific investigation depends on the identification and extraction of biomarkers, objective neurophysiological signatures that capture a system's functional state in real time [6]. These biomarkers act as the functional language of the brain, providing the feedback loop required for an autonomous platform to decode internal state dynamics and adjust experimental inputs [12].

#### Spike-Based Biomarkers

To achieve cellular-level resolution, closed-loop architectures extract feature metrics directly from single-unit activity and multi-unit spiking patterns recorded via high-density microelectrodes [12, 42].

- **Instantaneous Firing Rates:** The instantaneous spike frequency across isolated units provides a direct metric of biological excitation, enabling platforms to detect rapid shifts in local neuronal firing and deliver state-dependent modulation [12, 11].
- **Inter-Spike Intervals:** Quantifying the precise temporal spacing between successive action potentials enables real-time identification of burst activity, single-spike trains and refractory dynamics [43].
- **Spike Synchrony and Phase Locking:** Across multi-channel setups, spike timing relative to neighboring units or local field oscillations reveals population coding and spike timing-dependent plasticity mechanisms [43, 42].

#### Field-Potential and Oscillatory Biomarkers

LFPs reflect the synchronized activity of local neuronal populations, providing continuous low-frequency signals ideal for tracking brain states and executing phase-locked interventions [2].

- **Band Power and Spectral Rhythms:** Quantifying power within specific frequency bands (e.g., theta, beta, gamma) tracks macroscopic state transitions, such as motor intention, sleep stages or pathological hypersynchrony [2, 11].
- **Instantaneous Phase and Cyclic Position:** The instantaneous phase of a target rhythm (signaling the precise position along a wave, such as peak, trough or rising edge) acts as a dynamic indicator of network excitability, allowing stimulation to be timed to specific windows of responsiveness [2].

#### Transient and Evoked Biomarkers

Closed-loop platforms also monitor discrete, event-driven voltage trajectories that signal transient network transitions or direct responses to stimulation.

- **Evoked Compound Action Potentials (ECAPs):** In peripheral nerve and spinal setups, ECAPs measure the sum of synchronous single-fiber action potentials generated in response to an electrical stimulus, serving as a direct readout of real-time fiber recruitment [41].
- **Macroscale Evoked Responses:** Network-level readouts, such as motor evoked potentials or transcranial magnetic stimulation evoked potentials, track systemic cortical dynamics by measuring the latency, waveform geometry and peak-to-peak amplitude of transient responses [2].
- **Spontaneous Transient Anomalies:** Large-amplitude and short-duration voltage anomalies, such as interictal epileptiform discharges, sharp-wave ripples or sudden population bursts, signal rapid shifts in localized network synchrony, enabling immediate intervention before pathological propagation occurs [42].

### 2.4.3 Feature Extraction

To operationalize the neurophysiological biomarkers described above, adaptive neurotechnology relies on feature extraction as the mathematical engine that transforms high-dimensional raw electrophysiological time-series data into low-dimensional control variables. For high-density microelectrode recordings in in vivo animal models and neuroscientific research, this processing pipeline is optimized to compute target features in real time while minimizing latency [12]. Depending on the experimental paradigm, these control metrics range from spatial and network features, to spike detection, sorting and action potential metrics and data-driven latent state estimations.

#### Spatial and Network Feature Extraction

To isolate network interactions while filtering volume-conducted noise, architectures deploy advanced spatial and multi-channel processing. This includes computing frequency-specific coherence metrics to measure cross-regional connectivity and spatial filtering like the Surface-Laplacian to isolate local generators while attenuating artifacts [42, 2]. Macroscopically, real-time functional connectivity algorithms applied to fMRI or fNIRS datasets compute correlation matrices across preserved networks to infer hidden intent [48].

### Spike Detection, Spike Feature Sorting and Action Potential Metrics

Real-time architectures that prioritize spike detection and waveform feature extraction first bandpass-filter raw microelectrode streams (typically $(300-5000$ Hz)) to isolate high-frequency action potentials from low-frequency background activity [12, 4].

- **Real-Time Thresholding and Waveform Detection:** Discrete action potentials are typically isolated from background noise by applying amplitude thresholds, either static or adaptive, or via window discriminators, that check amplitude constraints across a specific time window [12, 43]. Following threshold crossing, candidate spike waveforms are extracted for downstream morphological and feature analysis [43].
- **Feature Clustering and Online Sorting:** To track individual single units and differentiate single-unit action potentials from unresolved multi-unit activity during in vivo animal experiments, low-dimensional feature vectors are extracted from candidate spike shapes [12, 43]. These features include morphological metrics, such as peak-to-trough amplitude (the voltage difference between peak depolarization and trough repolarization), half-width duration and maximum slope. Additionally, transformed features like Principal Component Analysis (PCA) projections or wavelet coefficients can also be obtained [12, 43]. The extracted feature space is then processed by real-time clustering algorithms (such as $k$-means or template matching) to achieve pulse-by-pulse spike sorting [12, 43].
- **Temporal Spike Metrics:** Once sorted, continuous spike trains are transformed into quantitative time-domain variables [43]. The system computes running instantaneous firing rates using small sliding windows or kernel density estimators, tracks inter-spike interval distributions to detect burst firing and calculates multi-channel cross-correlation to measure millisecond-scale synchrony across single-neuron ensembles [43, 42].

### Latent State Estimation and Deep Modeling

Finally, advanced architectures bypass deterministic feature engineering to estimate hidden latent brain states. Cognitive platforms like CLoSES deploy encoder-decoder models to infer attentional focus or memory quality from neural feature matrices [42]. At the cellular level, Long Short-Term Memory (LSTM) networks track non-linear neural signals by leveraging recurrent feedback loops to manage temporal context and infer unobservable biophysical states [6]. When interacting with complex models, LSTMs serve as real-time predictive engines to infer hidden membrane potentials and gating variables, updating stimulators pulse-by-pulse [6]. They also project impending pathological transitions, such as seizure onset or tremor bursts [48].

### 2.4.4 Signal Processing and Artifact Mitigation Pipelines

To provide the uncorrupted inputs required for real-time feature extraction, signal cleaning serves as a foundational step in closed-loop neuromodulation, essential for preserving signal fidelity and preventing stimulation or physiological artifacts from corrupting downstream models [12]. For instance, electrical stimulation couples directly into the conductive path of the tissue and electrode array, while optical stimulation introduces localized photoelectrical artifacts on metal electrodes [49]. Such artifacts can easily saturate recording amplifiers and obscure biomarkers such as single-unit spikes and Local Field Potentials (LFPs) [42, 36, 6]. Addressing this in real time typically requires a combination of hardware design, waveform engineering and digital signal processing.

Standard preprocessing applies band-pass filtering to isolate relevant oscillatory biomarkers [2] and notch filtering to eliminate powerline interference (typically 50 or 60 Hz). To meet strict real-time feedback constraints, these are commonly implemented using causal infinite impulse response architectures, such as Butterworth filters, which provide steep frequency roll-off while minimizing latency and phase delay [42]. When applied to isolate the action potential band, this targeted frequency rejection substantially reduces total signal power and root-mean-square (RMS) noise by suppressing high-amplitude low-frequency oscillations and out-of-band interference, bringing baseline RMS values down to $\sim 5\ \mu V_{RMS}$ [31]. Additionally, Gaussian smoothing is often applied to reduce rapid noise fluctuations [12]. Although these methods provide a baseline layer of signal enhancement, traditional filtering alone is generally insufficient to eliminate large stimulation-locked artifacts.

#### Hardware and Spatial Design Strategies

Minimizing the artifact's footprint before it reaches the digital stage is often the most effective protection against amplifier saturation. Bipolar sensing arrangements exploit common-mode rejection, cancelling a stimulation signal that appears nearly identical at both electrodes [42], an idea extended by symmetric tripolar montages, which position the recording anode and cathode around the stimulation contact, aided by segmented or directional electrodes [50]. Physically separating recording and stimulation sites, for instance recording cortically while stimulating subcortically, further reduces coupled current amplitude [42, 50].

#### Temporal Control and Waveform Engineering

Controlling stimulation timing and pulse shape allows systems to bypass the artifact window, eliminating the need for retrospective post-processing. A common approach is a temporal "blanking window" during and immediately after the pulse, implemented via analog switches,

such as those in active pixel sensor arrays, that physically disconnect recording channels during stimulation [42, 36, 39]. Time-division multiplexing achieves a similar effect interleaving recording and stimulation phases into distinct temporal windows [36]. Waveform shape matters as well, since sharp-edged square pulses contain high-frequency components that contaminate broad neural bands [50]. In contrast, smooth pulse shapes, such as Gaussian or sinusoidal pulses, produce artifacts that are much easier to isolate and subtract [39], while charge-balanced biphasic waveforms limit net charge accumulation and reduce lingering electrochemical tail artifacts [51].

#### Digital Removal of Residual Artifacts

When the artifact cannot be avoided physically, adaptive noise cancellation and state estimation models separate stimulation-induced interference from endogenous signals, a key strategy for preventing massive stimulation transients from saturating recording channels [39]. Modern implementations integrate these adaptive filters with Recurrent Neural Networks (RNNs) trained to learn artifact profiles or deploy Kalman and Unscented Kalman Filters to track true neuronal trajectories despite interface interference [39, 52, 44]. For single-unit spike detection specifically, a brief transient can be excised via a narrow (e.g., 3 ms) window around the pulse and reconstructed by interpolation [53, 54], while template subtraction routines strip away known stimulation morphologies directly from the data stream [53, 54, 11]. At the array level, spatial filters such as the Surface-Laplacian or common average referencing (CAR) remove artifacts shared across channels [2], with an analogous role played in optical recordings by matrix factorization methods, which subtract background fluorescence and light artifacts from true calcium transients [55].

The most advanced systems go further, using embedded AI (RNNs or LSTMs trained on the tissue-electrode interface's non-linear transfer function) running on low-latency edge hardware to predict and subtract the artifact pulse-by-pulse in real time [39].

### 2.4.5 Control Theory and Feedback Loops

In a closed-loop neural interface, the recorded signal is continuously monitored and processed to dynamically decide when and how to intervene. Depending on the experimental or clinical application, this decision can be driven by different control principles. The controller acts as the decision-making core, determining the timing, intensity and, where relevant, spatial distribution of the actuation output by applying one of several algorithmic strategies. These range from simple reactive switches to sophisticated model-based optimization frameworks capable of anticipating neural state transitions [6, 2].

### On/Off Thresholding

The most straightforward control strategy is threshold-based triggering. In a single-threshold policy, an action is triggered as soon as a monitored signal deviates past a critical baseline limit. Depending on the application, the system can be configured to respond to positive or negative deviations, for instance, when a signal's amplitude crosses above (or, in the case of extracellular spike detection, below, since action potentials manifest as negative deflections) a calibrated boundary [45, 42]. In spike-driven architectures, this is commonly implemented as event-driven detection, where an action is initiated upon detecting discrete action potentials or rapid increases in local spike rates [45, 42].

A primary challenge of simple thresholding is "trigger jitter" or high-frequency chattering, where noise or rapid bursts cause repeated unwanted re-triggering. To stabilize execution, spike-based architectures typically implement temporal refractory rules (such as a fixed dead-time window following a trigger event), ensuring that subsequent events within a narrow time frame are ignored [45]. Alternatively, when tracking continuous or smoothed control metrics (such as windowed firing rates or band power), systems often utilize dual-threshold hysteresis. In this framework, state transitions occur across separate upper activation and lower reset boundaries, creating a state-dependent memory effect that prevents chattering without needing explicit temporal lockouts [45, 53].

### Proportional-Integral-Derivative (PID) Control

For applications requiring granular and continuous adjustments, PID control is the established engineering standard [41]. The algorithm modulates stimulation intensity to minimize discrepancies between measured neural responses (like ECAP amplitude) and target values [41, 56]. Its output combines three terms: proportional (P) for immediate error response, integral (I) for accumulating past errors to eliminate steady-state offsets and derivative (D) for anticipating future error trends [57, 56]. Additionally, PID Neural Networks leverage backpropagation to dynamically update gain coefficients in real time, improving robustness, energy efficiency and adaptability to physiological drift over static PID approaches [57].

### Fuzzy Logic

Fuzzy logic control provides an intelligent alternative when the neural environment is too complex or nonlinear to be captured by a precise mathematical model [53, 54]. Instead of relying on thresholds, fuzzy inference systems map continuous inputs into overlapping membership functions and process them using expert-defined linguistic, for example, "if tremor is high and

medication is low" [57, 53]. These fuzzy sets are subsequently mapped back to precise numerical actuation values through defuzzification algorithms [53].

### 2.4.6 Neuro-Actuation and Stimulation Modalities

To establish effective closed-loop neural interfaces, implantable probes must integrate actuation mechanisms capable of delivering targeted physical, chemical or optical energy directly into brain tissue [6, 39]. While several non-invasive stimulation modalities exist, implantable platforms strictly rely on micro-scale actuation mechanisms integrated directly onto the probe shank to achieve high spatial resolution and localized therapeutic action. There are four primary tissue-facing actuation mechanisms:

- **Electrical Stimulation:** The most established and widely used approach, delivering microcurrent pulses via probe contacts to alter membrane potentials and modulate neural firing [12, 43, 44, 41]. However, high-voltage electrical stimulation generates substantial artifacts that saturate co-located recording channels [45].
- **Infrared Neural Stimulation:** An optical approach utilizing coherent infrared light delivered via micro-optical fibers to elicit spatially constrained thermal-excitatory neural responses without requiring genetic modification or introducing electrical recording artifacts [5].
- **On-Demand Chemical Delivery:** Integrates microfluidic channels or subcutaneous drug capsules alongside electrical probes to provide targeted responsive pharmacological stabilization in bimodal probe architectures [37].
- **Optical Stimulation:** The primary high-precision method for cell-type-specific neural perturbation, enabling microsecond-temporal control over targeted neural populations [39, 49], detailed below.

#### Optogenetic Actuation

Optogenetics combines genetic engineering with probe-integrated light delivery to control targeted neural populations with millisecond temporal precision [39, 49]. Target cells express light-sensitive microbial proteins, known as opsins. Illuminating excitatory opsins induces membrane depolarization and triggers action potentials [39, 49]. Conversely, red-shifted inhibitory opsins drive hyperpolarizing photocurrents that reliably suppress action potential generation [55].

Traditionally, light is delivered using optical fibers coupled to external lasers or LEDs. However, these fibers are bulky, restrict illumination to broad tissue volumes and lack the spatial resolution required to target individual cortical layers. To overcome these spatial limits, micro-LEDs (light-emitting diodes) can be directly integrated onto active CMOS probe shanks alongside recording sites [42]. When combined with real-time decoding pipelines, these integrated micro-LEDs enable cell-type-specific closed-loop control with minimal electrical artifact contamination, suppressing hypersynchronized epileptic seizure activity [**?**] or executing phase-locked optical inhibition to inhibit pathological oscillations before they manifest [58, **?**]. While translational clinical application remains constrained by viral vector safety concerns and localized thermal dissipation from micro-LEDs [**?**, **?**, **?**], integrated optoelectronic shanks represent the premier paradigm for high-resolution multi-site functional actuation.

### 2.4.7 Integrated Sensing, Stimulation and Closed-Loop CMOS Architectures

To enable real-time feedback in closed-loop neuromodulation, modern neural probes are increasingly designed to integrate both high-density recording and stimulation onto a single micro-array.

When integrated with high-speed digital logic, active CMOS probes form the foundation for low-latency autonomous closed-loop systems. In fundamental neuroscience, high-density arrays such as Neuropixels [31] stream broadband signals into real-time pipelines to trigger activity-dependent feedback and probe neural plasticity [31]. For optical neuromodulation, specialized multifunctional CMOS shanks incorporate micro-LEDs directly alongside recording pixels, creating self-contained optogenetic platforms capable of detecting and suppressing hypersynchronized epileptic seizures in real time [59].

Finally, to eliminate the latency and bandwidth bottlenecks of off-chip transmission, next-generation implants couple active probes directly with on-chip processing elements, including compute-in-memory architectures [56], low-power microcontrollers [44] and neuromorphic co-processors such as Loihi [60]. By executing real-time feature extraction and spike decoding directly at the neural interface, these integrated platforms enable artificial functional bridges across injured neural pathways with microsecond-latency feedback loop dynamics.

### 2.4.8 Development and Applications

While integrated CMOS architectures establish the physical substrate for real-time sensing, processing and actuation, deploying these systems relies on a translational pipeline that runs from theoretical computational validation to deployed clinical devices. This trajectory typically progresses from in-silico modelling, through in-vitro and cellular-level validation, to in-vivo

animal studies and finally into human research, established clinical therapeutics and ambulatory or wearable platforms. Reviewing this progression highlights how signal acquisition, feature extraction and control strategies are adapted at each stage, with a particular focus here on systems built around extracellular activity sensing via implantable probes.

### In-Silico and Theoretical Applications

Computational platforms are used to validate control algorithms before biological implementation, often through hardware-in-the-loop systems that combine general-purpose processors for modelling neural dynamics with dedicated signal processing hardware for real-time control [45, 52]. These platforms allow direct comparison of different control strategies, from simple thresholding and PID control to more adaptive approaches such as fuzzy logic, reinforcement learning or Kalman filtering, under controlled and reproducible conditions before any biological variability is introduced [45, 52, 56]. Their main value lies in providing a transparent and biologically interpretable testbed, which simplifies the verification and risk assessment later required for clinical translation [45, 52, 56, 12].

### Single-Unit Applications

Operating closed-loop systems at single-neuron resolution shifts modulation from broad network dynamics to individual action potentials and subthreshold synaptic events [6, 61]. Key application domains for single-unit closed-loop systems include:

- **Selective Neuromodulation and Targeted Stimulation:** Utilizing real-time computational spike sorting on multi-electrode arrays to drive fast selective optogenetic or electrical stimulation of specific target neurons [6].
- **Microcircuit Synchrony Modulation:** Operantly reinforcing firing synchrony within small localized clusters of neighboring neocortical neurons to investigate fine-scale circuit dynamics and functional connectivity [6].
- **Single-Cell Synaptic Plasticity Research:** Leveraging patch-clamp techniques, which inherently record signals directly from individual cells without needing computational spike sorting, to detect subthreshold events (such as excitatory postsynaptic currents or high-frequency bursts) and deliver contingent feedback with minimal latency ($\sim$ 50 ms) to drive and study cellular-level activity-dependent plasticity [61].

Together, these single-unit applications serve as a core platform in real-time neurofeedback frameworks to probe cellular learning mechanisms and fine-grained neural control [6, 61].

### Pre-Clinical Animal Models

Animal studies remain the standard for validating closed-loop hardware and control strategies under realistic biological conditions. Several implementations combine real-time extracellular or local field potential monitoring with fast electrical or optical intervention. Examples include systems detecting seizure activity to trigger electrical or pharmacological suppression [37], closed-loop deep brain stimulation systems predicting pathological oscillatory events to deliver targeted stimulation bursts [62] and closed-loop optogenetic platforms that monitor local field potentials or circuit-specific activity to drive targeted light-based intervention [55, 12].

### Clinical Implementation

In human clinical research and commercial therapeutics, closed-loop systems leverage both chronic fully implanted neurostimulators and temporary diagnostic recording arrays to execute real-time feature extraction and responsive stimulation [42]. These platforms generally operate by either triggering intervention when a monitored neural feature crosses a defined threshold or employing state-space models to continuously estimate latent brain states [42]. Transitioning from static open-loop protocols to on-demand responsive delivery significantly reduces side-effect profiles and extends device battery longevity [42, 46].

For instance, the NeuroPace Responsive Neurostimulation System [63], the first FDA-approved responsive closed-loop stimulator for drug-resistant epilepsy, monitors patient-specific electrocorticographic signatures to deliver targeted electrical bursts prior to seizure onset [42, 46]. In movement disorders, sensing-enabled deep brain stimulation (DBS) devices, such as Medtronic's Percept [64] and Newronika's AlphaDBS [65], continuously track pathological LFP biomarkers to dynamically adjust stimulation voltage, to suppress motor symptoms like rigidity and tremor in real time [50, 6, 46, 66]. Similarly, closed-loop spinal cord stimulation platforms, such as the Saluda Evoke system [41], monitor evoked compound action potentials (ECAPs) to maintain a consistent neural dose during postural movement, ensuring stable pain relief [41, 56]. Furthermore, responsive vagus nerve stimulators rely on peripheral autonomic proxies to trigger intervention, targeting systemic neuro-cardiac stabilization and seizure disruption [39, 37, 46]. Together, these commercial platforms illustrate a widespread clinical shift toward personalized adaptive neuromodulation validated across large-scale trials [12, 39, 11].

### Non-Invasive and Wearable Applications

More recently, closed-loop neuromodulation has extended into non-invasive ambulatory devices, generally relying on peripheral physiological signals (EMG, cardiovascular markers)

rather than direct neural recording to trigger stimulation, for applications such as bladder control, acute stress mitigation and sleep-related autonomic regulation [67, 51, 57]. While outside the scope of extracellular probe-based sensing, these examples illustrate the broader translational endpoint of the closed-loop paradigm discussed above.

### 2.4.9 Current Challenges

Despite the clinical and commercial success of modern adaptive platforms, the transition to fully autonomous high-density closed-loop interfaces remains hindered by significant biophysical and computational challenges.

A critical technical challenge is the residual management of stimulation artifacts. While the hardware gating and software cancellation routines described in Section 2.4.4 offer substantial attenuation, fully neutralizing these transients remains difficult in high-density configurations, largely because target biomarkers (single-unit spikes or local field potentials) are measured in microvolts, while the stimulation delivered nearby is several orders of magnitude larger, in volts [41]. This high voltage can saturate recording amplifiers, blinding the system during a window that is critical for real-time feedback and state estimation [41, 6, 2]. This is not limited to invasive systems. Moreover, the hardware and software strategies used to avoid these artifacts, such as blanking windows or time-division multiplexing, introduce their own trade-off: the resulting "decoding vacuum" during and after each pulse limits temporal resolution and can cause the system to miss fast neural dynamics [42, 6, 36]. Poorly filtered or residual artifacts also carry an operational risk, since they can be misidentified as genuine biomarkers and trigger unintended stimulation cycles [6].

Beyond artifact management, closed-loop systems also face a computing architecture bottleneck. General-purpose operating systems (Windows, macOS, Linux) rely on schedulers that lack the deterministic low-latency control that real-time closed-loop processing requires, which is why dedicated hardware implementations are generally favoured over software-only solutions for maintaining consistent timing. Analog circuits historically addressed this latency problem but at the cost of limited parameter flexibility. More recent platforms, such as the CorTec Brain Interchange System [68] and modern adaptive DBS frameworks [50], illustrate continued progress in this direction, but their reliance on specialized closed-source hardware limits scalability and constrains algorithmic experimentation, representing a further barrier to wider academic and clinical adoption of next-generation closed-loop neurotechnology [40].

# Chapter 3

# Materials and Methods

This chapter provides an overview of the materials and the methodology, as well as an explanation of adopted resources and strategies for the development of this dissertation work.

## 3.1 AI models

In accordance with institutional academic integrity guidelines, Artificial Intelligence (AI) platforms and Large Language Models (LLM) were utilized strictly as assistive and productivity-enhancing tools during the development of this dissertation. The core scientific contributions, conceptual design of the code pipelines and final textual interpretations remain entirely the work of the author. The specific deployment of these technologies was partitioned into two distinct operational workflows:

1. **Literature Synthesis and Document Analysis:** During the initial literature review phase, peer-reviewed source material was systematically gathered from academic databases, including Scopus and ScienceDirect. To streamline the extraction of high-density technical parameters, the compiled literature was uploaded to Google NotebookLM. This platform was used exclusively to index the source material, generate semantic cross-references and efficiently locate specific biophysical metrics within the author's retrieved reference library.

2. **Code Optimization and Algorithmic Support:** Google Gemini was utilized as a debugging and code-optimization assistant during the implementation of the digital signal processing and analysis algorithms. The model was leveraged to refine python scripts and

suggest syntactic fixes for runtime errors. All AI-generated code optimizations were manually verified and integrated by the author to ensure algorithmic correctness and system stability.

No generative AI tools were utilized to author the scientific text, formulate independent hypotheses or interpret empirical data.

## 3.2 SiNAPS probe

The SiNAPS (Simultaneous Neural Active Pixel Sensor) neural probe [32] served as the recording platform for all in vitro and in vivo datasets. Manufactured using a standard 0.18 $\mu$m CMOS process, SiNAPS probes integrate signal processing circuits directly beneath each electrode (Active Pixel Sensor concept). This modular design features 32-pixel modules on an 8-shank configuration (1024 channels, $< 30$ $\mu$m pitch, 4.12 mm$^2$ coverage). To minimize tissue damage, shanks are thinned to 30 - 50 $\mu$m [4, 32].

Each pixel integrates front-end local amplification and filtering, preventing electromagnetic interference and crosstalk. The circuit includes a DC-coupled low-noise amplifier, a first-order low-pass filter, a shared module-level active feedback loop for DC offset cancellation and an NMOS analog switch for direct electrode access. Electrode pads were electrodeposited with platinum (Pt) to lower impedance from $\sim$ 14 M$\Omega$ to 500 - 650 k$\Omega$ at 1 kHz. Signals are read sequentially via time-division multiplexing (TDM) and 12-bit ADC conversion is performed off-probe [4, 32].

Key technical performance metrics include an in-pixel gain of $\sim$ 45.4 dB, a sampling frequency of 20 kHz per channel and an input noise of $\sim$ 7.5 $\mu V_{RMS}$ (300 Hz - 7.5 kHz). Power consumption remains under 6 $\mu$W per pixel to avoid tissue heating [32].

## 3.3 NeuralGlider

The NeuralGlider Inserter [9] was used to insert the SiNAPS probe into the mouse cortex for in vivo experiments. It is a commercially available system that leverages ultrasonic microvibrations to enable precise, minimally invasive electrode implantation into neural tissue. While initially developed to lower insertion forces and prevent cortical dimpling during the deployment of flexible polymer arrays, which lack the structural stiffness to breach the pia mater without buckling, its scope has expanded to needle-based brain puncturing for delivering wireless neural microdevices. By utilizing ultrasonic micro-vibrations to temporarily reduce surface

friction and tissue drag, this approach enables ultra-low-force penetration across diverse substrate geometries, extending its utility to delicate rigid shuttles and injectors where minimizing tissue trauma and trajectory drift is paramount [9].

For the experiments, two main operational parameters were set on the NeuralGlider platform: a vibration intensity power of 1 W and an insertion speed of 20 $\mu$m/s.

## 3.4 Electrophysiological Data

Electrophysiology experiments were performed using the SiNAPS Research Platform, which included a custom data acquisition board, a digital control unit, a custom coax cable assembly with embedded electronics and data acquisition software running on a PC equipped with a Camera Link frame grabber. The digital control unit was built on an Opal Kelly XEM7310 integration module based on an Altera Cyclone IV FPGA. This hosted the SiNAPS IP core, which provided real-time hardware bandpass filter (second-order FIR filters with a high-pass cutoff of 300 Hz and a low-pass cutoff frequency of 5 kHz). Signals were acquired at a sampling frequency of 20 kS/s, with 12-bit resolution over a range of 2.4 V. The PC ran data acquisition software (SiNAPS Control Software, IIT, Italy) for data visualization and storage. This was developed in Microsoft Visual Studio Enterprise 2019 using the general-purpose high-level C# programming language and the .NET Framework 4.8.1.

### 3.4.1 Synthetic Data

A synthetic 100 s long extracellular recording was simulated using a neural signal simulator (Multichannel Systems MEA Signal Generator (MEA-SG)) and acquired through the lab-developed SiNAPS CMOS amplification front-end at a sampling frequency of 20 kS/s. This setup ensured identical recording and front-end signal processing parameters to actual in vivo experiments while maintaining precise ground-truth knowledge of the expected signals. The simulated signal modelled single-unit action potentials generated by two distinct active neurons with different spike features under controlled noise conditions.

### 3.4.2 In vitro Data

In vitro recordings were acquired in collaboration with the research group led by Prof. Michele Giugliano at UNIMORE. Electrophysiological signals were recorded from primary neuronal cultures grown on a passive 60-channel MEA, utilizing the custom lab-developed amplification front-end based on SiNAPS CMOS circuitry. Data acquisition was executed through the previously described SiNAPS acquisition platform and dedicated software at a sampling

rate of 20 kS/s. In this experimental setting, neuronal cells expressing the inhibitory channel rhodopsin Jaws [69] were used. To simulate and evaluate the suppression of epileptiform activity, a 590 nm wavelength LED stimulator was applied across a two-stage protocol. First, during a sustained activation stage of 1 s to induce temporary cell silencing, immediately followed by a subsequent 1 s control stage, during which the LED was rapidly pulsed (turned on and off) with high-frequency non-periodic intervals, to suppress post-silencing rebound bursting.

In the analyzed 100 s long recording, to account for photoelectric switching interference, a 5-sample window immediately following LED activation and deactivation was classified as optogenetic artifact zone. This window duration was selected empirically via visual inspection of the raw voltage traces for this dataset.

### 3.4.3 In vivo Data

In vivo neural activity was recorded from mouse cortex using the SiNAPS probe inserted via the NeuralGlider system. Acquisition was performed with the SiNAPS acquisition platform, at a 20 kS/s sampling rate. The impact of two main factors on acute signal quality was assessed: dura integrity (intact dura vs. durotomy) and NeuralGlider actuation state (ultrasonic microvibration ON vs. OFF). The summary of the recordings is presented in Table 3.1.

Table 3.1: Summary of experimental in vivo recording sessions categorized by NeuralGlider actuation state and dura integrity. The **Condition** column specifies whether ultrasonic actuation was active (ON) or inactive (OFF) during probe insertion, as well as dura mater status (*intact* vs. *removed*). The **Date** column is formatted as YYMMDD. The **Duration** column lists total recording time in seconds. The **Hemisphere** column designates the targeted brain side (left or right), and the **Half** column identifies whether the data corresponds to the first half (1) or second half (2) of the continuous recording session.

| Condition (Actuation / Dura) | Date | Duration (s) | Hemisphere | Half |
|---|---|---|---|---|
| **ON / With** | 260325 | 1211 | LEFT | 1 |
| | 260401 | 1800 | RIGHT | 1 |
| | 260401 | 1800 | RIGHT | 2 |
| | 260415 | 1800 | RIGHT | 1 |
| | 260415 | 1800 | RIGHT | 2 |
| | 260506 | 1800 | RIGHT | 1 |
| | 260506 | 1800 | RIGHT | 2 |
| | 260507 | 1800 | LEFT | 1 |
| | 260512 | 1800 | LEFT | 1 |
| | 260512 | 1800 | LEFT | 2 |
| **ON / Without** | 260429 | 1082 | LEFT | 1 |
| | 260430 | 1800 | RIGHT | 1 |
| | 260430 | 1800 | RIGHT | 2 |
| | 260513 | 1800 | LEFT | 1 |
| | 260513 | 1800 | LEFT | 2 |
| **OFF / With** | 260325 | 1800 | RIGHT | 1 |
| | 260325 | 1800 | RIGHT | 2 |
| | 260401 | 1800 | LEFT | 1 |
| | 260401 | 1800 | LEFT | 2 |
| | 260415 | 1800 | LEFT | 1 |
| | 260415 | 1478 | LEFT | 2 |
| | 260416 | 1800 | RIGHT | 1 |
| | 260416 | 1800 | RIGHT | 2 |
| | 260506 | 1800 | LEFT | 1 |
| | 260506 | 1800 | LEFT | 2 |
| | 260507 | 1800 | RIGHT | 1 |
| | 260507 | 1800 | RIGHT | 2 |
| | 260513 | 1800 | RIGHT | 1 |
| | 260513 | 1800 | RIGHT | 2 |
| **OFF / Without** | 260429 | 1800 | RIGHT | 1 |
| | 260429 | 1800 | RIGHT | 2 |

### 3.4.4 Closed-loop simulation data

Synthetic signals containing a single sinusoidal wave repeated at a 4Hz frequency were generated with an oscilloscope to bench-mark real-time threshold crossing detection. The 4 Hz repetition rate provided a deterministic trigger with a 250 ms inter-event period, ensuring that individual threshold-crossing events were fully processed without risk of buffer accumulation or event overlap. Furthermore, evaluating threshold crossings on a known periodic signal allowed for precise quantification of the pipeline's intrinsic temporal jitter and hardware latency across all channels. The oscilloscope output was connected directly to the acquisition system and the same signal was recorded on all device channels.

## 3.5 Software Environment

To process the acquired electrophysiological data streams, all signal processing pipelines and feature extraction algorithms described in this chapter were implemented using Python, version 3.13. Development, debugging and environment management were conducted within the Visual Studio Code, version 1.101.0.

### 3.5.1 Feature Extraction from In Vivo Data

Signal detection and spike sorting were performed on the in vivo data offline using Kilosort4 [70], an automated framework that employs drift correction and template matching to isolate single-unit activity from high-density extracellular recordings. Spikes were detected across all channels and clustered into distinct neural units, which were categorized as either single-unit activity (SUA or "good" units, representing well-isolated action potentials from individual neurons) or multi-unit activity (MUA, reflecting unresolved signals from multiple neighboring cells). Kilosort4 was used to directly extract metrics including total unit counts, "good" unit yield, unit template amplitudes and unit contamination rates. The contamination rate quantifies refractory period violations ($\leq$ 1.5 ms) derived from the spike auto-correlogram, estimating the percentage of false-positive spikes within a cluster to distinguish well-isolated single units ($<$ 10% contamination) from multi-unit activity (MUA) [70].These metrics were evaluated across two experimental factors, dura integrity (intact vs. durotomy) and actuation state (on vs. off), to confirm that microvibration-assisted insertion reliably yields high-quality single-unit recordings suitable for downstream analysis.

In contrast to the sorting output, root-mean-square (RMS) and signal-to-noise ratio (SNR) were calculated programmatically. The RMS of the continuous raw recording was computed

over consecutive 1-second chunks across the recording channels as:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x[n]^2} \tag{3.1}$$

Additionally, the SNR for each peak-aligned spike waveform was computed as the peak-to-peak amplitude ($V_{\text{p-p}}$) divided by twice the standard deviation of its pre-spike baseline noise segment:

$$\text{SNR} = \frac{V_{\text{p-p}}}{2\sigma_{\text{noise}}} \tag{3.2}$$

This SNR calculation was evaluated at two levels: globally across all sorted "good" units to track session-wide stability and individually for specific target units to assess single-unit reliability over time.

### 3.5.2 Statistical Analysis

Features extracted from the in vivo data were statistically evaluated using the `scipy.stats` module from the SciPy library. Two distinct statistical frameworks were employed to evaluate signal stability over time and assess performance variations across experimental conditions. All statistical hypothesis tests were conducted using a two-tailed significance threshold of $\alpha = 0.05$, corresponding to a 95% confidence level.

To evaluate the continuous stability of recording quality over time, a linear regression analysis was applied to track the signal-to-noise ratio (SNR) across session duration. By modeling SNR as a function of continuous time $t$, the rate of change in signal quality was quantified by the trendline slope $\beta_1$. To determine whether this slope reflected a genuine directional trend or random variance, a $t$-statistic was calculated by dividing the estimated slope by its standard error [71]:

$$t = \frac{\beta_1}{\text{SE}(\beta_1)} \tag{3.3}$$

The corresponding $p$-value was derived to evaluate the probability of observing such a slope under the null hypothesis of constant signal quality ($\beta_1 = 0$) [71]. A statistically significant result ($p < 0.05$) indicates a systematic directional trend, reflecting either signal degradation ($\beta_1 < 0$) or baseline stabilization ($\beta_1 > 0$), whereas a non-significant result ($p \geq 0.05$) indicates that the signal remained stable over time without a statistically detectable trend ($\beta_1 = 0$).

To assess discrete differences in single-unit performance metrics, including total unit count, single-unit count, SUA yield percentage (% Good), "good" units template amplitude ($\text{Amp}_{\text{Good}}$) and "good" units contamination rates ($\text{Contam}_{\text{Good}}$), two-tailed independent samples $t$-tests were

conducted. First, to determine whether single-unit yield and quality varied across recording duration, performance metrics were compared between the first and second halves of each recording session across distinct NeuralGlider actuation states and dura conditions. Second, to directly assess the impact of microvibration-assisted insertion, independent $t$-tests were performed comparing actuated (ON) versus non-actuated (OFF) experiment conditions during the initial recording phase (Half 1) with intact dura. For all tests, the test statistic $t$ was computed as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.4}$$

where $\bar{X}_1$ and $\bar{X}_2$ represent group sample means, $n_1$ and $n_2$ are sample sizes and $s_p$ is the pooled standard deviation [71]. The resulting $t$-statistic was evaluated against Student's $t$-distribution, where $p < 0.05$ was considered statistically significant [71].

### 3.5.3 Data filtering

To condition the streamed electrophysiological data and eliminate out-of-band noise, digital filtering routines were integrated prior to further analysis.

The acquisition system provides simultaneous access to both the unconditioned raw wideband signal and the online bandpass-filtered stream (300 - 5000 Hz). All real-time processing, spike detection and analytical results presented utilize the online filtered data by default, unless explicitly noted that raw wideband recordings were evaluated. This conditioning isolates action potential activity while suppressing low-frequency baseline drift and high-frequency noise, as addressed it Section 2.4.3.

For the in vitro dataset, a Common Average Reference (CAR) was applied to mitigate common-mode interference. Because the switching of the LED generates an electrical artifact that couples synchronously across all recording channels, this shared component was estimated by calculating the ensemble average across channels. Then, subtracting it from each individual trace, the artifact is effectively removed, while leaving the underlying spiking activity largely intact.

Together, these preprocessing steps ensure that the signals entering the detection stage are restricted to the frequency band relevant to spiking activity and, where applicable, free from stimulation-related artifacts, providing a cleaner and more reliable basis for the analysis that follows.

### 3.5.4 Spike Detector

Effective spike detection is essential for spike-based closed-loop processing because it isolates neural firing from background noise, ensuring that real-time feedback is driven by genuine activity [10]. The spike detection framework was developed and evaluated using three methodologies. This design process prioritizes minimal computational bandwidth by transitioning from standard amplitude-based detection to a novel hardware-efficient energy operator cascade that eliminates the necessity for bulky area-smoothing filters [10].

#### Method 1: Simple Root-Mean-Square Thresholding

As an initial baseline validation, a classic amplitude-thresholding pipeline was implemented. The global root-mean-square (RMS) value of the conditioned signal is computed, establishing a static negative threshold defined by:

$$T_{\text{simple}} = A \times \text{RMS}_{\text{signal}} \tag{3.5}$$

where $A$ represents a user-defined threshold multiplier tuned to balance detection sensitivity against false positives and false negatives based on expected signal amplitudes. A signal is detected as an extracellular spike whenever the voltage drops below $-T_{\text{simple}}$. To isolate true physiological action potentials from sustained threshold crossings, a local peak-finding algorithm groups consecutive samples below the threshold and extracts only the local minimum point (negative peak) as the definitive spike timestamp.

#### Method 2: Standalone Amplitude Slope Operator (ASO) Processing

To enhance transient spike features prior to thresholding, the use of non-linear energy operations was explored. In this configuration, two independent non-linear operators were evaluated to determine their efficacy in isolating high-frequency extracellular features. The first mathematical approach implements an Absolute Differential Operator (ADO) designed to capture sharp voltage shifts over a localized interval, defined by:

$$\text{ADO}(n) = |x(n) - x(n - k_s)| \tag{3.6}$$

where $k_s$ represents a discrete temporal delay parameter that dictates the look-back window for the differential calculation [10].

Alternatively, to account for the rapid rate of voltage change while managing out-of-band noise, the pipeline evaluates a standalone modified Amplitude Slope Operator (ASO). To ensure

a strictly positive energy profile suitable for unipolar thresholding, the absolute value is taken across the entire operator product:

$$\mathrm{ASO}(n) = |x(n) \cdot (x(n) - x(n - k_a))| \tag{3.7}$$

where $k_a$ is a secondary discrete temporal delay parameter optimized to emphasize the characteristic slope of extracellular action potentials [10]. Mathematically, while the difference term $(x(n) - x(n - k_a))$ effectively tracks the instantaneous slope of the voltage trace, multiplying this value by the instantaneous amplitude $x(n)$ yields a non-linear metric of signal energy. This multiplication suppresses low-amplitude background noise transients even if they possess sharp slopes, while exponentially amplifying genuine high-amplitude biological spikes that exhibit both high slopes and large voltage magnitudes.

For the final implementation of this methodology, the system utilizes the standalone ASO profile, calculating a static threshold defined by:

$$T_{\mathrm{ASO}} = B \times \mathrm{RMS}_{\mathrm{ASO}} \tag{3.8}$$

where $B$ is a scaling multiplier [10]. Because the thresholding is applied directly to the positive-going $\mathrm{ASO}(n)$ energy wave rather than the raw voltage signal, a specialized peak-alignment routine is required to isolate the definitive spike timestamp. When the $\mathrm{ASO}(n)$ signal exceeds $T_{\mathrm{ASO}}$, the system clusters multiple consecutive crossings caused by a single action potential within a physiological refractory period of 2 ms (Section 2.1.1). The algorithm then maps these grouped indices back to the corresponding window of the raw conditioned signal, executing a local minimum search to extract the true physiological peak. Unlike the strict-adjacency grouping used in Method 1, this refractory-based clustering is necessary because the non-linear ASO energy profile can produce brief non-contiguous threshold crossings around a single action potential. For this reason, grounding the merge window in the neuron's physiological refractory period prevents these from being miscounted as multiple spikes.

#### Method 3: The Target Adaptive Cascaded

The final implementation integrates the ASO energy architecture of Method 2 with a dynamic hardware-optimized adaptive thresholding pipeline to maintain robust performance in the presence of non-stationary background noise. Rather than deriving threshold metrics from the processed energy operators, this architecture dynamically tracks the noise floor of the underlying online bandpass-filtered signal.

The algorithm segments the incoming raw stream into consecutive windows of $M$ samples and calculates the windowed root-mean-square value ($\text{RMS}_{\text{window}}$) [10]. Computing the root-mean-square keeps the windowed noise metrics in the linear voltage domain ($\mu$V), which is highly advantageous for hardware implementation, since it prevents the exponential scaling issues and extreme values that would occur if the threshold multiplier were applied directly to a raw mean-square ($\mu\text{V}^2$) baseline.

To prevent the background noise baseline from being artificially inflated, the windowed values are processed through a 3-point moving median filter [10]:

$$\sigma_{\text{noise}}(n) = \text{median}(\text{RMS}_{\text{window}}(n-1), \text{RMS}_{\text{window}}(n), \text{RMS}_{\text{window}}(n+1)) \tag{3.9}$$

This operation ensures that transient energy bursts, whether from true physiological neural spikes or non-biological artifacts, are excluded from the noise estimation math as statistical outliers, preserving a stable baseline threshold that allows the system to accurately detect the neural spikes. The final dynamic threshold profile is generated by scaling this robust noise estimate by a scaling factor, $T_{\text{target}}(n) = C \cdot \sigma_{\text{noise}}(n)$ [10]. A definitive spike is flagged whenever the condition $ASO(n) > T_{\text{target}}(n)$ is satisfied [10]. Following the threshold comparison, the refractory clustering routine isolates the true physiological local minimum, verifying that the selected peak aligns with a valid negative-going extracellular potential ($x(n_{\text{peak}}) < 0$).

#### 3.5.4.1 Parameter Optimization

The detection scheme described earlier depends on a small number of parameters that are not obvious to set manually: the operator's lag ($k_a$), the threshold scaling constant ($C$) and, for the adaptive-threshold variant used as a comparison method, the length of the window over which the local noise level is estimated ($M$). An exhaustive grid search was carried out for the in vitro dataset parameter optimization, testing every combination of a predefined set of candidate values for the three parameters and scoring each one against a shared performance criterion, so that the final choice reflected the best joint configuration. Each parameter directly governs a specific operational trade-off within the detection pipeline:

- The lag, $k_a$, sets how many samples apart the two samples compared by the ASO are. $k_a$ is selected to match the characteristic timescale of the action potential's sharp voltage drop, effectively accentuating the transient dynamics of the spike waveform.

- The scaling constant, $C$, sets how far above the estimated noise floor a sample must rise before being treated as a candidate spike. Using a low value, more events are detected at

the cost of letting more noise and artifacts through, while using a high value reduces the occurrence of false positives at the cost of discarding genuine smaller spikes.

- The window length, $M$, determines how the noise floor is estimated over time for the adaptive-threshold method. The signal is divided into consecutive windows of $M$ samples, the root-mean-square amplitude is computed for each window and a short three-point moving median is applied across windows to smooth out sudden jumps. In this sense, a small $M$ tracks local changes in noise (such as those caused by the LED itself) more closely, while a large $M$ produces a steadier but slower-reacting estimate.

Because no ground truth was available for the in vitro dataset, the parameter search could not rely on standard precision/recall metrics against a reference sorter. Instead, the LED trigger channel was used as an indirect reference to identify time windows contaminated by stimulation artifacts. Specifically, any detection occurring within a 5-sample safety margin immediately following an LED turn-ON or turn-OFF transition was classified as an optogenetic artifact. This approach carries distinct limitations. Because true spike times remain unknown, it measures artifact suppression rather than genuine detection sensitivity, risking parameter choices that over-filter actual spikes. Additionally, true physiological action potentials occurring within the safety window are inevitably misclassified as artifacts. Lastly, because the trigger channel only marks discrete switching events, it cannot account for noise-induced false positives during baseline or steady-state stimulation.

Using this definition, candidate parameter combinations were evaluated using two derived quantities across the recording:

- Clean detection rate: Detections per second during artifact-free periods.
- LED-locked artifact rate: Detections per second during the cumulative LED transition windows (the 5-sample post-switching intervals).

The ratio between these two rates, measuring how much more likely a detection was to occur during clean periods than during LED switching windows, was multiplied by the clean detection rate to form a single balanced metric. This ensured that an optimal parameter set was both selective (strongly suppressing detections during LED transition windows) and productive (maintaining a robust yield of genuine spikes), rather than achieving high selectivity by indiscriminately suppressing all neural activity.

$$n^{\circ}\text{clean}/t_{\text{clean}} \cdot \frac{n^{\circ}\text{clean}/t_{\text{clean}}}{n^{\circ}\text{LED}/t_{\text{LED}}} \tag{3.10}$$

A total of six candidate values for $k_a$ $(4, 6, 8, 10, 12, 14)$, six for $C$ $(50, 100, 150, 200, 250, 275)$ and four for $M$ $(32, 64, 128, 256)$ were tested, giving 144 parameter combinations. For every combination, the detector was run over the full recording, its detections were split into clean and artifact-related events using the LED-based criteria and the balanced score described above was computed. In parallel with this main criterion, two secondary rankings were also tracked for reference, mainly to check that the chosen combination was not an extreme outlier on either end. These were the combination yielding the highest number of clean detections and the one yielding the fewest LED-related artifacts. Because multiple parameter combinations successfully achieved zero artifact detections, the combination with the lowest parameter values was selected to ensure a conservative baseline.

The combination that maximized the balanced performance score across the full recording was adopted as the final operating point for this dataset and carried forward into the remaining diagnostic steps. Working this way, optimizing over the complete session rather than a short snippet and scoring against an artifact-based criterion drawn directly from the LED channel, meant that the final parameters were chosen specifically to suppress stimulation-related artifacts without overly sacrificing genuine detections, which was the main practical concern for this particular dataset.

### 3.5.5 Real-Time Spike Detection Pipeline

To enable real-time closed-loop control, neural spikes were causally extracted from continuous streaming signals using a chunk-based processing framework that mimicked hardware data acquisition. Incoming multi-channel signals (1024 channels) were processed sequentially without look-ahead capability in small fixed-size buffers ($N_{\text{samples}} \in [16, 512]$).

For the in vitro dataset, spatial noise was firstly removed by subtracting the instantaneous array-wide Common Average Reference (CAR), computed frame-by-frame across all channels. This step adds negligible latency, ensuring low-overhead pre-processing within strict real-time deadlines.

Spike detection was subsequently performed on the absolute value of the ASO operator applied to the cleaned signal stream. To ensure mathematical continuity of the ASO metric across buffer boundaries, an overlap buffer storing the trailing samples ($k_a$) of the preceding chunk was prepended to each new frame, preventing edge discontinuities that would otherwise cause false positives or missed detections.

Candidate spikes were flagged using a channel-specific threshold, calibrated during an initial baseline window (2 s for the synthetic and in vitro datasets, 10 s for the in vivo dataset) by

scaling the root-mean-square (RMS) amplitude of each channel's baseline signal. This per-channel calibration accounts for spatial variations in noise levels caused by electrode impedance, recording depth and local tissue conditions. Consecutive above-threshold samples occurring within a 2 ms refractory window were grouped into a single candidate event. A candidate event was confirmed as a valid spike only if its local trough within that interval was strictly negative, with the index of the minimum voltage designated as the spike timestamp.

Buffer size directly dictates the available execution budget imposed by the sampling rate. Thus, various chunk sizes were evaluated by measuring average per-chunk processing times against available frame durations, ensuring the complete pre-processing and detection sequence executed comfortably within real-time time constraints without sacrificing artifact rejection.

### Validation Across Different Datasets

The detector was evaluated on three different kinds of recordings, each posing a different validation challenge.

For synthetic recordings, because they are simple enough to inspect visually, the real-time detector's output was verified directly against the waveforms as a low-risk sanity check before moving to real data. For the real-time detector, a fixed lag of $k_a = 4$ samples and a scaling constant of $c = 100$ were chosen empirically, as the synthetic signals allowed straightforward visual confirmation of reliable detection. This initial setup served primarily as a controlled baseline to validate the detection logic before performing formal parameter searches.

For the LED optogenetic dataset, where no sorted ground truth was available, the real-time detector (using a fixed threshold) was instead benchmarked against the offline adaptive-threshold variant to quantify the performance trade-off required for real-time latency. For the real-time detector, the parameters obtained from the offline optimization were used ($k_a$ and $C$). Detected events from both implementations were matched within a small time window of 1 ms to categorize them as mutual or unique detections. Additionally, all events were cross-referenced with the LED trigger channel, flagging spikes within a short safety window of 5 samples after LED onset or offset as stimulation artifacts to evaluate clean and artifact-contaminated detections separately.

For in vivo chronic recordings from a multi-shank probe, the detector's output was benchmarked against spikes obtained offline with Kilosort4 [70], restricted to the units it labeled as "good" (single-unit activity) to exclude ambiguous multi-unit activity and noise, ensuring the detector was evaluated against well-isolated and defined single-neuron action potentials. Since a real neuron's spike is often visible, at reduced amplitude, on several neighboring electrodes and not exclusively on the one where it peaks, a detected spike was considered a match to a

Kilosort4 spike if it fell within a short time tolerance of 2 ms of a reference spike whose peak channel was within a small radius of electrodes from the one being evaluated (3 channels). From these matches, precision (the fraction of detected events that corresponded to a real, sorted unit), recall (the fraction of real units that were successfully detected) and their harmonic mean, the $F_1$ score, were computed per channel. This same ground truth was firstly used the other way around. Rather than fixing the detector's two main parameters (the ASO operator's lag ($k_a$) and the threshold scaling factor ($c$)) arbitrarily, a small grid search was run over candidate values on a short calibration segment (10 s) of each channel to select the combination that maximized the $F_1$ score against Kilosort4 before running the detector on the full recording. This per-channel optimization was necessary because this recording lacked a uniform noise or artifact profile across channels variation in impedance and local tissue conditions, which meant that a single parameter pair could not perform optimally everywhere.

### 3.5.6 Real-Time Spike Sorting and Clustering

Because a single electrode records signals from multiple neighboring neurons, spike sorting is used to match each detected spike to the specific cell that produced it (Section 2.4.3). Once a spike is detected, a short waveform snippet around its trough is extracted (3 ms) and three simple interpretable features are computed from it. These are the spike's amplitude (depth of the trough), its peak-to-trough ratio (how large the rebound after the trough is, relative to the trough itself) and a rough estimate of its width (the number of samples between the trough and the following peak). These are simple biophysically motivated descriptors, deliberately chosen instead of more abstract representations such as PCA components, since amplitude, sharpness and duration are known to vary systematically between neurons and offer low computational complexity for real-time processing [54]. A reduced two-feature version (amplitude and peak-to-trough ratio only, dropping width) was also tested, to check whether the extra feature was contributing to separability or simply adding computational cost.

Because the sorter must operate in real time, it cannot afford to re-cluster the entire dataset every time a new spike arrives. Instead, all the sorters implemented here follow the same two-phase logic. During a short warm-up period, incoming spikes are simply pooled and temporarily left unlabeled, while the system waits until enough events have accumulated (50 spikes) to have a representative sample of the ongoing activity. Once this calibration quota is reached, clustering is run once on this reference pool to define the initial set of units and all the buffered spikes are retroactively relabeled accordingly. From that point on, every new spike is classified by comparing its features against the reference pool using a nearest-neighbor vote (the label is decided by a small majority vote among its 5 closest neighbors in feature space), rather than by

rerunning the clustering algorithm. This design keeps the heavy computational step (the actual clustering) confined to a single calibration moment, while the steady-state classification of each individual spike remains a fast constant-time operation compatible with the acquisition system's timing budget.

### Clustering Algorithms

To achieve low-latency real-time spike classification, a unified two-phase pipeline architecture was implemented across both synthetic and in vivo datasets. The pipeline operates by first running an offline Calibration Phase on an initial warm-up buffer of spikes to discover neural units and establish a reference library, followed by an online Streaming Phase that classifies incoming spikes on-the-fly.

**Calibration Phase** During the initial warm-up period, raw baseline spikes are collected without prior labels. Because ground-truth unit assignments are unavailable, unsupervised clustering is performed to discover cluster identities across the reference pool:

- **Synthetic Dataset ($k$-Means Calibration):** For the synthetic benchmark datasets, $k$-Means clustering ($k$) was used as a fast baseline model. $K$-Means partitions unlabelled feature vectors into $k$ pre-specified clusters by iteratively minimizing the distance between each point and its assigned cluster centroid. The algorithm initializes $k$ centroids in feature space, assigns each baseline spike to its nearest centroid using Euclidean distance and updates each centroid position to the mean of its assigned points until positions stabilize [72]. While computationally fast, its main limitation is that the number of clusters ($k$) must be specified a priori, a reasonable constraint for synthetic validation where unit count is known, but unsuitable for exploratory in vivo recordings.
- **In Vivo Dataset (Louvain Modularity Calibration):** To overcome the fixed-cluster limitation in experimental data, Louvain community detection was utilized during calibration. Rather than forcing data points into pre-defined spherical groups, his method constructs an unsupervised $k$-nearest neighbors ($k$-NN) graph where each calibration spike is a node connected to its $k$ closest neighbors in feature space. The algorithm then partitions the graph into communities by maximizing modularity, a metric that compares the density of intra-community connections against a random network baseline. A resolution parameter controls the scale of community detection, tuning whether fine-grained or coarse-grained unit groups emerge (set to 0.6) [73]. The primary advantage is that the number of active neural units is discovered naturally from the topological structure of the data rather than imposed beforehand.

**Streaming Phase** Once initial cluster labels are established, the reference feature vectors and their assigned unit IDs are stored in memory as a calibrated exemplar pool. The pipeline then transitions to a real-time streaming mode using $k$-NN classification. During streaming, each newly detected spike is classified on-the-fly by calculating its Euclidean distance to all reference exemplars in the calibrated buffer and assigning the majority class label among its $k$-closest neighbors [72]. Using instance-based $k$-NN lookup during streaming provides two distinct advantages. First, unlike the single-centroid matching used in $k$-means, evaluating distance against individual reference exemplars allows the classifier to form flexible non-spherical decision boundaries around complex biological features [72]. Second, $k$-NN eliminates the iterative computational loops required by algorithms like Louvain community detection, ensuring low-latency determinism so that per-spike processing latency remains strictly within the real-time execution budget per chunk.

### Sorting Pipelines

Both single-channel and multi-channel pipelines were implemented. The single-channel sorter classifies each spike using only the features computed from the channel on which it was detected. To evaluate the effect of feature richness on real-time budget and unit separation on the synthetic dataset, three single-channel feature sets were implemented:

- **1D Raw Amplitude:** Utilizes solely the negative peak amplitude of the detected spike. In a 1D feature space, complex cluster geometry is not a concern, as distance-based partitioning does not suffer from high-dimensional shape distortions or non-spherical cluster assumptions. Consequently, $k$-Means was retained for the streaming phase, as classifying 1D feature vectors against cluster centroids reduces to simple scalar threshold comparisons, eliminating $k$-NN sorting overhead [72].
- **2D Feature Set:** Combines the absolute peak amplitude with the peak-to-trough ratio (P2T), capturing waveform asymmetry.
- **3D Feature Set:** Expands the 2D set by incorporating a biophysical spike width, measured as the sample distance from the negative trough to the subsequent positive peak.

For the in vivo experimental data, a single-channel pipeline with the 3D Feature Set and a multi-channel pipeline were compared. The multichannel sorter instead builds a spatial footprint for each spike. Because a single neuron produces attenuated signals across multiple neighboring electrodes [51], the feature vector was constructed as a 7-channel spatial footprint ($\pm 3$ channels around the peak) using trough amplitudes normalized relative to the central channel.This

spatial signature is often more discriminative than single-channel waveform shape alone, particularly when distinct neurons exhibit similar action potential dynamics on their primary recording channel. Because the spatial attenuation profile across adjacent channels depends directly on the neuron's physical distance, depth and dendritic orientation relative to the electrode array, multichannel footprints reliably distinguish units that would otherwise appear identical on a single electrode [51]. The trade-off is computational, since the multichannel approach requires reading and processing several channels for every single detected spike, rather than one, which increases the per-spike processing time and therefore the pressure on the real-time budget.

### 3.5.7 Validation with Ground Truth

For the in vivo dataset, the output of each online sorter was compared against an offline reference obtained with Kilosort4, restricted to units it classified as "good." Since the cluster labels produced by an unsupervised method have no inherent correspondence to the reference labels (cluster "0" from one method will not mean the same thing as cluster "0" from another), the best possible one-to-one mapping between predicted and reference clusters was found using the Hungarian algorithm, which solves this assignment problem optimally. It works by constructing a cost matrix between all possible pairs and systematically subtracting the smallest values from each row and column until a complete set of zero-cost, one-to-one matches can be selected [74]. Sorting accuracy was then computed as the fraction of correctly matched spikes under this optimal mapping. In parallel, the processing time was measured, so that sorting accuracy could always be interpreted alongside its computational cost.

This set of pipelines (a simple fixed-$k$ $k$-means baseline, a graph-based Louvain sorter with an automatically determined number of units and their single- versus multichannel variants) allowed a direct comparison between sorting accuracy and computational cost under the same real-time constraints. This comparison was essential to identify which combination of feature set, clustering method and channel scope offered the best balance between sorting quality and the strict timing budget imposed by the closed-loop acquisition system described in the previous section.

## 3.6 Closed-loop System Architecture

To support state-sensitive neurostimulation, closed-loop platforms must process neural signals and trigger feedback within narrow physiological time windows, without excessive delays

from concurrent tasks like visualization and logging. Therefore, this section describes the architecture, software design and operational latency of the closed-loop system. The objective was to achieve a low-latency closed-loop neural platform with high spatiotemporal resolution.

The closed-loop platform was designed to acquire neural signals, process them online and drive an effector with minimal latency while simultaneously supporting live visualization and data logging (Figure 3.1). The system architecture centers around an Opal Kelly FPGA-driven acquisition unit, the same module utilized by the SiNAPS acquisition system, as the closed-loop architecture was specifically designed for integration with the SiNAPS platform. This unit interfaces directly with the experimental preparation and distributes incoming neural data in parallel to both a visualization workstation and a real-time closed-loop processing workstation.

The experimental interface completes the physical feedback loop. The probe/electrode array acts as the signal source acquired by the FPGA, while the effector is driven by the FPGA output, with its biological response re-measured by the probe.

### Data Acquisition and Distribution

The FPGA continuously acquires probe signals and mirrors the digitized data stream across two independent communication channels to prevent computational load from interfering with latency requirements:

- **Camera Link stream:** Routed to a frame grabber on the visualization workstation running the SiNAPS Control Suite for real-time monitoring and data logging control, as mentioned in Section 3.4.
- **USB stream:** Routed via the FPGA's USB/FrontPanel endpoint to a custom Python script on the closed-loop workstation.

### Closed-Loop Feedback Path

On the closed-loop workstation, incoming data is evaluated against a detection criterion (threshold crossing). Upon event detection, the script issues a trigger command over USB to the FPGA, which converts it into an actuation output delivered to the effector. Three functionally distinct signal types were defined and color-coded for clarity in the figures: data streams in blue (probe acquisition, Camera Link and USB data), visualization control commands in orange (start/stop play and record) and closed-loop feedback/actuation signals in green (detection triggers and actuation output).

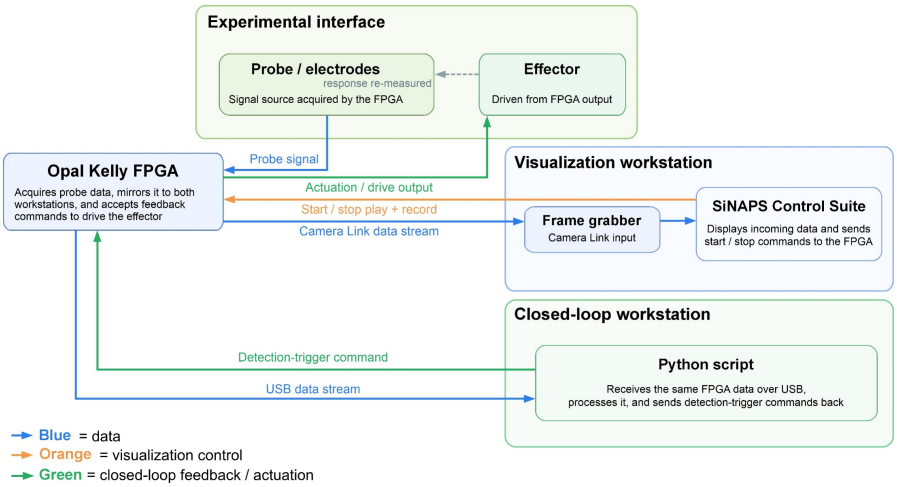



Figure 3.1: Closed-loop system implementation. Blue arrows represent data flow, orange arrows represent visualization control flow and green arrows represent closed-loop feedback and actuation signal paths.

## Closed-Loop Software Implementation

The closed-loop Python pipeline uses a multiprocessing architecture to execute acquisition and processing concurrently (Figure 3.2):

- **Main Supervisor Process:** Initializes inter-process communication queues, manages stop events, launches subprocesses and handles optional foreground plotting. Key parameters include selected channel (`selCh` = 0) and detection mode (e.g., threshold).
- **`acquireData()` Process:** Configures the FPGA, reads probe settings and streams raw data chunks (64 × 512 words; 32 frames of 1024 samples) into `raw_data_queue`. It also consumes `control_queue` commands to issue actuation signals to the FPGA.
- **`processData()` Process:** Consumes `raw_data_queue` chunks, parses frames and performs data integrity checks. In threshold mode, samples exceeding the threshold push a detection triggered event to `control_queue` for FPGA execution.
- **`plotData()` Process:** Displays selected channel data on an interactive Matplotlib interface using a downsampled 1-second ring buffer.

Process roles are isolated asynchronously using dedicated queues, guaranteeing a low-latency path for the detection-to-actuation cycle.

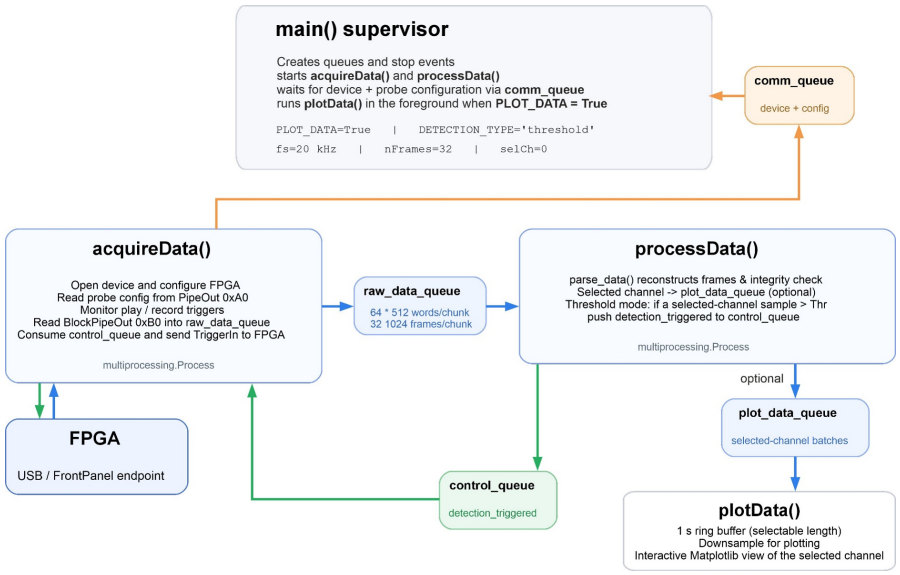



Figure 3.2: Python Script implemented for Closed-loop Control. Blue arrows represent data flow, orange arrows represent visualization control flow and green arrows represent closed-loop feedback and actuation signal paths.

# Chapter 4

# Results and Discussion

This chapter presents the experimental results obtained across the different components of this work, spanning insertion performance, closed-loop system integration and several independent applications of the real-time signal processing pipeline.

## 4.1 Neural Glider - probe insertion

To determine whether ultrasonic microvibrations mitigate mechanical insertion trauma, this section evaluates in vivo electrophysiological recordings acquired using the NeuralGlider.

Spike detection and sorting were performed offline using Kilosort4 [70], which extracts metrics including total unit counts, “good” unit yield, unit template amplitudes and unit contamination rates to be compared across two experimental factors: dura integrity and actuation state, as summarized in Table 4.1.

To evaluate the impact of mechanical insertion parameters on recording stability, isolated single-unit metrics were compared across NeuralGlider actuation states (Actuated vs. Unactuated) and dura integrity conditions (Intact Dura vs. Durotomy). Across sessions, total unit yields ranged from 269 to 938 detected clusters, with the proportion of high-quality isolated units (% Good) spanning 8.8% to 33.8%. Actuated insertions consistently yielded robust unit metrics, reaching peak "good"-unit yields above 30% (e.g., up to 33.8% in the Actuated / Intact Dura condition and 30.5% in the Actuated / Durotomy condition). Isolated single units demonstrated clear signal isolation over background multi-unit activity, with mean "good"-unit amplitudes ($Amp_{\text{Good}}$) consistently outperforming MUA amplitudes ($Amp_{\text{MUA}}$) across all conditions (typically 14 - 22 $\mu$V vs. 10 - 14 $\mu$V, respectively). Furthermore, single-unit contamination

Table 4.1: Single-unit quality metrics and yield across NeuralGlider actuation states and dura conditions. For each recording session, metrics include total detected units, isolated "good" units, yield percentage (% Good), peak-to-peak signal amplitude (*Amp*) for "good" units and MUA, and contamination values (*Contam*) for "good" units and MUA.

| **Actuation** | **Dura** | **Total** | **Good** | **% Good** | $\text{Amp}_{\text{Good}}$ [$\mu$V] | $\text{Amp}_{\text{MUA}}$ [$\mu$V] | $\text{Contam}_{\text{Good}}$ [%] | $\text{Contam}_{\text{MUA}}$ [%] |
|---|---|---|---|---|---|---|---|---|
| OFF | Intact | 430 | 96 | 22.33 | 14.40 | 11.64 | 3.84 | 80.69 |
| OFF | Intact | 471 | 112 | 23.78 | 15.01 | 11.45 | 4.33 | 93.20 |
| OFF | Intact | 415 | 99 | 23.86 | 14.38 | 10.91 | 4.72 | 79.31 |
| ON | Intact | 444 | 135 | 30.41 | 22.24 | 12.32 | 3.18 | 319.35 |
| OFF | Intact | 529 | 97 | 18.34 | 16.95 | 13.72 | 3.18 | $1.51 \times 10^{13}$ |
| OFF | Intact | 938 | 83 | 8.85 | 15.50 | 11.37 | 2.27 | 131.02 |
| OFF | Intact | 402 | 106 | 26.37 | 15.57 | 10.74 | 2.64 | $3.87 \times 10^{13}$ |
| ON | Intact | 293 | 45 | 15.36 | 11.23 | 10.40 | 2.74 | $1.50 \times 10^{14}$ |
| ON | Intact | 269 | 35 | 13.01 | 9.95 | 10.46 | 1.23 | $1.22 \times 10^{14}$ |
| OFF | Intact | 466 | 102 | 21.89 | 16.29 | 11.27 | 3.58 | $1.06 \times 10^{13}$ |
| OFF | Intact | 550 | 159 | 28.91 | 17.90 | 13.61 | 3.45 | $2.28 \times 10^{12}$ |
| OFF | Intact | 482 | 141 | 29.25 | 18.84 | 13.65 | 2.98 | 342.26 |
| ON | Intact | 424 | 110 | 25.94 | 15.14 | 11.28 | 3.60 | $2.06 \times 10^{13}$ |
| ON | Intact | 437 | 116 | 26.54 | 16.36 | 14.11 | 3.83 | $1.42 \times 10^{13}$ |
| OFF | Intact | 378 | 78 | 20.63 | 14.43 | 16.08 | 2.02 | $5.67 \times 10^{13}$ |
| OFF | Intact | 387 | 68 | 17.57 | 14.69 | 11.13 | 2.57 | $7.77 \times 10^{13}$ |
| ON | Durotomy | 395 | 93 | 23.54 | 18.43 | 12.07 | 4.10 | $1.50 \times 10^{12}$ |
| OFF | Durotomy | 511 | 130 | 25.44 | 19.37 | 13.92 | 3.96 | $8.85 \times 10^{11}$ |
| OFF | Durotomy | 531 | 113 | 21.28 | 21.08 | 13.84 | 3.08 | $2.44 \times 10^{12}$ |
| ON | Durotomy | 337 | 50 | 14.84 | 10.19 | 12.86 | 2.49 | $1.11 \times 10^{13}$ |
| ON | Durotomy | 486 | 71 | 14.61 | 13.46 | 11.88 | 2.34 | $3.15 \times 10^{11}$ |
| OFF | Intact | 528 | 113 | 21.40 | 22.13 | 12.35 | 3.27 | $2.93 \times 10^{11}$ |
| OFF | Intact | 504 | 99 | 19.64 | 20.58 | 38.27 | 4.26 | $6.44 \times 10^{10}$ |
| ON | Intact | 639 | 132 | 20.66 | 18.81 | 13.51 | 3.13 | 238.41 |
| ON | Intact | 793 | 156 | 19.67 | 17.12 | 12.47 | 2.00 | 173.07 |
| ON | Intact | 929 | 135 | 14.53 | 12.74 | 10.01 | 2.30 | 96.12 |
| OFF | Intact | 609 | 153 | 25.12 | 20.85 | 13.26 | 4.88 | $4.19 \times 10^{10}$ |
| OFF | Intact | 648 | 171 | 26.39 | 21.75 | 12.65 | 4.91 | $1.09 \times 10^{13}$ |
| ON | Intact | 571 | 193 | 33.80 | 16.09 | 16.03 | 4.34 | $2.11 \times 10^{13}$ |
| ON | Intact | 482 | 144 | 29.88 | 17.29 | 13.46 | 4.41 | $7.68 \times 10^{12}$ |
| ON | Durotomy | 557 | 170 | 30.52 | 18.02 | 13.04 | 4.44 | $9.14 \times 10^{12}$ |
| ON | Durotomy | 544 | 122 | 22.43 | 19.61 | 13.60 | 4.66 | $8.36 \times 10^{11}$ |
| OFF | Intact | 478 | 81 | 16.95 | 14.48 | 11.81 | 2.78 | 868.98 |
| OFF | Intact | 401 | 80 | 19.95 | 17.92 | 14.33 | 4.42 | $3.06 \times 10^{13}$ |

rates ($\text{Contam}_{\text{Good}}$, typically $< 5\%$) were substantially lower than those observed for multi-unit clusters ($\text{Contam}_{\text{MUA}}$), confirming high sorting purity. These data confirm stable single-unit isolation suitable for real-time processing.

Focusing on a single unit reveals high waveform consistency across individual spikes, closely matching the extracted template waveform (Figure 4.1).

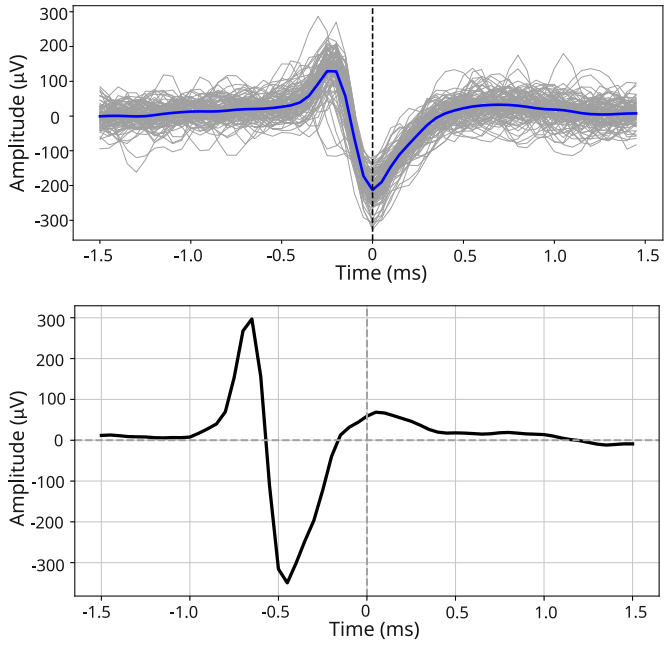



Figure 4.1: Analysis of spike waveforms, main-channel template morphology and signal-to-noise ratio (SNR) stability across the recording session for a representative single unit acquired via NeuralGlider insertion. **Top:** Aligned single-unit waveform overlay: overlay of all detected action potentials for Unit 4 on Channel 6 (gray), along with the mean spike waveform (blue trace), demonstrating cleaned baseline alignment and consistent peak morphology. **Bottom:** Kilosort template waveform: dominant Kilosort template for Unit 4 evaluated on primary Channel 6, highlighting the characteristic action potential shape used for unit assignment.

To address potential temporal signal degradation and characterize the baseline electrophysiological environment, single-unit signal-to-noise ratios (SNRs) and wideband root-mean-square (RMS) amplitudes were evaluated across the recording session.

Tracking individual single-unit SNRs across one session (Figure 4.2A) reveals a gradual decline over time, reflecting a slight degradation in isolated unit signal quality. However, the global single-unit SNR averaged across all recording channels (Figure 4.2B) remains largely stable, aside from an initial transient fluctuation. An overall mean SNR of 3.03 (corresponding

to 9.63 dB) confirms robust signal quality and reliable spike isolation within the action potential band, closely matching the state-of-the-art benchmark of 9 *dB* discussed in Section 2.3.

Conversely, the elevated unfiltered RMS values illustrated in Figure 4.2C (approximately 10 151 $\mu$V) reflect the total signal power of the unconditioned wideband recording prior to processing. Evaluating the unfiltered RMS provides a direct measure of the raw physical recording environment, capturing the total in-band and out-of-band signal amplitude before any digital conditioning occurs. As established in Section 2.4.4, this high amplitude (standing several orders of magnitude above the $\sim 5\,\mu V_{RMS}$ action potential band benchmark) is primarily driven by low-frequency Local Field Potentials (LFPs), electrochemical DC offsets and motion artifacts. Reporting this raw baseline illustrates the full scale of out-of-band background activity that must be attenuated by bandpass filtering prior to single-unit spike extraction.

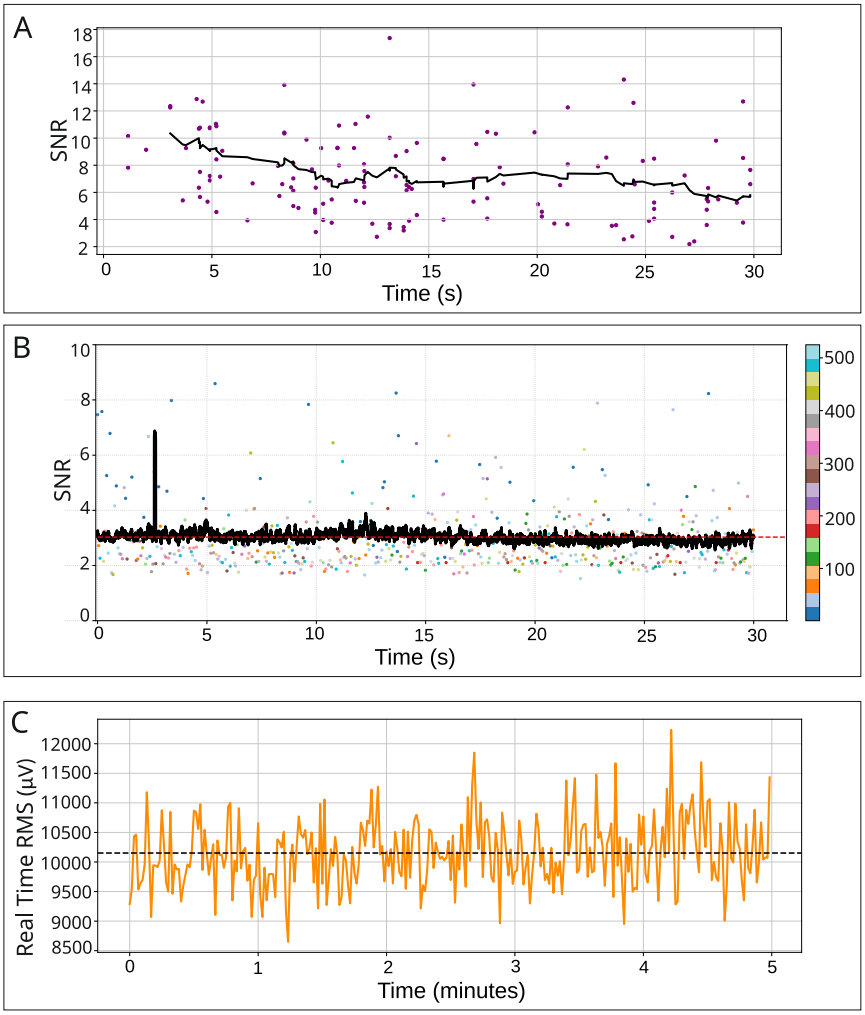



Figure 4.2: Assessment of global single-unit signal quality across all identified units and raw background noise tracking during a representative NeuralGlider recording session. (A) Single-Unit SNR Tracking Across Session: Time-course scatter plot of one single-unit individual spike SNRs evaluated on Channel 6, with a moving average trend line illustrating temporal signal quality and stability throughout the recording (B) Global Single-Unit SNR Stability Across Session: Scatter plot of individual spike SNRs computed across all identified single units on their respective primary channels (color-coded by Unit ID), overlaid with a rolling trend line to illustrate overall signal stability over time. (C) Real-Time Background Noise RMS: Continuous root-mean-square (RMS) trace of the unfiltered signal on Channel 6 over the course of the recording, depicting baseline noise level variations.

To investigate the dynamics of signal quality and probe-tissue integration, Signal-to-Noise Ratio (SNR) trajectories were evaluated across all experimental conditions.

Analysing these trajectories revealed that SNR generally decreased over time (Table 4.2), indicating a degradation in signal quality. This downward trend is also evident in the temporal trajectories of the mean SNR shown in Figure 4.3 for the intact-dura conditions (both with and without actuation). This acute decline is driven primarily by insertion-induced micro-trauma, edema and localized vascular leakage, as detailed in Section 2.3.

However, four sessions demonstrated a statistically significant increase in SNR over the observation period across experimental conditions. In line with the recovery mechanisms outlined in Section 2.3, this temporal improvement reflects an acute stabilization phase during which surrounding tissue recovers from mechanical shock, edema clears and tissue relaxation restores electrode-neuron proximity.

Conversely, two sessions exhibited no statistically significant temporal variation in SNR, indicating stationarity in mean signal quality. This stationarity suggests a highly stable electrode-tissue interface and a constant baseline noise floor throughout the recording [8].

Taken together, these divergent trajectories reflect a spectrum of acute tissue responses governed by the local degree of insertion trauma. Depending on the extent of initial mechanical disruption, the microenvironment surrounding the probe can either suffer progressive signal degradation driven by localized edema and cellular depression, enter a recovery phase as acute trauma resolves and tissue settles, or achieve immediate equilibrium characterized by stationary signal quality [9, 8].

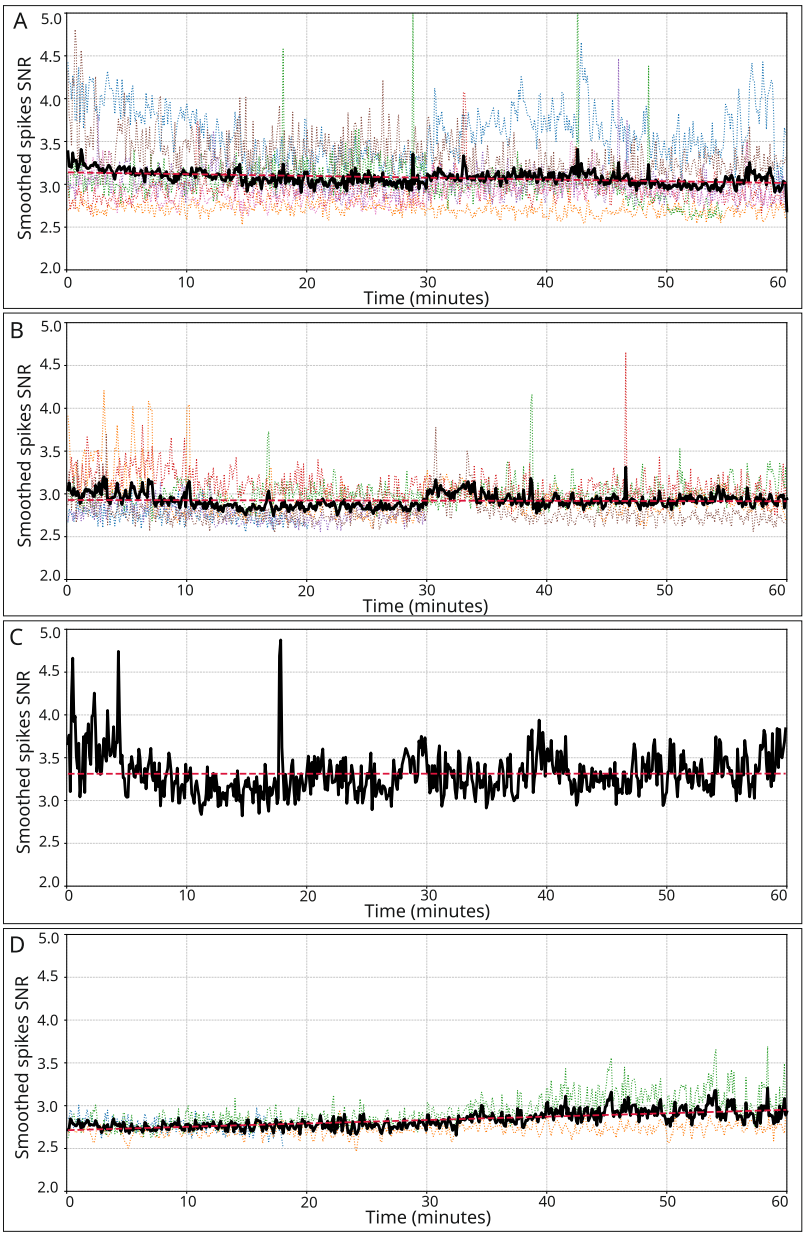



Figure 4.3: Smoothed spike Signal-to-Noise Ratio (SNR) trajectories calculated across all identified single units on their respective primary channels, categorized by dura status and actuation state during NeuralGlider insertion. Colored lines represent individual recording sessions, while solid black lines denote the ensemble average across recordings.(A) Non-Actuated Insertion with Intact Dura: Temporal evolution of unit SNRs recorded during control insertions performed without mechanical actuation through the intact dura. (B) Actuated Insertion with Intact Dura: Temporal evolution of unit SNRs recorded during micro-actuated insertions performed through the intact dura. (C) Non-Actuated Insertion with Dura Removed: Temporal evolution of unit SNRs recorded during control insertions performed without mechanical actuation following durotomy. (D) Actuated Insertion with Dura Removed: Temporal evolution of unit SNRs recorded during micro-actuated insertions performed following durotomy.

Table 4.2: Signal-to-Noise Ratio (SNR) and Trend Analysis across Experimental Sessions

| Session Identifier | Mean SNR | Trend Slope | $p$-value | Significant |
|---|---|---|---|---|
| **Actuation, With Dura** | | | | |
| 260325 | 2.7581 | −0.00380 | $1.28 \times 10^{-5}$ | TRUE |
| 260401 | 2.9530 | −0.00597 | $1.24 \times 10^{-19}$ | TRUE |
| 260415 | 3.0125 | 0.00159 | $2.44 \times 10^{-6}$ | TRUE |
| 260506 | 3.0886 | −0.00380 | $1.23 \times 10^{-40}$ | TRUE |
| 260507 | 2.7858 | −0.00523 | $1.21 \times 10^{-20}$ | TRUE |
| 260512 | 2.7831 | −0.00135 | $4.30 \times 10^{-13}$ | TRUE |
| **Actuation, No Dura** | | | | |
| 260429 | 2.7843 | −0.00749 | $2.17 \times 10^{-11}$ | TRUE |
| 260430 | 2.7282 | 0.00129 | $3.99 \times 10^{-7}$ | TRUE |
| 260513 | 2.9230 | 0.00692 | $8.94 \times 10^{-126}$ | TRUE |
| **No Actuation, With Dura** | | | | |
| 260325 | 3.5708 | −0.00468 | $3.58 \times 10^{-13}$ | TRUE |
| 260401 | 2.7131 | −0.00121 | $1.08 \times 10^{-9}$ | TRUE |
| 260415 | 3.0561 | −0.00377 | $1.37 \times 10^{-15}$ | TRUE |
| 260416 | 2.9417 | 0.00115 | 0.00591 | TRUE |
| 260506 | 3.0134 | −0.00234 | $6.25 \times 10^{-11}$ | TRUE |
| 260507 | 3.3286 | −0.00601 | $1.12 \times 10^{-36}$ | TRUE |
| 260513 | 2.8993 | 0.00055 | 0.14672 | FALSE |
| **No Actuation, No Dura** | | | | |
| 260429 | 3.3223 | −0.00013 | 0.78125 | FALSE |

To isolate the specific impact of each experimental variable on recording performance, direct pairwise comparisons were conducted across actuation state (Actuated vs. Unactuated) and recording half ($0-30$ min *vs*. $30-60$ min) for different parameters (Figure 4.4).

Analyzing the temporal dynamics across all experimental conditions, independent-samples $t$-tests revealed no statistically significant differences in unit and single-unit metrics between the first and second halves of the recording ($p > 0.05$) (Table 4.3). This indicates that the insertion yielded robust overall signal and unit stability over time, despite the divergent SNR trends observed previously.

Furthermore, when evaluating the specific impact of actuation in intact dura, independent-samples $t$-tests similarly revealed no statistically significant differences in single-unit metrics between the actuated and unactuated conditions ($p > 0.05$), demonstrating that micro-actuation did not measurably compromise or alter electrophysiological recording performance or signal quality (Table 4.4).

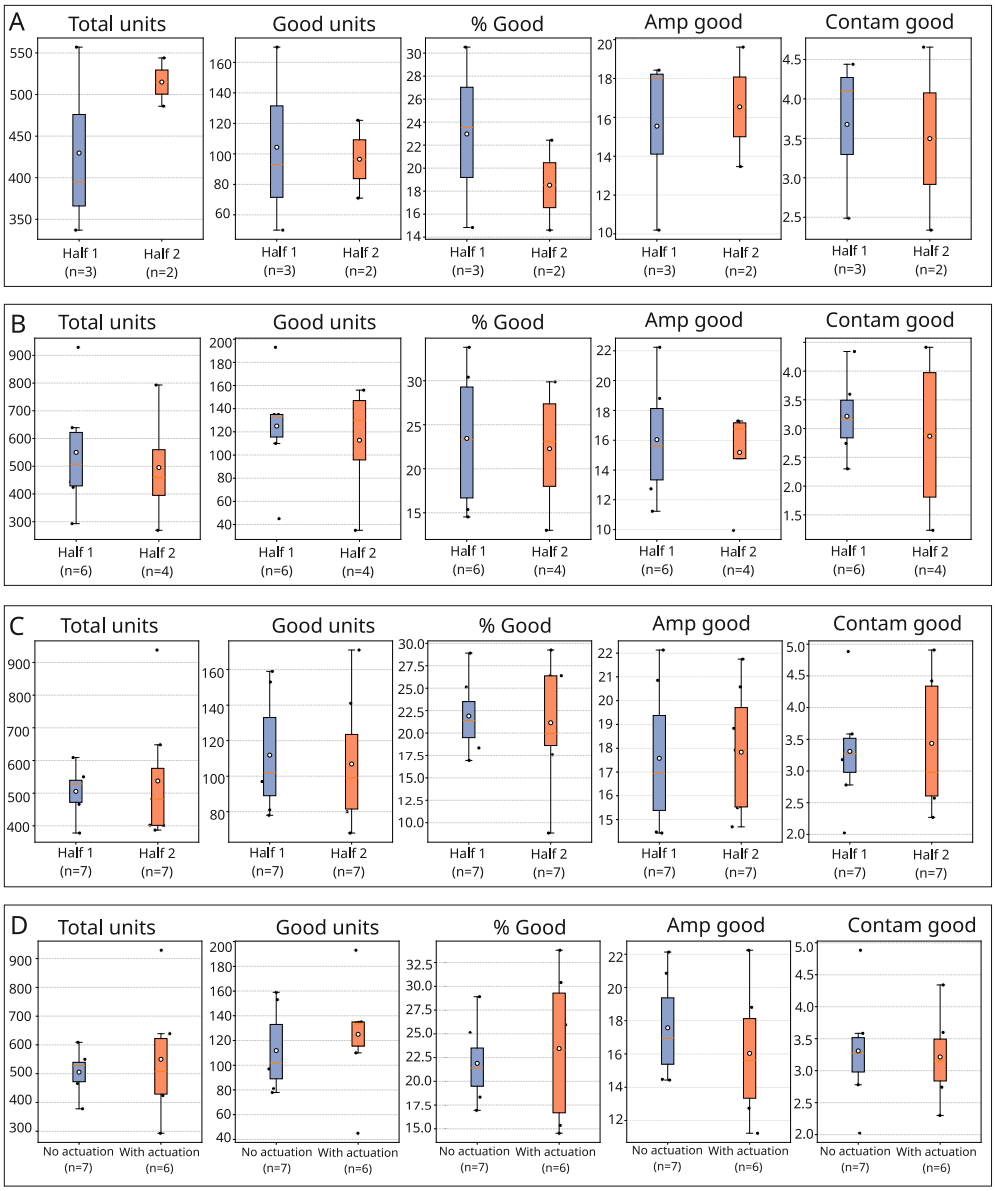



Figure 4.4: Boxplot comparisons of signal and unit metrics evaluated during the first ($0-30$ min) and second ($30-60$ min) halves of in vivo recording sessions across different NeuralGlider actuation states and dura conditions. For each boxplot, the red horizontal line denotes the median value, the white marker indicates the sample mean, black dots represent individual data points, the box boundaries enclose the interquartile range (IQR) and the black whiskers extend to the last data point within the $1.5 \times$ IQR. **(A)** Actuated insertion without dura: comparison of metrics between the first and second halves of the recording under the micro-actuated condition following durotomy. **(B)** Actuated insertion with intact dura: comparison of metrics between the first and second halves of the recording under the micro-actuated condition through the intact dura. **(C)** Unactuated insertion with intact dura: comparison of metrics between the first and second halves of the recording under the unactuated control condition through the intact dura. **(D)** Baseline comparison of actuation effects (first half): comparative distribution of parameters between actuated and unactuated insertions during the initial recording half with an intact dura.

Table 4.3: Independent-samples $t$-test results comparing single-unit performance metrics between the first half ($0-30$ min) and second half ($30-60$ min) across NeuralGlider actuation states and dura conditions.

| **Condition** | **Metric** | $t$**-statistic** | $p$**-value** | **Result** |
|---|---|---|---|---|
| Actuated / Durotomy | Total Units | $-1.186$ | 0.3309 | Not Significant |
| | "Good" Units | 0.181 | 0.8682 | Not Significant |
| | % "Good" | 0.743 | 0.5133 | Not Significant |
| | $\text{Amp}_{\text{Good}}$ ($\mu$V) | $-0.243$ | 0.8269 | Not Significant |
| | $\text{Contam}_{\text{Good}}$ (%) | 0.138 | 0.9060 | Not Significant |
| Unactuated / Intact Dura | Total Units | $-0.400$ | 0.7001 | Not Significant |
| | "Good" Units | 0.269 | 0.7923 | Not Significant |
| | % "Good" | 0.246 | 0.8112 | Not Significant |
| | $\text{Amp}_{\text{Good}}$ ($\mu$V) | $-0.170$ | 0.8682 | Not Significant |
| | $\text{Contam}_{\text{Good}}$ (%) | $-0.243$ | 0.8126 | Not Significant |
| Actuated / Intact Dura | Total Units | 0.386 | 0.7116 | Not Significant |
| | "Good" Units | 0.365 | 0.7276 | Not Significant |
| | % "Good" | 0.237 | 0.8196 | Not Significant |
| | $\text{Amp}_{\text{Good}}$ ($\mu$V) | 0.357 | 0.7316 | Not Significant |
| | $\text{Contam}_{\text{Good}}$ (%) | 0.433 | 0.6881 | Not Significant |

Table 4.4: Independent-samples $t$-test results comparing single-unit performance metrics between unactuated and actuated insertions in the intact dura condition during the first recording half ($0-30$ min).

| **Comparison** | **Condition** | **Metric** | $t$**-statistic** | $p$**-value** | **Result** |
|---|---|---|---|---|---|
| Unactuated vs. Actuated | Intact Dura | Total Units | $-0.471$ | 0.6542 | Not Significant |
| | | "Good" Units | $-0.569$ | 0.5841 | Not Significant |
| | | % "Good" | $-0.435$ | 0.6764 | Not Significant |
| | | $\text{Amp}_{\text{Good}}$ ($\mu$V) | 0.772 | 0.4598 | Not Significant |
| | | $\text{Contam}_{\text{Good}}$ (%) | 0.216 | 0.8330 | Not Significant |

Overall, the NeuralGlider system demonstrated high recording stability across all tested conditions. Although it was initially hypothesized that micro-actuation would yield statistically significant improvements in unit yield and signal fidelity by mitigating insertion trauma, the differences between the actuated and unactuated conditions did not reach statistical significance ($p > 0.05$). This lack of significance may be attributed to a combination of modest sample sizes, post-insertion tissue relaxation masking subtle mechanical advantages over short recording windows, or baseline insertion techniques already achieving near-optimal signal isolation in this preparation. Additionally, due to the limited number of experimental observations under durotomy conditions, findings for these specific groups were treated as exploratory rather than

generalizable trends.

Nevertheless, the absence of significant temporal degradation between the first and second recording halves confirms robust short-term stability. Furthermore, while acute recordings evaluate immediate mechanical entry, the primary benefits of microvibration-assisted insertion (such as reduced chronic glial scar formation and improved long-term tissue integration) may be more pronounced in chronically implanted devices [9]. Future investigations across extended implantation timelines will be necessary to test this hypothesis.

## 4.2 Synthetic data - spike detection

Following the validation of in vivo signal quality and longitudinal stability using the NeuralGlider, the next crucial requirement for downstream closed-loop neurostimulation is the accurate detection and extraction of relevant neural features, such as individual action potentials, from continuous extracellular recordings. To enable real-time closed-loop applications, spike detection algorithms must reliably isolate these action potentials from background noise while maintaining minimal computational overhead. Accordingly, this section evaluates and compares three candidate spike detection strategies using synthetic extracellular datasets. The objective was to validate detection accuracy under controlled baseline conditions and confirm that dynamic thresholding provides precise event detection, establishing a reliable foundation for feature extraction and online single-unit sorting.

The comparative evaluation of spike detection algorithms was conducted using synthetic data with controlled baseline noise. Consequently, all three approaches (simple amplitude thresholding (Figure 4.5A), static Amplitude Slope Optimization (ASO) thresholding (Figures 4.5B) and ASO with an adaptive dynamic threshold (Figures 4.5C)) achieved precise visual alignment with synthetic neural spikes, once their parameters ($k_a = 4$, $c = 100$, $m = 256$) were properly tuned. This test validates the core implementation of each algorithm under ideal baseline conditions.

Superimposing the extracted waveform templates (Figure 4.5D, top) demonstrates two morphologically distinct neural units, differentiated primarily by peak amplitude and repolarization kinetics. As shown in the bivariate feature distributions (Figure 4.5D, bottom), $k$-means clustering applied to paired temporal metrics (peak-to-trough duration, half-width) and geometric metrics (amplitude, slope) provides strong linear separability between Neuron 1 and Neuron 2. This distinct separation confirms that combining ASO detection with adaptive thresholding provides precise temporal alignment of detected events, enabling accurate feature extraction and robust single-unit classification for real-time closed-loop applications.

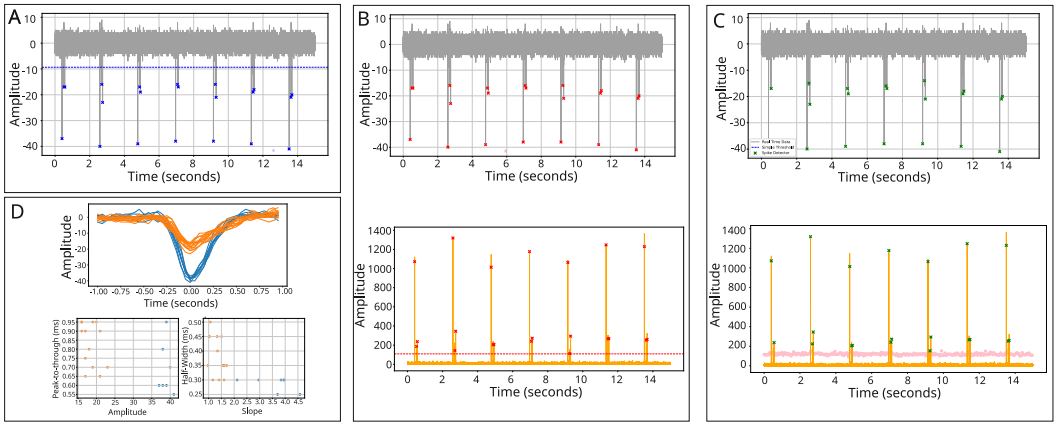



Figure 4.5: Comparative evaluation of spike detection methods operating on a single-channel simulated extracellular voltage trace, alongside template matching and spike feature characterization. **(A) Simple voltage threshold detection:** Continuous extracellular trace (gray) overlaid with a fixed amplitude threshold (blue dashed line) and detected spike events (blue markers). **(B) Static ASO thresholding: Top:** Signal trace with spike events detected using the static Amplitude Slope Operator (ASO) algorithm (red markers). **Bottom:** Transformed ASO signal (orange trace) with peak events (red markers) exceeding the fixed detection threshold (red dashed line). **(C) Adaptive dynamic ASO thresholding: Top:** Integrated detection pipeline showing validated spike detections (green markers) derived from the ASO algorithm combined with an adaptive noise-based threshold. **Bottom:** Transformed ASO signal (orange trace) with detected peaks (green markers) evaluated against the dynamic noise-adaptive threshold (pink trace). **(D) Waveform clustering and feature extraction: Top:** Extracted spike waveforms detected via dynamic ASO thresholding, categorized into two distinct neural clusters using $k$-means clustering: Neuron 1 (blue, $n = 7$) and Neuron 2 (orange, $n = 13$). **Bottom:** Bivariate scatter plots illustrating cluster separation between Neuron 1 (blue) and Neuron 2 (orange) across amplitude vs. peak-to-trough duration (left panel) and slope vs. half-width (right panel).

## 4.3 In vitro data - artifact removal and spike detection

To address the challenge of signal contamination during optical stimulation, this section systematically evaluates a multi-stage signal conditioning framework with the objective of quantifying the progressive suppression of optogenetic artifacts across each processing stage and demonstrating that effective artifact elimination preserves genuine firing dynamics. Accurate single-unit isolation requires spike detection pipelines to reliably separate biological action potentials from non-biological interferences. However, extracellular recordings during optogenetic

experiments are routinely corrupted by high-amplitude photoelectric artifacts with spectral profiles overlapping neural action potential bands.

As shown in Figure 4.6A, large-amplitude excursions in the continuous trace, caused by LED onset and offset transitions, trigger frequent threshold crossings that do not conform to physiological action potential kinetics. Subsequent template alignment (Figure 4.6B) and feature space mapping (Figure 4.6C) effectively partition these spurious signals from genuine neural activity. Cluster analysis clearly identifies two distinct signal populations: genuine single-unit activity (Cluster 2, $n = 400$, orange) and a high-amplitude artifact population (Cluster 1 / Artifact, $n = 108$, blue). Genuine action potentials display characteristic physiologically constrained waveform profiles with consistent peak-to-trough durations and steep depolarization slopes, whereas the artifact cluster is characterized by abnormally large negative amplitude swings (reaching $-150\,\mu$V) and distorted non-physiological kinetics. However, the separation is not entirely absolute. Due to partial feature overlap near the cluster boundary, a small number of events near the interface are likely misclassified.

The distinct separation in the amplitude vs. peak-to-trough domain highlights the severe impact of non-biological artifacts when processing real-time filtered data. While genuine spikes (Cluster 2, orange) cluster in the lower-amplitude, shorter-duration region (bottom-left), artifact events (Cluster 1, blue) occupy a markedly distinct space characterized by elevated amplitudes and prolonged peak-to-trough durations (top-right). In contrast, the slope vs. half-width feature space proved less conclusive and offered little diagnostic utility. Instead of forming distinct clusters, the artifact events and genuine spikes are intermixed without a clear spatial pattern, demonstrating that these two metrics alone are insufficient to distinguish optical transients from neural activity.

Even the advanced detection approach of Amplitude Slope Optimization (ASO) paired with an adaptive noise threshold fails to exclude these events in unconditioned real-time data, as these high-amplitude noise transients exceed adaptive dynamic boundaries and are mistakenly detected as primary neural spikes. This highlights the absolute necessity of rigorous signal preprocessing prior to feature detection.

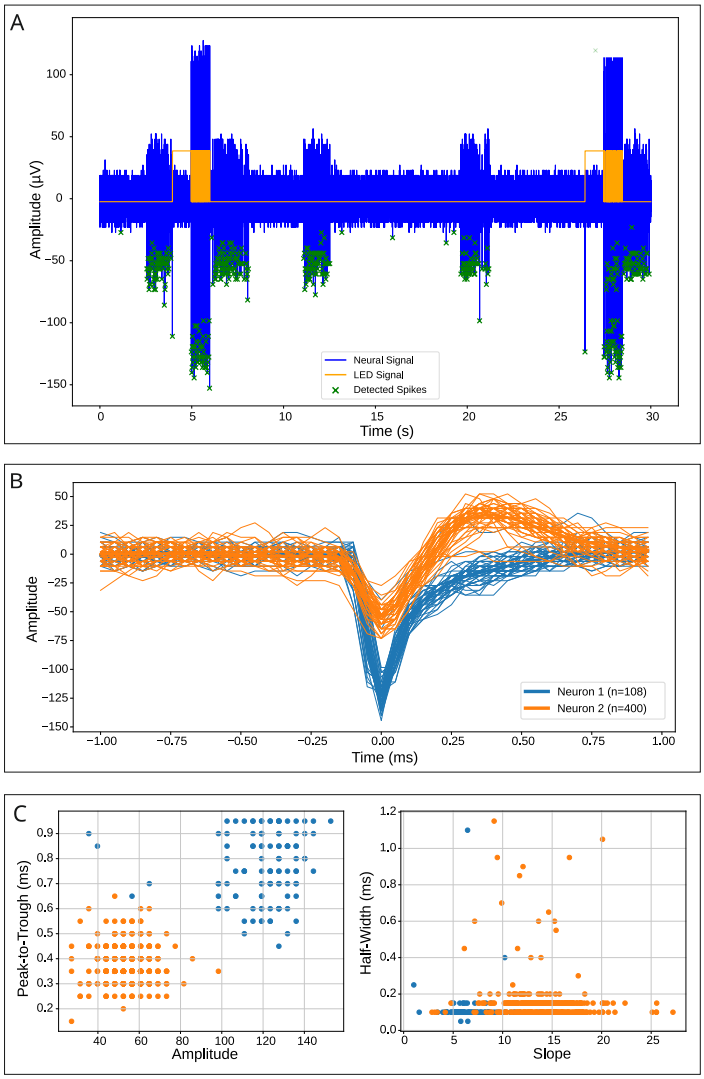



Figure 4.6: Overview of spike detection performance on real-time extracellular data, along with template shape comparisons and feature space distributions for extracted single units. (A) Spike Detection Window Overview: Continuous real-time neural signal (blue trace) overlaid with LED signal windows (orange trace) and detected spike events (green crosses). (B) Extracted Spike Waveform Templates: Superimposed action potential waveforms sorted into two distinct clusters using *k*-means clustering, displaying clear differences in peak depolarization amplitude and trough duration.(C) Feature Space Separation: Bivariate feature scatter plots showing cluster distribution between the 2 clusters, across Amplitude vs. Peak-to-Trough Duration (left panel) and Slope vs. Half-Width (right panel).

Optogenetic stimulation introduces correlated photoelectric artifacts across recording sites on multi-channel arrays, as addressed in Section 2.4.6. As demonstrated in the channel profile inspection (Figure 4.7A), LED switching triggers highly synchronized voltage transients across Channels 11 - 16. Because these photoelectric transients affect all recording channels simultaneously, with the same polarities and similar amplitudes, computing a Common Average (CA) provides an effective spatial representation of the shared noise floor.

Thus, subtracting the CA trace from individual signals significantly enhances recording quality. In the real-time recording (Figure 4.7B), large light-induced photovoltaic transients and elevated background fluctuations contaminate the neural baseline, leading to false-positive spike detections during LED onset and offset. Following CAR (Figure 4.7D), these shared transient artifacts are strongly attenuated, effectively flattening the baseline noise floor (Figure 4.7C) without distorting localized biological action potentials. Isolating activity from widespread optical interference stabilizes adaptive detection thresholds and prevents artifact-driven trigger errors. However, residual voltage transients remain around LED activation and deactivation transitions, underscoring the need for further algorithmic refinement to achieve complete artifact elimination.

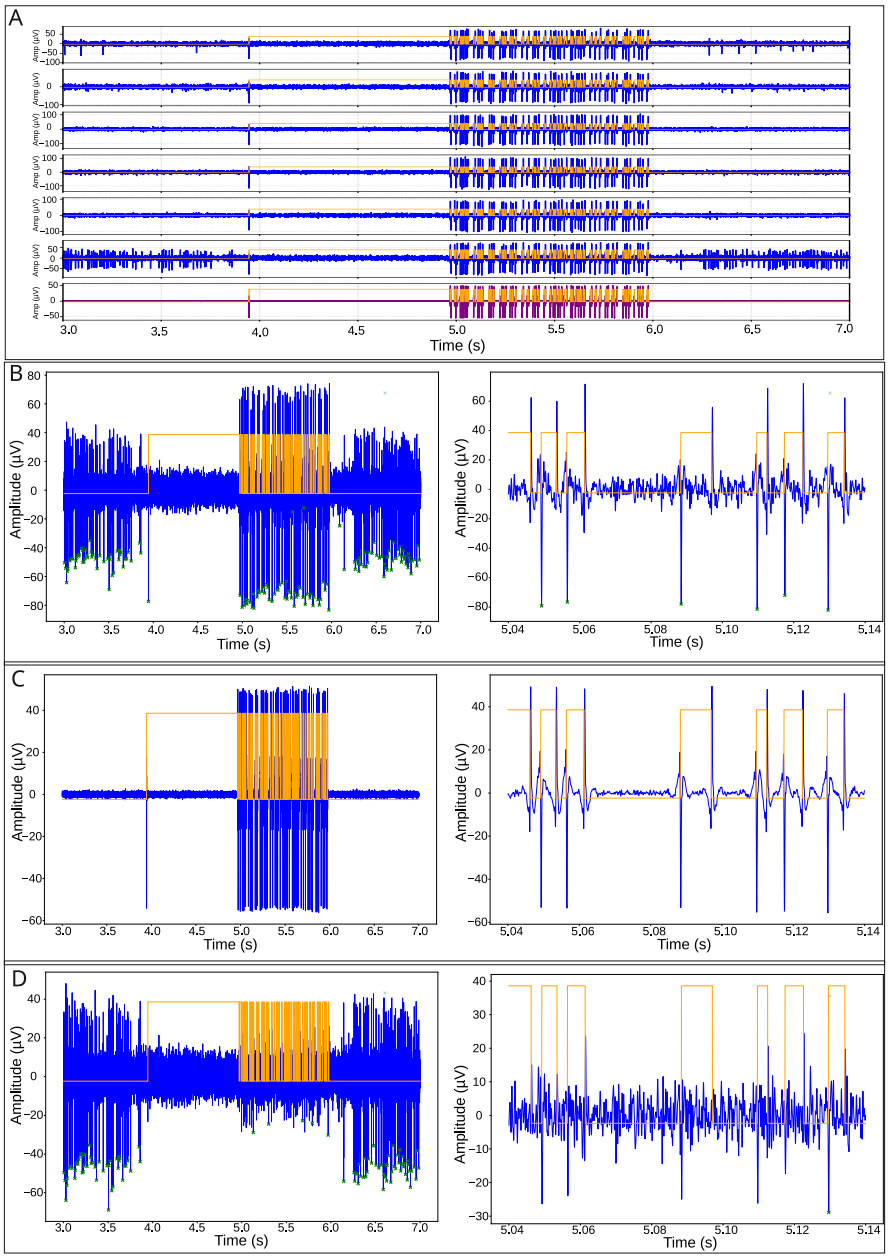



Figure 4.7: Evaluation of Common Average across all channels during optogenetic LED stimulation. **(A) Array Matrix Inspection Profile**: Multichannel neural traces (blue; Channels 11–16, top to bottom) shown alongside the calculated Common Average signal (purple, bottom), highlighting shared optical switching artifacts during LED pulses (orange overlay). **(B) Raw Channel 16 Traces**: Global evaluation of raw Channel 16 across a 4-second window (left; 3.0 - 7.0 s) alongside a high-resolution window (right; 5.04 - 5.14 s). **(C) CA Background Noise Floor**: Global evaluation of the CA noise floor across the 4-second window (left; 3.0 - 7.0 s) alongside the high-resolution window (right; 5.04 - 5.14 s). **(D) CAR-Cleaned Channel 16**: Global evaluation of CAR-cleaned Channel 16 across the 4-second window (left; 3.0 - 7.0 s) alongside the high-resolution window (bottom right; 5.04 - 5.14 s), demonstrating reduced artifact amplitudes.

While Common Average Referencing (CAR) significantly attenuates widespread photoelectric noise, fine-tuning detection parameters ($k_a$, $C$, $M$) remains essential to resolve local neural activity near LED switching transitions. Manual parameter settings reveal a clear trade-off between sensitivity and specificity. Permissive thresholds prioritizing maximum spike yield ($k_a = 8, C = 50, M = 32$) incorrectly classify residual photoelectric transients as neural events (Figure 4.8A, red markers), detecting 4645 events and 80 artifacts as spikes. Conversely, overly conservative parameters ($k_a = 4, C = 275, M = 128$) prioritize the elimination of all optical contamination (Figure 4.8B), but risk omitting low-amplitude biological spikes, detecting 592 spikes and 0 artifacts.

The balanced configuration ($k_a = 12, C = 250, M = 64$) successfully resolves this trade-off (Figure 4.8C - E), completely suppressing LED switching transients, while incurring only a slight loss of neural spikes in artifact-free regions. The extracted spike waveforms exhibit tight physiological alignment ($n = 805$, middle right), while high-resolution micro-profiles (bottom panels) confirm a consistent detection without false positives, even during rapid optogenetic stimulation trains.

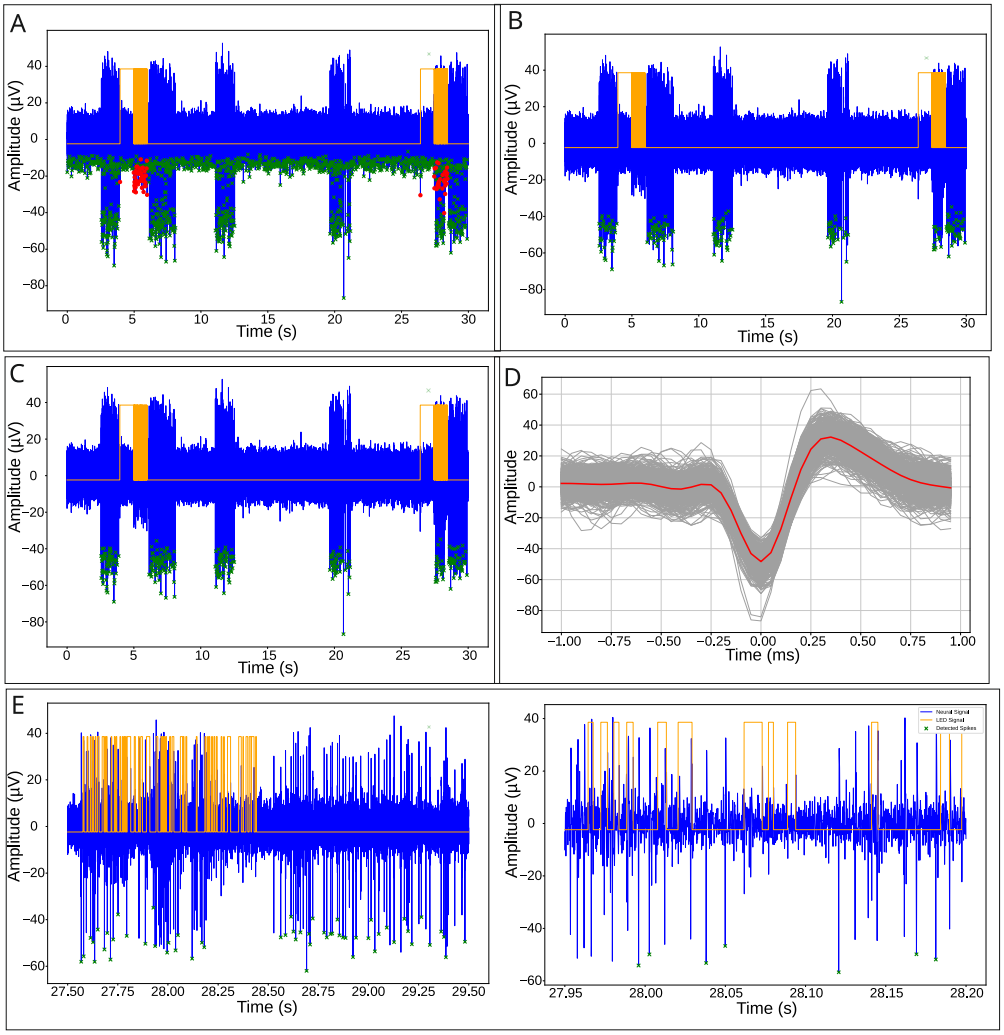



Figure 4.8: Algorithmic Parameter Optimization for Post-CAR Spike Detection and Optical Artifact Suppression. (A) Neural traces (blue) overlaid with LED signal (orange) for low detection threshold configuration ($k_a = 8, C = 50, M = 32$) maximizing true neural spike detection green points, but suffering from severe false-positive detections triggered by LED switching transients red points. (B) High threshold configuration ($k_a = 4, C = 275, M = 128$) eliminating all optogenetic switching artifacts, at the expense of reduced true spike sensitivity. (C) Top General Performance Profile: Optimal balanced parameter set ($k_a = 12, C = 250, M = 64$), achieving high real spike detection while completely rejecting LED switching noise across the full 0–30 s trace. (D) Action Potential Waveform Overlay: Mean template ($n = 805$ detected units) and individual waveforms extracted using the balanced parameter set, displaying physiological action potential kinetics. (E) Winner Micro-Profile: Zoomed-in views (27.50–29.50 s, left and 27.95–28.20 s, right) illustrating stable and artifact-free spike detection during high-frequency LED pulse trains.

Evaluating spike extraction metrics across sequential signal processing stages demonstrates how spatial referencing and parameter tuning collectively mitigate high-amplitude noise while preserving most physiological neural activity (Table 4.5). The initial real-time recording captured 1219 candidate events and 107 identified light-emitting diode (LED) artifacts (8.78% artifact rate) with a high voltage spread (Mean Artifact Amplitude $= 122.69\,\mu$V) compared to the clean spikes (Mean Clean $= 53.15\,\mu$V). Introducing Common Average Referencing (CAR) achieved substantial spatial artifact rejection. Because optogenetic switching artifacts corrupt the recording array concurrently, subtracting the instantaneous spatial ensemble average across channels isolated and eliminated correlated global noise. CAR drastically reduced the detected artifact count from 107 down to 10 (0.94% artifact rate). Subsequently, fine-tuning the extraction hyperparameters fully eliminated these remaining 10 LED artifacts (0.0% artifact rate). By increasing the absolute detection thresholding parameters, false-positive crossing events were completely rejected, leaving 805 clean, isolated action potentials with a tight amplitude distribution (SD $= 6.21\,\mu$V). The total elimination of artifact contamination ensures high signal purity and protects downstream sorting algorithms from false-positive unit classifications, at the cost of a minor reduction in true neural spikes within artifact-free regions

Table 4.5: Spike extraction metrics across signal processing stages, for in vitro data.

| | | | | **Amplitude** ($\mu$V) | | |
|---|---|---|---|---|---|---|
| **Stage** | **Clean Spikes** | **LED Artifacts** | **Artifact Rate (%)** | **Overall Std** | **Mean Artifact** | **Mean Clean** |
| Raw Signal | 1112 | 107 | 8.78 | 22.1260 | 122.6904 | 53.1474 |
| CAR | 1050 | 10 | 0.94 | 7.2122 | 27.9376 | 46.7075 |
| Optimized Parameters | 805 | 0 | 0.00 | 6.2050 | 0.0000 | 48.3274 |

To evaluate the temporal dynamics of the isolated neuronal activity, firing rate analysis was performed on a single representative channel (Channel 16). Following optimal parameter configuration, the temporal profile (Figure 4.9) reveals clear firing rate modulations across the session, confirming that post-CAR processing preserves underlying biological spiking patterns without significant signal attenuation.

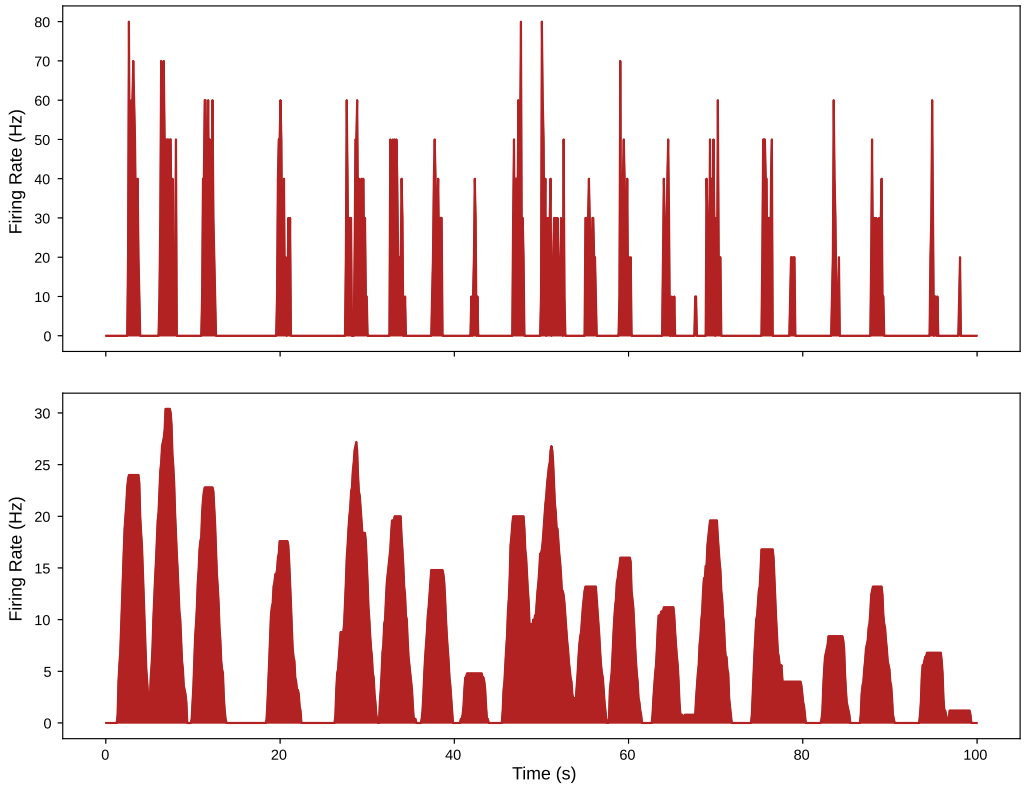



Figure 4.9: Single-channel firing rate dynamics on Channel 16. (Top) Unsmoothed raw firing rate computed across the recording session (100 ms). (Bottom) Smoothed firing rate profile calculated using a 25-sample moving average filter.

Overall, the integration of Common Average Referencing with targeted parameter optimization completely eliminated light-induced photoelectric artifacts (0.00% residual artifact rate), while maintaining high single-unit yield. Consequently, this pipeline provides a validated artifact-free foundation for robust feature extraction and real-time closed-loop neural decoding.

## 4.4 Real time spike detection and sorting

To enable activity-dependent neurostimulation and real-time decoding, electrophysiological pipelines must execute spike detection, feature extraction and multi-unit classification under strict latency budgets. While offline frameworks achieve high accuracy by processing whole-session recordings using multi-pass optimization and KS4, online systems must operate causally on continuous streaming data frames. Therefore, this section evaluates the real-time performance of the detection and sorting pipeline across synthetic, in vitro and in vivo datasets.

The objective was to benchmark algorithmic execution latency against physical streaming constraints, quantify online spike detection and unit-sorting trade-offs against offline references and validate spatial-temporal scalability across high-density probe arrays.

### 4.4.1 In vitro data

The real-time spike detection pipeline was first validated on the in vitro dataset. Utilizing the ASO algorithm allows to isolate high-slope voltage deflections characteristic of genuine action potentials (Figure 4.10B). As demonstrated in Figure 4.10A, the threshold boundary separates high-energy spike transients from background baseline noise.

To optimize the real-time execution of the spike detection pipeline, the algorithm was benchmarked across a range of chunk sizes ($N_{\text{samples}} \in [16, 512]$) (Table 4.6). Evaluating latency performance against available processing budgets is critical for closed-loop neural interfaces, where data must be processed faster than the physical duration of each incoming frame. Across all tested buffer configurations, the detection performance remained completely invariant, consistently identifying 1,066 clean spikes and falsely triggering on 3 LED artifacts (Table 4.6). Crucially, every chunk size comfortably met its respective time budget, with average processing times staying well below the maximum allowable duration per frame (Table 4.6). However, $N_{\text{samples}} = 32$ represented the optimal trade-off point between computational overhead and system responsiveness.

At $N_{\text{samples}} = 16$, repeating the Python loop thousands of times creates heavy overhead, raising the average processing time to 0.3775 ms (47.2% of its 0.8000 ms budget) (Table 4.6). Increasing the buffer size to $N_{\text{samples}} = 32$ resolves this bottleneck, since it cuts down on loop repetitions and lets the CPU process data points together in parallel batches, achieving the fastest computation time of 0.2602 ms per chunk (16.3% of its 1.6000 ms budget) (Table 4.6). For larger chunks ($N_{\text{samples}} \geq 64$), computation time increases because each chunk carries a heavier data workload. More importantly, larger buffers force the system to wait longer just to collect the data before processing can start, jumping from 1.6 ms for 32 samples to 25.6 ms for 512 samples (Table 4.6). Therefore, $N_{\text{samples}} = 32$ provides the best trade-off, minimizing overall hardware delay while leaving an 83.7% CPU safety margin for reliable real-time execution.

Extending this evaluation across the entire electrode array via a real-time firing rate heatmap (Figure 4.10C) provides crucial spatial context. While individual channel traces confirm local detection precision, the population heatmap reveals distinct spatial heterogeneity across the probe. Specific channel sub-groups (e.g., channels around 10–15) display synchronized firing rate elevations, whereas adjacent regions remain silent. Tracking array-wide spatiotemporal

activity in real time validates the scalability of the detection pipeline, proving its utility for multi-channel closed-loop optogenetic applications.

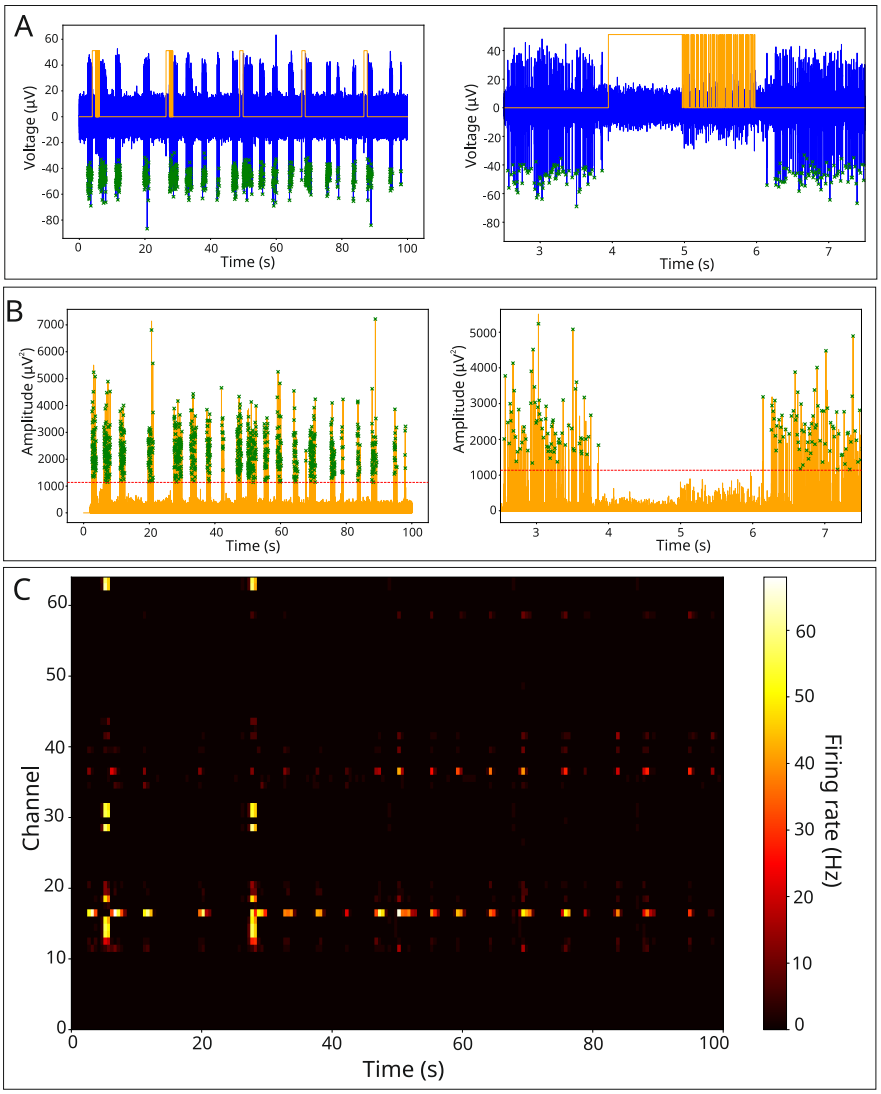



Figure 4.10: Evaluation of real-time detection performance on Channel 16 alongside multi-channel population activity across a 100 s optogenetic recording session. **(A) Real-time spike detection on Channel 16:** Full session (left) and high-resolution (0 - 10 s, right) voltage traces displaying filtered neural signals, auxiliary LED stimulation pulses (orange steps) and real-time detected spikes (green points). **(B) Amplitude Slope Operator (ASO) profile on Channel 16:** ASO metric profile across 0 - 100 s (left) and 0 - 10 s (right), where the fixed detection threshold (red dashed line) isolates high-frequency spike slopes (green crosses) from baseline activity. **(C) Real-time firing rate population heatmap:** Spatiotemporal population firing rate across all recording channels (Channels 1 - 64) over the full 100 s session.

Table 4.6: Real-time processing latency (average execution time per data chunk), execution budget (chunk time duration) and total spike detection yield across the entire recording for different buffer chunk sizes using in vitro data.

| $N_{\text{samples}}$ | Avg Chunk Time (ms) | Time Budget (ms) | Clean Spikes | LED Artifacts |
|---|---|---|---|---|
| 16 | 0.3775 | 0.8000 | 1066 | 3 |
| 32 | 0.2602 | 1.6000 | 1066 | 3 |
| 64 | 0.2647 | 3.2000 | 1066 | 3 |
| 128 | 0.3065 | 6.4000 | 1066 | 3 |
| 256 | 0.3256 | 12.8000 | 1066 | 3 |
| 512 | 0.4494 | 25.6000 | 1066 | 3 |

To evaluate online detection accuracy, the real-time pipeline was benchmarked against the offline adaptive moving median detector on Channel 16, a representative channel (Figure 4.11). The real-time detector captured 1069 events compared to 805 detected by the offline method (Table 4.7), while demonstrating strong overall alignment across the 100-second session (Figure 4.11, top). Crucially, 805 of these detections were confirmed mutual events (green circles), with zero events unique to the offline method, demonstrating that the online RMS calibration identified every neural spike detected by the offline benchmark.

The yield discrepancy (264 events unique to the real-time pipeline, blue markers) highlights the distinct algorithmic trade-offs between static calibration and dynamic moving-median thresholding. The offline benchmark continuously updates its threshold using a short trailing window to track local noise floor fluctuations. While this dynamic adaptation maximizes conservative artifact rejection, high-density spike bursts temporarily elevate the local median, raising the threshold and causing the detector to filter out lower-amplitude or closely spaced action potentials. Conversely, the real-time pipeline employs a static threshold established during initial calibration. This fixed threshold maintains sensitivity during intense firing bursts, maximizing action potential capture during rapid neural events at the cost of higher overall event counts and decreased selectivity against transient in-band artifacts.

Furthermore, performance during active LED optical pulsing (25-30 s) illustrates how both approaches handle photostimulation artifacts (Figure 4.11, bottom). While spatial Common Average Referencing (CAR) successfully suppressed the vast majority of optogenetic noise across both pipelines, the fixed real-time threshold registered 3 false positives at sharp stimulus transitions (cyan markers), whereas the adaptive offline threshold successfully absorbed these transients. Ultimately, both strategies prove highly effective for their intended operational domains, with the real-time causal pipeline providing consistent low-latency sensitivity suitable for closed-loop control applications.

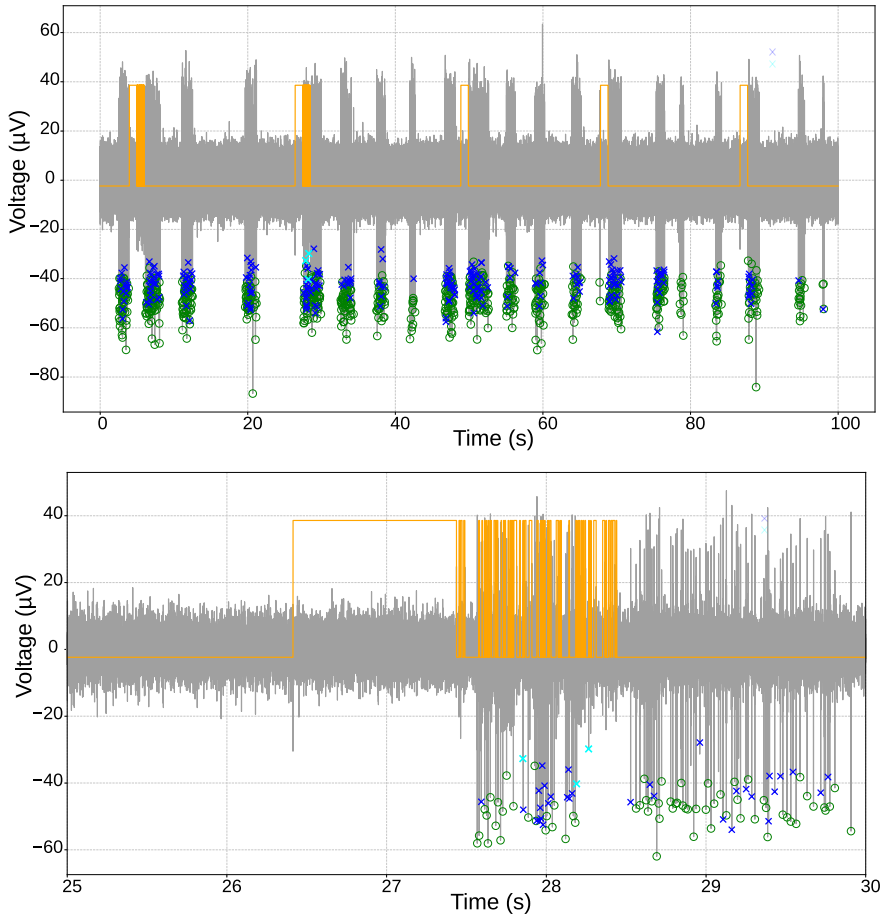



Figure 4.11: Performance comparison between the real-time spike detector and the offline adaptive moving median benchmark on Channel 16. (Top) Full 100-second continuous recording showing filtered neural signals (gray), optical stimulation intervals indicated by the LED context signal (orange), mutually detected clean spikes (green circles), real-time unique clean spikes (blue crosses) and real-time unique optical artifacts (cyan crosses). (Bottom) High-resolution 5-second window (25 - 30 s) illustrating detector behavior during dynamic optogenetic LED pulsing.

Table 4.7: Spike detection comparison report for in vivo data

| **Algorithm Strategy** | **Spikes Detected** |
|---|---|
| Real-Time Fixed Threshold | 1069 |
| Offline Adaptive Moving Median | 805 |
| **Cross-Comparison Breakdown** | |
| Confirmed Mutual Detections | 805 |
| Unique to Offline | 0 |
| Unique to Real-Time (Clean Spikes) | 261 |
| Unique to Real-Time (Artifacts) | 3 |

### 4.4.2 Synthetic data

Detecting that a spike occurred is only half of the problem. In most cortical or subcortical recordings, a single electrode picks up activity from more than one neuron in its vicinity, so the detected events are a mixture of different cells firing at different times. Spike sorting is the step that tries to disentangle this mixture, assigning each detected event to the neuron that most likely produced it, based on the shape of its waveform. This distinction could be an improvement for closed-loop systems designed to investigate activity-dependent plasticity and targeted neural modulation, as described in section 2.4.8.

To evaluate online multi-unit isolation capabilities, the real-time pipeline was integrated with 1-feature $k$-means classification on the synthetic dataset. (Figure 4.12). The continuous recording trace confirms that the fixed-threshold ASO metric reliably separates spikes from background noise over time, while the two distinct marker colors demonstrate clear amplitude-based cluster separation across the entire session (Figure 4.12A). The waveforms profiles validate the sorting accuracy, separating the detected activity into two distinct unit clusters: Unit 0 ($N = 91$, red) and Unit 1 ($N = 45$, blue) (Figure 4.12B). Unit 0 corresponds to smaller-amplitude action potentials averaging a -20 $\mu$V trough, whereas Unit 1 captures higher-energy units with deep negative deflections around -40 $\mu$V. The sharp separation between waveform templates highlights that combining causal ASO detection with a K-Means model achieves low-latency multi-unit sorting without exceeding real-time streaming constraints.

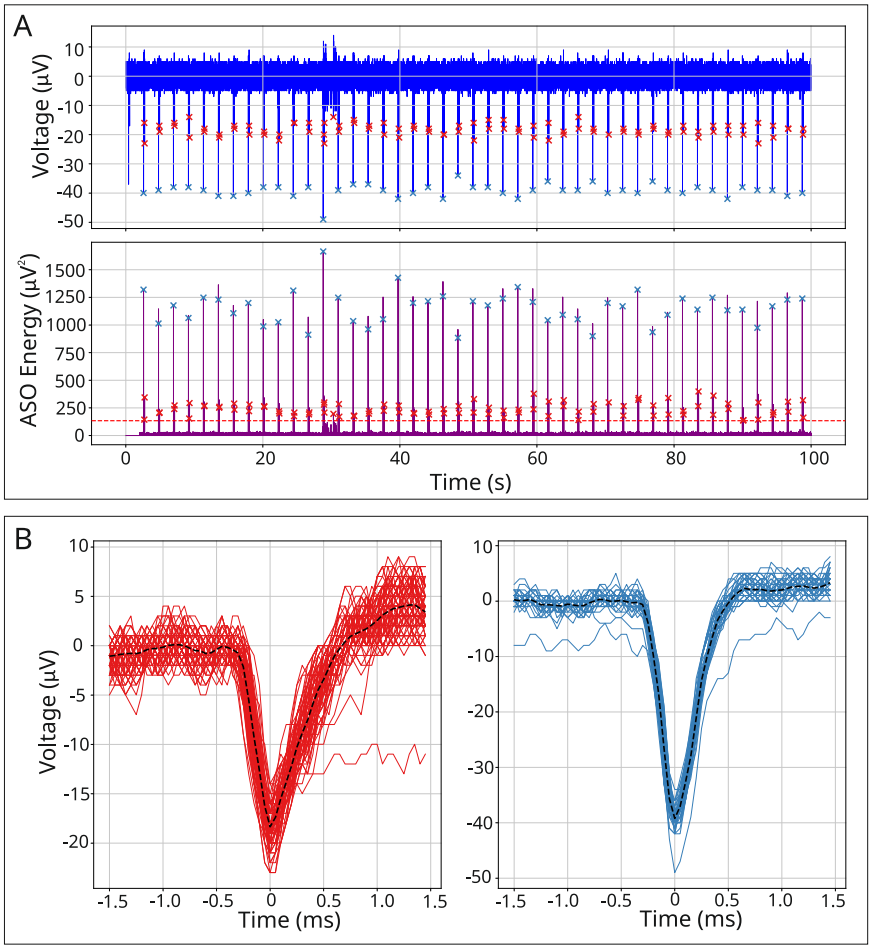



Figure 4.12: Real-time single-feature $k$-means spike sorting evaluated on Channel 5 of the synthetic dataset. **(A) Spike detection and cluster assignment: Top:** Filtered neural voltage trace over a 100 s session displaying detected action potentials classified into distinct clusters: Unit 0 (red markers) and Unit 1 (blue markers). **Bottom:** Corresponding Amplitude Slope Operator (ASO) metric trace evaluated against a fixed detection threshold of 134.1 (red dashed line). **(B) Extracted unit waveform profiles:** Isolated spike waveforms for Unit 0 ($n = 91$, left) and Unit 1 ($n = 45$, right), illustrating clear physiological segregation by amplitude and waveform morphology.

Evaluating the real-time 2-feature classification pipeline (Amp + P2T) on synthetic data demonstrates how combining signal magnitude with biophysical waveform morphology impacts online unit segregation (Figure 4.13). The top and middle panels show continuous spike detection and unit tracking across the 100-second session (Figure 4.13). The fixed ASO energy threshold successfully isolates action potentials above background noise. Throughout most of the trace, red cross markers map to lower-amplitude events ($\approx -18\,\mu$V) while blue cross

markers track larger-amplitude spikes ($\approx 40\,\mu$V). The extracted waveform profiles support this physical separation, grouping the activity into Unit 0 ($N = 90$, red) for moderate depolarizations and Unit 1 ($N = 46$, blue) for high-amplitude deflections around - $40\,\mu$V. The feature space matrix illustrates the underlying cluster distribution driving these predictions (Figure 4.13, bottom). Peak amplitude acts as the dominant discriminant, separating events centered at $18\,\mu$V (Unit 0) and $40\,\mu$V (Unit 1). Adding the Peak-to-Trough (P2T) ratio provides a secondary morphological axis. While Unit 1 maintains a tightly constrained lower P2T ratio ($\approx 0.1$), Unit 0 spans a broader range (0.2 - 0.5).

However, comparing these outcomes to the 1D K-Means amplitude sorter reveals a key trade-off when introducing shape-dependent metrics in a streaming environment. While the 2-feature approach incorporates biophysical shape information, it increases sensitivity to localized signal corruptions. Specifically, at $t \approx 28\,$s, a low-amplitude event was misclassified as Unit 1 (blue markers) despite its small trough, probably because of a distortion on the post-trough peak caused by a high-frequency noise burst. This noise is further reflected in the Unit 0 waveform profiles, which retain visible trailing outliers that deviate from the ensemble mean. In contrast, 1D K-Means avoids shape-based distortions entirely, maintaining strict amplitude boundaries at the cost of discarding non-amplitude morphological traits.

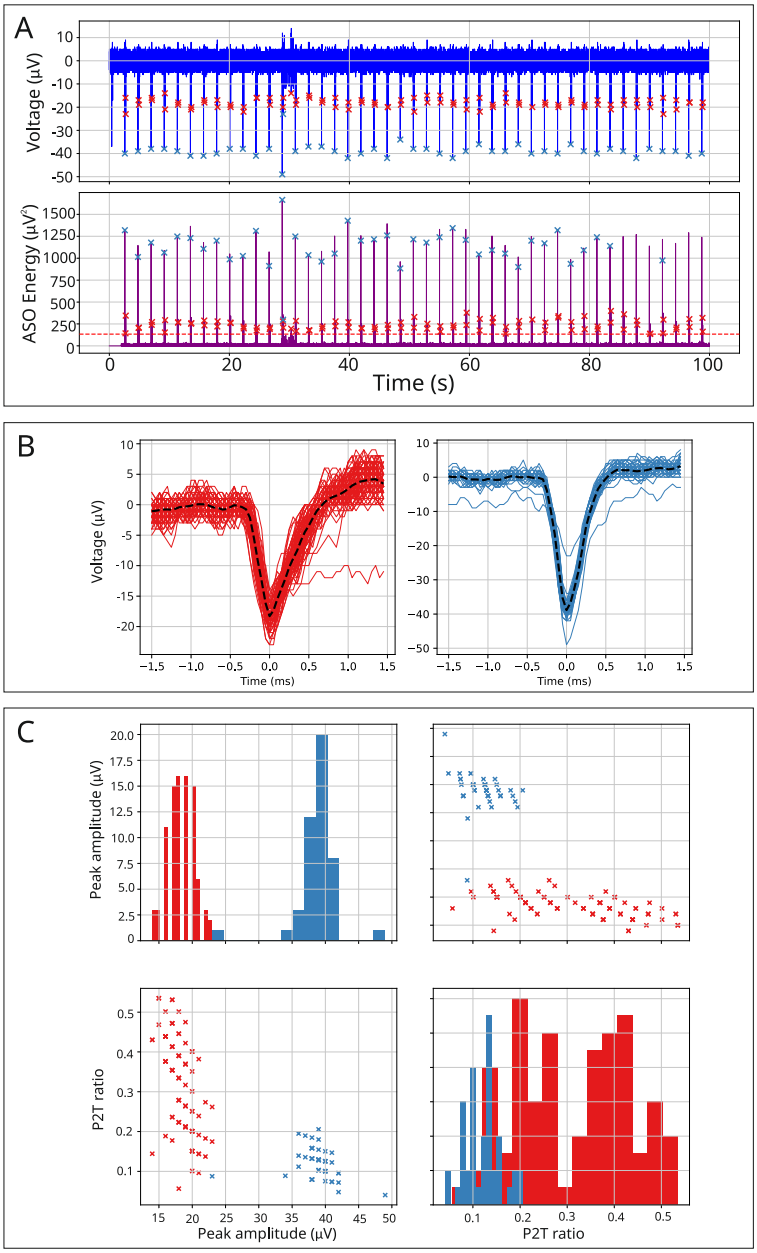



Figure 4.13: Real-time two-feature spike sorting evaluated on Channel 5 of the synthetic dataset. **(A) Spike detection and cluster assignment: Top:** Filtered continuous voltage stream displaying isolated spikes classified into Unit 0 (red markers) and Unit 1 (blue markers). **Bottom:** Corresponding Amplitude Slope Operator (ASO) metric trace evaluated against a fixed detection threshold (134.1). **(B) Extracted unit waveform profiles:** Isolated spike waveforms for Unit 0 ($n = 90$, left) and Unit 1 ($n = 46$, right), displaying distinct morphological separation. **(C) Feature space distribution and marginal density:** Pairwise feature space scatter matrix and marginal histograms comparing peak amplitude ($\mu$V) and peak-to-trough (P2T) ratio.

Evaluating the 3-feature pipeline (Peak Amplitude, P2T Ratio, T2P Duration) on Channel 5 demonstrates how scaling feature dimensionality influences multi-unit classification (Figure 4.14). The top continuous voltage trace and ASO metric track show stable spike detection across the 100-second session. The 3-feature model reverses the unit indexing relative to prior sorters. Here, Unit 0 ($N = 46$, red) represents the high-amplitude spike population ($\approx -40\,\mu$V), while Unit 1 ($N = 90$, blue) captures the lower-amplitude events ($\approx -18\,\mu$V). Despite the label swap, the population counts (46 vs. 90) remain strictly aligned with the spike counts observed in the 2-feature sorting (46 vs. 90). Similar to the 2-feature sorter, the high-frequency artifact at $t \approx 28.5$ s corrupts time-domain shape estimators. In the continuous trace, a low-amplitude event during this burst is misclassified into the large-amplitude Unit 0 cluster (red markers at $-50\,\mu$V drop). Pure 1D K-Means avoids this misassignment because it evaluates peak energy without relying on waveform duration or inflection points.

The isolated waveform plots reveal that adding Trough-to-Peak (T2P) Duration (Width) does not resolve the trailing shape outliers seen in the 2-feature model. The pairwise matrix highlights that Peak Amplitude remains the primary separator, creating non-overlapping clusters along the amplitude axis. While P2T Ratio and T2P Duration provide extra biophysical dimensions, their marginal distributions overlap significantly. While adding a third feature increases computational overhead without providing orthogonal variance, the 2-feature space offers an optimal balance between morphological richness and efficiency. Nevertheless, for this synthetic dataset, the 1D K-Means sorter achieves the most accurate performance, as pure amplitude thresholding works better to separate the clean waveforms.

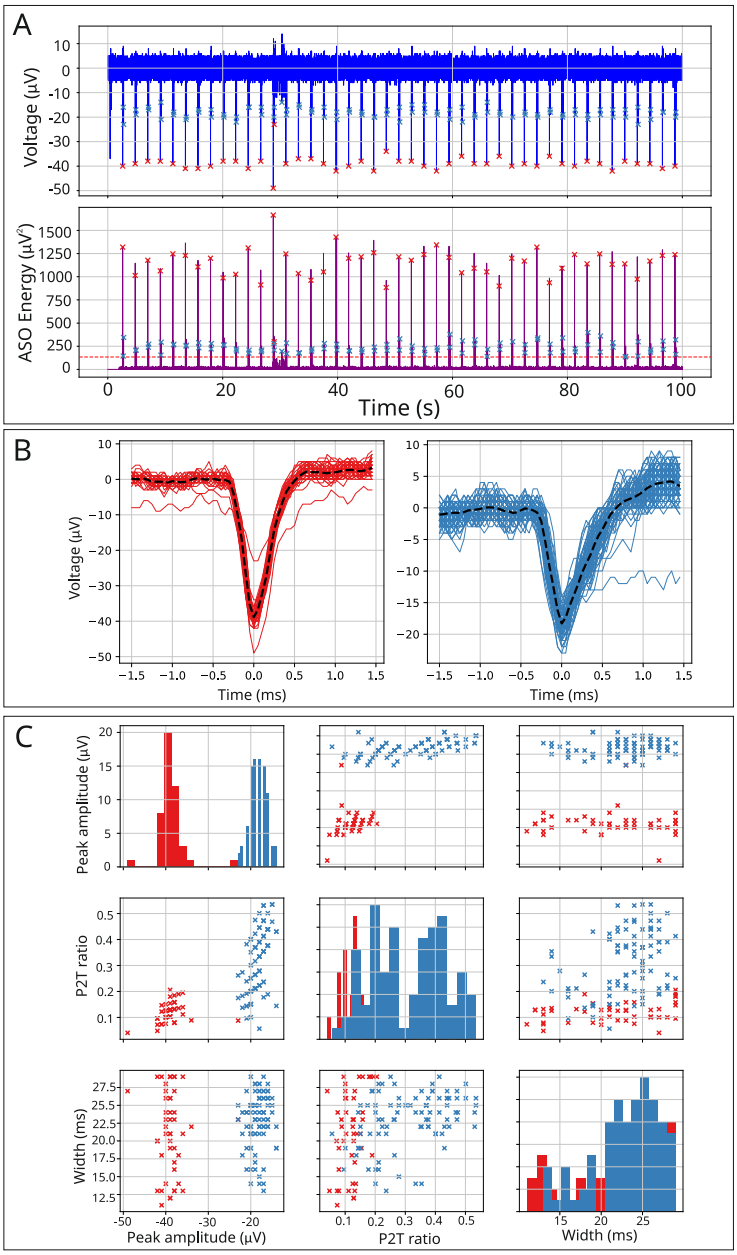



Figure 4.14: Real-time three-feature spike sorting evaluated on Channel 5 of the synthetic dataset. **(A) Spike detection and cluster assignment: Top:** Filtered continuous voltage trace showing isolated spikes categorized into Unit 0 (red markers) and Unit 1 (blue markers). **Bottom:** Amplitude Slope Operator (ASO) detection trace evaluated against a fixed threshold (134.1). **(B) Extracted unit waveform profiles:** Waveform profiles displaying Unit 0 ($n = 46$, high amplitude, left) and Unit 1 ($n = 90$, low amplitude, right), illustrating clear morphological separation. **(C) Feature space matrix and marginal distributions:** Pairwise $3 \times 3$ feature scatter matrix and marginal distributions for peak amplitude ($\mu$V), peak-to-trough (P2T) ratio and trough-to-peak duration (width).

### 4.4.3 In vivo data

Evaluating the real-time ASO detection framework on in vivo data highlights the sensitivity trade-offs inherent to online single-channel thresholding compared to complex offline sorters (Figure 4.15). Across both Channel 413 and Channel 14, the streaming ASO engine establishes stable energy thresholds ($\approx 7000$ and $\approx 9000$, respectively) after the initial calibration period (10s), where ka and c values are defined. The ASO metric tracks reliably capture rapid energy transients, flagging candidate action potentials in real time.

However, comparing streaming detections against the offline Kilosort4 (KS4) benchmark reveals clear differences in detection philosophy and sensitivity. The real-time ASO detector identifies a significantly higher density of candidate spikes across the 100-second window than KS4. On Channel 413 ($k_a = 6, c = 300$), ASO triggers on numerous low-amplitude voltage deflections that fall close to the local noise floor. On Channel 14 ($k_a = 12, c = 350$), this tendency is even more pronounced due to the higher background activity. Consequently, the threshold calculated from the 10-second calibration window proved too low for the overall recording. Because this brief calibration segment underrepresented long-term noise variance, it led to elevated false-positive rates over time.

Kilosort4 employs multi-channel spatial drift correction, template matching and global graph-based optimization offline, allowing it to aggressively filter out marginal events and retain only high-confidence unit firings [70]. In contrast, the single-channel ASO algorithm operates causally with zero lookahead, prioritizing zero-latency detection over selectivity. Consequently, while the real-time ASO engine guarantees low-latency event triggering, these results demonstrate that single-channel thresholding captures a broader spectrum of sub-threshold activity than offline pipelines. As expected, a real-time causal detector cannot achieve the same level of selectivity as a full offline pipeline like Kilosort4, making subsequent feature filtering or classification essential to reject false-positive noise or MUA triggers prior to decoding.

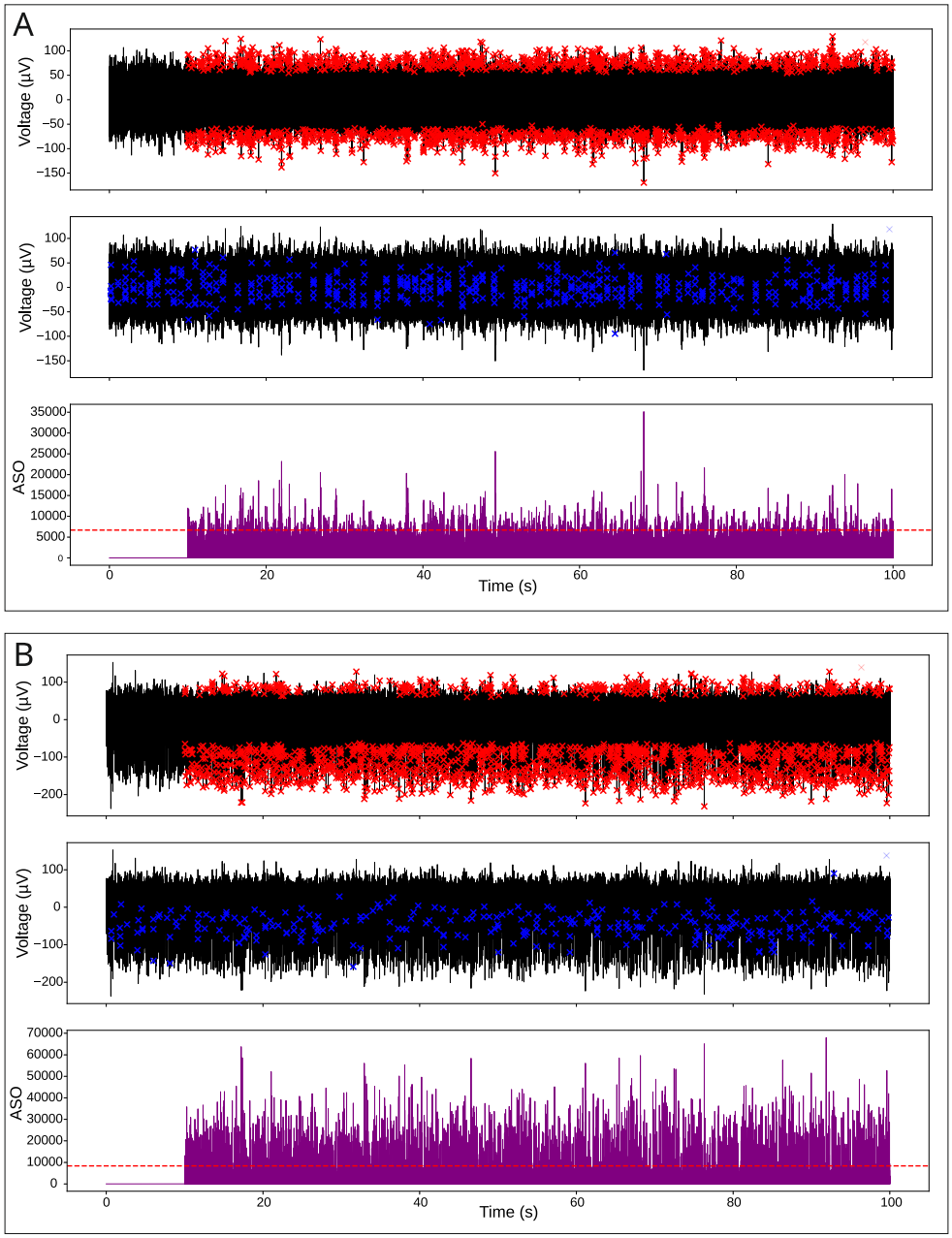



Figure 4.15: Validation of real-time Amplitude Slope Operator (ASO) spike detection against offline Kilosort4 (KS4) benchmarks across two representative channels of the NeuralGlider dataset. **(A) Channel 413 performance** ($k_a = 6$, $c = 300$)**: Top:** Filtered neural trace with real-time ASO detections (red $\times$). **Middle:** Offline KS4 benchmark spike annotations (blue $\times$). **Bottom:** Streaming ASO energy metric trace evaluated against its dynamic threshold. **(B) Channel 14 performance** ($k_a = 12$, $c = 350$)**: Top:** Filtered neural trace with real-time ASO detections. **Middle:** Corresponding offline KS4 benchmark spike annotations. **Bottom:** Streaming ASO energy metric trace. Note that offline KS4 benchmark spike timestamps are not trough-aligned like the real-time ASO detections.

Quantitative evaluation against Kilosort4 highlights this operational divergence across the five channels with the highest amplitude on "good" units (Table 4.8). Across channels 14, 413, 871, 875 and 880, the single-channel ASO framework yielded high nominal false-positive rates relative to KS4's template-matched output, resulting in low overall $F_1$ scores (0.000-0.138). For example, on Channel 14, ASO detected 1971 candidate events against KS4's 280 ground-truth units, yielding 155 true positives (Recall = 0.554), 1816 false positives (Precision = 0.079) and 125 false negatives. Crucially, the KS4 ground truth was restricted to curated single units to establish a rigorous benchmark for evaluating detector performance on well-isolated events, a big advantage for downstream classification and high-fidelity closed-loop control, as described in section 2.4.8. Consequently, these high false-positive counts do not simply represent baseline noise but also surrounding multi-unit activity (MUA) and lower-amplitude spikes that KS4 deliberately discards during single-unit sorting. Furthermore, the causal detector struggled to capture a significant portion of ground-truth single-unit firings, highlighting a key area for future improvement.

Despite capturing a broader spectrum of candidate events, the streaming engine consistently met strict hardware execution limits. Achieving average processing times of 0.14 ms per chunk on standard channels (and 1.16 ms on Channel 871), the pipeline remained well within the 1.60 ms real-time latency budget. These metrics confirm that despite trading strict single-unit selectivity for multi-unit sensitivity, single-channel ASO's low computational overhead guarantees stable execution for real-time applications.

Table 4.8: Spike detection validation metrics and real-time processing latency per channel benchmarked against KS4. Performance metrics include channel ID, spike counts detected by the real-time algorithm versus KS4, true positives (TP), false positives (FP), false negatives (FN) and $F_1$ score. Real-time execution feasibility is quantified by comparing the average running time per chunk against the total chunk duration.

| Channel | Detected | KS4 | TP | FP | FN | F1 | Avg Chunk (ms) | Budget (ms) |
|---|---|---|---|---|---|---|---|---|
| 871 | 4239 | 3 | 1 | 4238 | 2 | 0.000 | 1.16 | 1.60 |
| 14 | 1971 | 280 | 155 | 1816 | 125 | 0.138 | 0.14 | 1.60 |
| 875 | 3887 | 533 | 114 | 3773 | 419 | 0.052 | 0.14 | 1.60 |
| 413 | 1896 | 512 | 72 | 1824 | 440 | 0.060 | 0.14 | 1.60 |
| 880 | 1367 | 643 | 48 | 1319 | 595 | 0.048 | 0.14 | 1.60 |

Comparing the single-channel and multi-channel sorting pipelines against offline Kilosort4 (KS4) ground truth reveals important structural and operational differences (Figure 4.16). The

single-channel Louvain sorter uses a 3-feature vector comprising peak amplitude, P2T ratio and T2P duration. On Channel 413, it splits candidate activity into two main populations, Unit 0 ($N = 298$, low amplitude) and Unit 1 ($N = 214$, high amplitude). While this matches the two primary units identified by KS4 (Units 42 and 43), evaluating the extracted waveform profiles reveals notable intra-class variance and trailing baseline noise. Single-channel morphology remains susceptible to local signal fluctuations, causing distinct firing events with similar trough depths to merge into single broad clusters.

By incorporating spatial footprints across a 7-channel neighborhood ($\pm 3$ channel radius centered on Channel 413), the multi-channel sorter evaluates relative voltage decay across adjacent probe sites. This extra spatial dimension over-segments Channel 413 activity into four spatial communities, Unit 0 ($N = 154$), Unit 1 ($N = 121$), Unit 2 ($N = 8$, low-yield artifact) and Unit 3 ($N = 229$). Rather than misclassifying spikes, the multi-channel pipeline separates sub-populations based on current spread across the shank.

Mapping the predicted clusters across all 1024 channels on the spike raster reveals that the multichannel sorter fails to form biologically meaningful units. The assigned clusters exhibit no temporal synchrony across channels, grouping non-simultaneous spikes across distant regions of the probe into the same cluster identity. While the single-channel Louvain sorter offers ultra-low latency and low computational overhead, it lacks the feature depth required to resolve overlapping waveforms on a single contact. Conversely, while the multi-channel Louvain sorter incorporates spatial geometry, it ultimately produces non-physiological clusters by grouping temporally uncorrelated spikes across distant channels. Consequently, neither pipeline achieves reliable real-time unit isolation under streaming conditions.

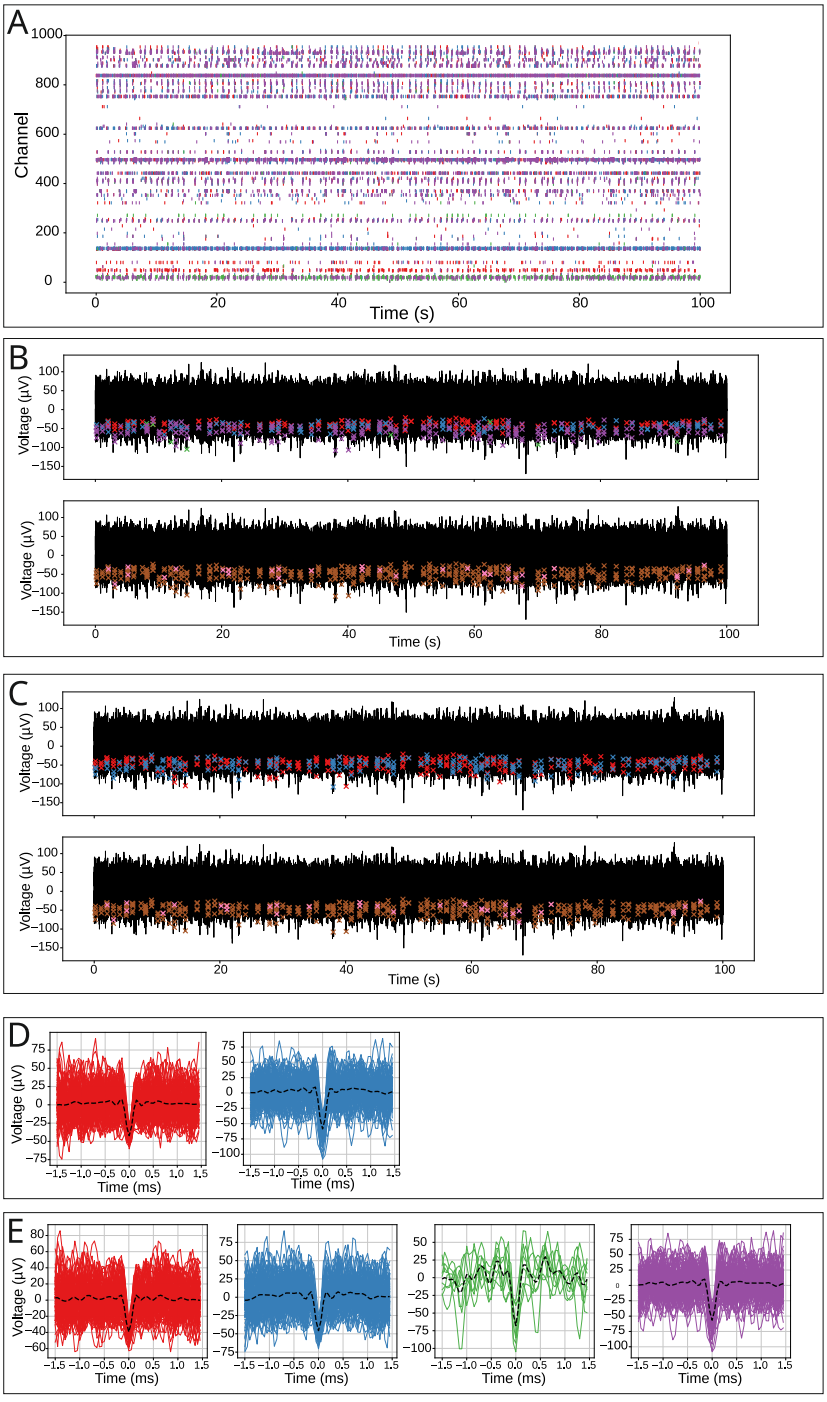



Figure 4.16: Real-time streaming Louvain community clustering vs. Kilosort4 (KS4) ground truth on in-vivo data Channel 413. (A) Real-time multichannel Louvain clustering spike raster plot across the entire 1024-channel array. (B) Comparative evaluation displaying multi-channel real-time cluster prediction (top) and Kilosort4 (KS4) offline ground-truth benchmark spikes (bottom). (C) Comparative evaluation displaying single-channel real-time cluster prediction (top) and Kilosort4 (KS4) offline ground-truth benchmark spikes (bottom). (D) Waveform profiles from the single-channel real-time clustering, displaying Unit 0 ($N = 298$, red) and Unit 1 ($N = 214$, blue). (E) Waveform profiles from the multi-channel real-time clustering, displaying Unit 0 ($N = 154$, red), Unit 1 ($N = 121$, blue), Unit 2 ($N = 8$, green) and Unit 3 ($N = 229$, purple).

To evaluate real-time performance metrics, numerical results between a single-channel Louvain clustering pipeline on channel 413 and a multichannel pipeline operating across all probe channels were compared (Table 4.9). Single-channel Louvain demonstrated strong targeted sorting performance, identifying 2 predicted units that directly matched the 2 local KS4 "good" units on Channel 413, with an accuracy of 58.3%. Conversely, multi-channel Louvain across all array heavily under-clustered the recording, finding only 4 global units out of 105 KS4 units, with a 10.2% accuracy, illustrating the challenge of real-time spatial feature extraction across high-density channels without offline processes. Crucially, both sorting pipelines maintained low average latencies ($\approx 0.14$ ms per chunk) within the 1.60 ms streaming budget, though single-channel sorting experienced higher variance during peak spike bursts (max $= 21.65$ ms). These findings highlight single-channel Louvain clustering as a promising approach for real-time unit isolation, contingent upon optimizations that reconcile sorting accuracy with strict low-latency constraints.

Table 4.9: Performance and processing latency of real-time single-channel (Channel 413) vs. multichannel Louvain spike sorting against Kilosort4. Quantitative metrics include single-unit detection counts for KS4 and the real-time pipeline, sorting accuracy relative to ground-truth KS4 labels and real-time processing latency statistics (mean, standard deviation and maximum latency per spike).

| **Metric** | **Channel 413** | **Multichannel** |
|---|---|---|
| KS4 "Good" Units | 2 | 105 |
| Predicted Units Found | 2 | 4 |
| Spikes Accuracy vs. KS4 | 58.3% | 10.2% |
| Mean Latency (ms) | 0.1480 | 0.1384 |
| Std Latency (ms) | 0.9518 | 0.0408 |
| Max Latency (ms) | 21.6540 | 5.6577 |

In summary, the ASO detection framework consistently operates well within strict real-time latency budgets ($< 1.60$ ms), providing a reliable foundation for low-latency event triggering, despite lower $F_1$ scores in vivo that underscore the need for further algorithmic refinement in complex neural environments. Conversely, online spike sorting pipelines fail to produce physiologically meaningful clusters across high-density arrays, requiring algorithmic refinement before they can replace offline curation in closed-loop decoding applications.

## 4.5 Closed-loop system

Closed-loop paradigms require real-time processing to ensure that therapeutic feedback is delivered within narrow physiological time windows. Therefore, this section evaluates the operational latency of the closed-loop architecture to determine whether spike detection algorithms and stimulus trigger generation operate with sufficient temporal precision for real-time applications.

The system's operational latency is defined as the total time delay from the moment a signal threshold crossing is detected in the Python script (on the closed-loop workstation) to the delivery of the stimulus at the Effector (routed through the Detection-trigger command and actuation / drive output green lines), as shown in Figure 4.17. A representative example of the two signals is shown in Figure 4.18, top. Evaluation across the entire recording session (Figure 4.18, bottom) reveals a tight latency distribution with a mean value of 3.038 ms and a maximum delay consistently below 4.0 ms . Furthermore, the system exhibited minimal temporal jitter, characterized by a standard deviation of 0.453 ms . This low latency and sub-millisecond jitter demonstrate high temporal precision and response uniformity, confirming that the closed-loop architecture delivers highly predictable and reproducible feedback crucial for real-time applications.

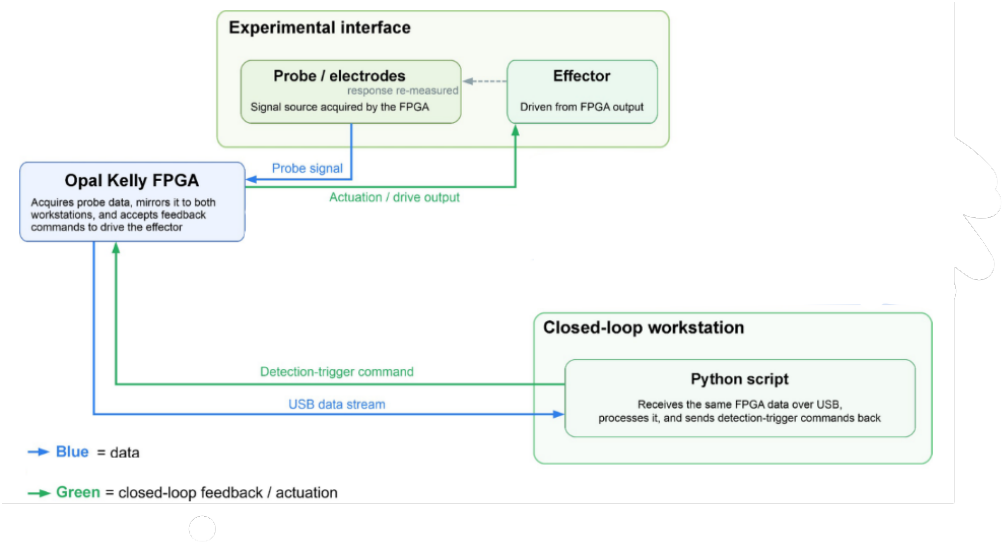



Figure 4.17: Closed-loop system implementation focused on the latency definition. Blue arrows represent data flow and green arrows represent closed-loop feedback and actuation signal paths.

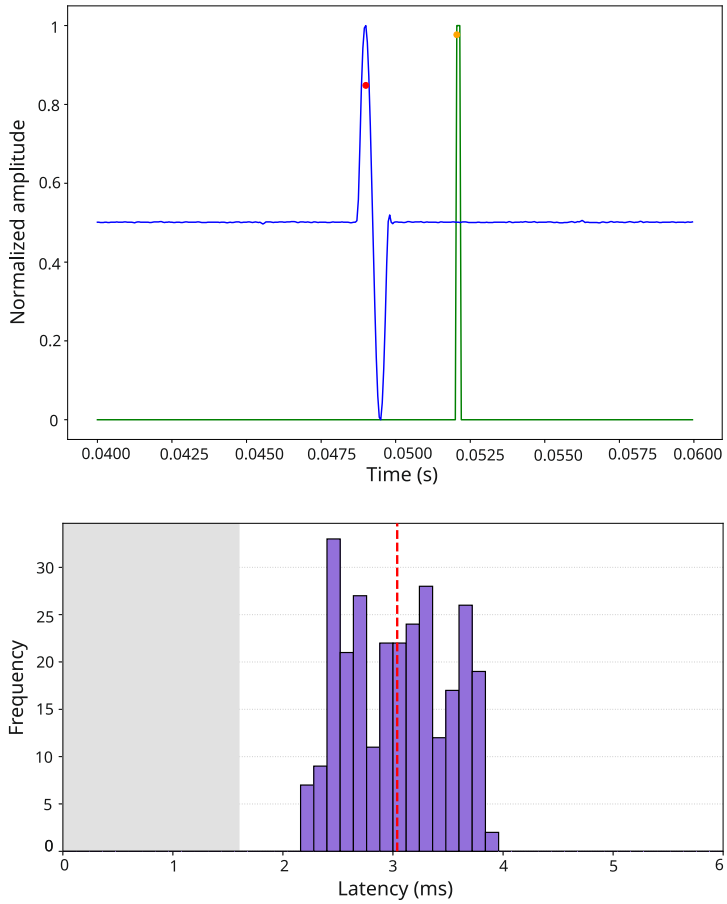



Figure 4.18: Evaluation of real-time spike detection and stimulation feedback timing during closed-loop operation. (Top) Threshold Detection and Triggered Stimulation Timing: Representative trace showing a simulated neural signal (blue) and the corresponding event-triggered stimulation pulse on the auxiliary channel (green), with threshold crossing points highlighted to illustrate single-event operational delay. (Bottom) Overall System Latency Distribution: Histogram displaying the distribution of closed-loop system latency across the full signal recording, depicting the data acquisition chunk size (1.6 ms) and the mean response latency (3.038 ms).

## 4.6 Summary

By bringing together precise insertion hardware via the NeuralGlider, the high-channel-density Sinaps probe, causal preprocessing for (stimulation) artifacts suppression and low-latency real-time detection using the ASO operator, this work establishes a framework for closed-loop CMOS-based electrophysiology. Naturally, transition to a complex in vivo execution highlights clear avenues for refinement. The mechanical benefits of actuation remain to be fully proven

on a larger dataset, simple online thresholding is not sufficient to achieve single-unit isolation and real-time graph clustering still struggles to match offline multi-channel sorting accuracy on high-density recordings. Nevertheless, by achieving sub-millisecond execution budgets while extracting relevant neural features, such as the firing rate, this system demonstrates its viability as a starting point for closed-loop neuroscience experiments.

# Chapter 5

# Conclusion

This dissertation sets out to build and validate a complete workflow for closed-loop neural interfacing using high-density tissue penetrating neural probes, starting with the optimization of the physical insertion of devices and progressively moving towards the implementation of real-time spike detection and sorting algorithms. Over the course of this work, four main components were brought together: the NeuralGlider insertion platform, the SiNAPS CMOS probe, a low-latency closed-loop acquisition architecture built around an Opal Kelly FPGA and a family of real-time signal processing algorithms for spike detection and sorting. Each of these algorithms was evaluated on its own merits, first on controlled synthetic data, then on in vitro recordings, culminating in in vivo validation against an offline reference sorter.

Taken together, the results show that it is possible to close the loop between neural acquisition and actuation with sub-4-millisecond latency, well within the timing margins required for meaningful closed-loop intervention at the single spike level. They also show, just as importantly, where the current approach still falls short. As expected, single-channel thresholding alone cannot deliver single-unit isolation on dense in vivo recordings, and the real time clustering, while promising in principle, still struggles to match the accuracy of heavier offline sorters once the number of channels and units grows. These limitations reflect the inherent trade-offs of operating under strict real-time constraints, alongside the complex challenge of selecting and optimizing appropriate algorithms within a limited development timeframe. Indeed, real-time pipelines are fundamentally designed to prioritize low-latency execution rather than duplicate the unconstrained retrospective processing of offline benchmarks. Resolving these real-time sorting trade-offs remains an open intricate engineering problem that warrants dedicated methodological refinement in future work.

The closed-loop architecture developed in this work achieved its central goal, a working end-to-end pipeline capable of acquiring, processing and acting on neural data with low and consistent latency, while still supporting parallel visualization and recording. The mean system latency of 3.038 ms, with a jitter of 0.453 ms, confirms that the dual-workstation design and the queue-based multiprocessing architecture behind it are good choices for this kind of application.

Conversely, the NeuralGlider evaluation did not confirm the initial hypothesis that ultrasonic actuation would produce a statistically significant improvement in recording quality or unit output over non-actuated insertions. This is a valuable negative result, suggesting that for this specific probe geometry and cortical target, passive insertion techniques may already approach optimal acute signal isolation. However, given the limited sample sizes involved, this observation should be treated as preliminary. Furthermore, advantages of microvibration-assisted insertion are likely to manifest over longer time horizons. Reducing early tissue displacement and inflammation could lead to enhanced long-term tissue integration and superior signal stability in chronic implantation scenarios, a hypothesis that warrants validation in extended longitudinal studies.

On the signal processing side, the adaptive noise-tracking threshold demonstrated effectiveness and robustness, particularly in the presence of optogenetic artifacts. Common average referencing proved essential for the in vitro dataset, reducing the artifact rate from close to 9% down to below 1% before parameter tuning and to zero afterward. The systematic grid search used to select the detection parameters, rather than tuning them manually, was a deliberate methodological choice that ensured the final parameter selection to be reproducible and explainable.

Finally, in vivo results were more mixed, but also informative. Benchmarking the real-time detector against Kilosort4 exposed exactly what was expected, a fast, causal and single-channel detector will always be more permissive than an offline, multi-channel and template-matching sorter, because it does not have access to the same spatial and temporal context. Similarly, the comparison between single-channel and multichannel Louvain sorting revealed two distinct operational limitations. Nevertheless, it is important to emphasize that this represents an initial proof-of-concept implementation that can be improved with further algorithmic optimization. Mapping these trade-offs represents one of the primary contributions of this work.

## 5.1 Future Work

Several directions naturally follow from the limitations identified above. On the sorting side, the most immediate improvement would be a hybrid approach that combines the low latency of single-channel detection with a lightweight spatial correction step, rather than relying purely on either a single channel or the full array at once. This could help the multichannel sorter avoid

grouping temporally uncorrelated spikes from distant electrodes into the same cluster. Testing the pipeline on a larger and more diverse in vivo dataset would also help clarify if the low sorting accuracies observed here are simply a consequence of the specific recording conditions used in this work.

From an insertion perspective, future evaluations of the NeuralGlider should focus on the superficial probe electrodes, where the FBR is typically most pronounced near the cortical surface. Furthermore, extending these comparisons across a larger number of recording sessions, diverse brain regions and varied probe geometries will provide a more statistically robust determination of whether micro-actuation meaningfully reduces insertion trauma in practice. In particular, long-term chronic studies will be crucial to determine if minimizing acute entry trauma successfully reduces the subsequent foreign body response and preserves signal quality over extended timelines, extending the microvibration benefits beyond the results already reported in the literature.

Beyond refining what has already been implemented, an interesting extension of this work would be to enrich the closed-loop decision criterion itself. The current system triggers actuation based purely on spike detection or simple threshold crossings. Incorporating additional biomarkers, such as local field potential power in specific frequency bands, single-unit activity or population-level firing rate synchrony, could allow the closed-loop system to respond to more biologically meaningful events rather than isolated spikes alone. This would move the system from a purely reactive spike detector toward a more context-aware closed-loop controller, which is ultimately the direction this kind of platform needs to grow in. To this end, applying this low-latency pipeline to closed-loop single-unit optogenetic neuromodulation will provide a powerful tool to investigate local neural interactions and activity-dependent plasticity.

# Referências


[1] György Buzsáki, Costas A Anastassiou, e Christof Koch. The origin of extracellular fields and currents—eeg, ecog, lfp and spikes. *Nature reviews neuroscience*, 13(6):407–420, 2012.

[2] Giuseppe Varone, Mana Biabani, Sara Tremblay, Joshua C. Brown, Elisa Kallioniemi, e Nigel C. Rogasch. The golden age of online readout: Eeg-informed tms from manual probing to closed-loop neuromodulation. *NeuroImage*, 322:121543, 2025.

[3] Patrick Ruther e Oliver Paul. New approaches for cmos-based devices for large-scale neural recording. *Current opinion in neurobiology*, 32:31–37, 2015.

[4] Gian Nicola Angotzi, Mihály Vöröslakos, Nikolas Perentos, Joao Filipe Ribeiro, Matteo Vincenzi, Fabio Boi, Aziliz Lecomte, Gabor Orban, Andreas Genewsky, Gerrit Schwesig, et al. Multi-shank 1024 channels active sinaps probe for large multi-regional topographical electrophysiological mapping of neural dynamics. *Advanced Science*, 12(16):2416239, 2025.

[5] Qian Shen, Xiong Zhou, Yuting Wang, Jiaqi Shao, Qiwei Chen, Zhi Gang Cheng, Xumin Ding, Zi-Han Guo, Daqian Hei, Wei Ji, Qiang Wang, Ruifeng Kan, Guanxing Li, Ping Chen, Haiyang Li, Junhao Lin, Nan Lin, Dairong Liu, Feng Luo, Yifan Bao, Bai Song, Haisheng Su, Huan Wang, Xiang Wang, Yajun Wang, Pan An, Yuxi Wang, Sen Yang, Wenqing Yao, Yu Zhou, Qian Cheng, Qing Huan, Bin Ren, Yong Wang, Xianghui Xue, Xuefeng Fu, e Kai Wu. Frontiers and trends in advanced scientific instruments. *Advanced Scientific Instruments*, 1(1):100006, 2026.

[6] Farhad Farkhondeh Tale Navi, Soomaayeh Heysieattalab, Dhakshin S. Ramanathan, Mohammad Reza Raoufy, e Mohammad Ali Nazari. Closed-loop modulation of the self-regulating brain: A review on approaches, emerging paradigms, and experimental designs. *Neuroscience*, 483:104–126, 2022.

[7] Fabio Boi, Andrea Locarno, Joao F Ribeiro, Raffaella Tonini, Gian Nicola Angotzi, e Luca Berdondini. Coupling sinaps high-density neural recording cmos-probes with optogenetic light stimulation. Em *2021 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, páginas 1–4. IEEE, 2021.

[8] Alberto Perna, Gabor Orban, Luca Berdondini, e Joao F Ribeiro. Force measurements to advance the design and implantation of cmos-based neural probes. *IEEE Transactions on Biomedical Engineering*, 72(5):1731–1739, 2024.

[9] Cassandra Acebal, Gurleen Kainth, Harry Villanueva, Ava Hajjar, Valery Fuhrer, Paul Wanczuk, Alexander S Lim, Shriya Srinivasan, Kevin J Otto, e Adam Khalifa. A method for minimally-invasive injection of wireless microdevices into brain tissue. Em *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, páginas 1–4. IEEE, 2024.

[10] Gerardo Saggese, Efstratios Zacharelos, e Antonio Giuseppe Maria Strollo. Low power spike detector for brain-silicon interface using differential amplitude slope operator. Em *2022 17th Conference on Ph. D Research in Microelectronics and Electronics (PRIME)*, páginas 301–304. IEEE, 2022.

[11] Yan Li, Yingnan Nie, Zhaoyu Quan, Han Zhang, Rui Song, Hao Feng, Xi Cheng, Wei Liu, Xinyi Geng, Xinwei Sun, Yanwei Fu, e Shouyan Wang. Brain-machine interactive neuromodulation research tool with edge ai computing. *Heliyon*, 10(12):e32609, 2024.

[12] Jaleh Bagheri Hamzyan Olia, Arasu Raman, Chou-Yi Hsu, Ahmad Alkhayyat, e Alireza Nourazarian. A comprehensive review of neurotransmitter modulation via artificial intelligence: A new frontier in personalized neurobiochemistry. *Computers in Biology and Medicine*, 189:109984, 2025.

[13] Richard Yang, Heather D. Orser, Kip A. Ludwig, e Brandon S. Coventry. msdf-dft: An ultra-low energy discrete fourier transform architecture for closed-loop neural sensing. *Journal of Systems Architecture*, 175:103749, 2026.

[14] United Nations. The 17 goals | sustainable development, 2024. Accessed: April 29, 2024.

[15] James F Cavanagh. Electrophysiology as a theoretical and methodological hub for the neural sciences. *Psychophysiology*, 56(2):e13314, 2019.

[16] Jürgen K Mai e George Paxinos. *The human nervous system*. Academic press, 2011.

[17] Jiawei Zhang. Basic neural units of the brain: neurons, synapses and action potential. *arXiv preprint arXiv:1906.01703*, 2019.

[18] Mark W Barnett e Philip M Larkman. The action potential. *Practical neurology*, 7(3):192–197, 2007.

[19] R Quian Quiroga, Zoltan Nadasdy, e Yoram Ben-Shaul. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural computation*, 16(8):1661–1687, 2004.

[20] György Buzsáki. *Rhythms of the Brain*. Oxford university press, 2006.

[21] Mikhail A Lebedev e Miguel AL Nicolelis. Brain-machine interfaces: from basic science to neuroprostheses and neurorehabilitation. *Physiological reviews*, 2017.

[22] Bradley J Edelman, Shuailei Zhang, Gerwin Schalk, Peter Brunner, Gernot Müller-Putz, Cuntai Guan, e Bin He. Non-invasive brain-computer interfaces: state of the art and trends. *IEEE reviews in biomedical engineering*, 18:26–49, 2024.

[23] Paras R Patel, Elissa J Welle, Joseph G Letner, Hao Shen, Autumn J Bullard, Ciara M Caldwell, Alexis Vega-Medina, Julianna M Richie, Hope E Thayer, Parag G Patil, et al. Utah array characterization and histological analysis of a multi-year implant in non-human primate motor and sensory cortices. *Journal of neural engineering*, 20(1):014001, 2023.

[24] Jess Bartels, Dinal Andreasen, Princewill Ehirim, Hui Mao, Steven Seibert, E Joe Wright, e Philip Kennedy. Neurotrophic electrode: method of assembly and implantation into human motor speech cortex. *Journal of neuroscience methods*, 174(2):168–176, 2008.

[25] Alberto Perna, Gian Nicola Angotzi, Luca Berdondini, e João Filipe Ribeiro. Advancing the interfacing performances of chronically implantable neural probes in the era of cmos neuroelectronics. *Frontiers in neuroscience*, 17:1275908, 2023.

[26] Henry H Radamson, Huilong Zhu, Zhenhua Wu, Xiaobin He, Hongxiao Lin, Jinbiao Liu, Jinjuan Xiang, Zhenzhen Kong, Wenjuan Xiong, Junjie Li, et al. State of the art and future perspectives in advanced cmos technology. *Nanomaterials*, 10(8):1555, 2020.

[27] Giovanni Pietro Seu, Gian Nicola Angotzi, Fabio Boi, Luigi Raffo, Luca Berdondini, e Paolo Meloni. Exploiting all programmable socs in neural signal analysis: A closed-loop control for large-scale cmos multielectrode arrays. *IEEE Transactions on Biomedical Circuits and Systems*, 12(4):839–850, 2018.

[28] Khalil Najafi e Kensall D Wise. An implantable multielectrode array with on-chip signal processing. *IEEE Journal of Solid-State Circuits*, 21(6):1035–1044, 1986.

[29] Jin Ji e Kensall D Wise. An implantable cmos circuit interface for multiplexed microelectrode recording arrays. *IEEE Journal of solid-state circuits*, 27(3):433–443, 1992.

[30] Khalil Najafi, Jin Ji, e Kensall D Wise. Scaling limitations of silicon multichannel recording probes. *IEEE transactions on biomedical engineering*, 37(1):1–11, 1990.

[31] James J Jun, Nicholas A Steinmetz, Joshua H Siegle, Daniel J Denman, Marius Bauza, Brian Barbarits, Albert K Lee, Costas A Anastassiou, Alexandru Andrei, Çağatay Aydın, et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, 2017.

[32] Gian Nicola Angotzi, Fabio Boi, Aziliz Lecomte, Ermanno Miele, Mario Malerba, Stefano Zucca, Antonino Casile, e Luca Berdondini. Sinaps: An implantable active pixel sensor cmos-probe for simultaneous large-scale neural recordings. *Biosensors and Bioelectronics*, 126:355–364, 2019.

[33] George Dimitriadis, Joana P Neto, Arno Aarts, Andrei Alexandru, Marco Ballini, Francesco Battaglia, Lorenza Calcaterra, Susu Chen, Francois David, Richárd Fiáth, et al.

Why not record from every electrode with a cmos scanning probe? *BioRxiv*, página 275818, 2018.

[34] Carolina Mora Lopez, Alexandru Andrei, Srinjoy Mitra, Marleen Welkenhuysen, Wolfgang Eberle, Carmen Bartic, Robert Puers, Refet Firat Yazicioglu, e Georges GE Gielen. An implantable 455-active-electrode 52-channel cmos neural probe. *IEEE Journal of Solid-State Circuits*, 49(1):248–261, 2013.

[35] K. Seidl, S. Herwik, Y. Nurcahyo, T. Torfs, M. Keller, M. Schüttler, H. Neves, T. Stieglitz, O. Paul, e P. Ruther. CMOS-based high-density silicon microprobe array for electronic depth control in neural recording. Em *Proceedings of the IEEE International Conference on Micro Electro Mechanical Systems (MEMS)*, páginas 407–410. IEEE, 2009.

[36] Hongliang Song, Xuewu Yang, Jingming Hou, Wanling Jiang, e Xing Wang. A prosthetic wrist proprioceptive feedback closed loop system based on electrotactile substitution for transradial amputees. *Procedia Computer Science*, 271:287–293, 2025. International Conference on Biomimetic Intelligence and Robotics 2025.

[37] Gen Li, Ye Tian, Luyue Jiang, Shuang Jin, Yifei Ye, Yunxiao Lu, Haoyang Su, Yingkang Yang, Xiaoling Wei, Zhitao Zhou, Tiger H. Tao, e Liuyang Sun. A bimodal closed-loop neuromodulation implant integrated with ultraflexible probes to treat epilepsy. *Biosensors and Bioelectronics*, 271:117071, 2025.

[38] Richárd Fiáth, Adrienn Lilla Márton, Ferenc Mátyás, Domonkos Pinke, Gergely Márton, Kinga Tóth, e István Ulbert. Slow insertion of silicon probes improves the quality of acute neuronal recordings. *Scientific Reports*, 9(1):111, 2019.

[39] Yifan Yu, Zhe Wang, Nils B. Kroemer, Liangwen Zhang, Lin Lu, e Jie Sun. Closed-loop brain-body interface: integrating brain-computer interfaces and peripheral nerve stimulation for adaptive neuromodulation. *Science Bulletin*, 70(24):4128–4132, 2025.

[40] Weitong Liu, Bin Deng, Siyuan Chang, Fei Su, Shanshan Li, Jiang Wang, Jilei Li, Meirong Ma, Yi Yang, e Chen Liu. In-silico, adaptive closed-loop electrophysiology platform for real-time neuromodulation. *Biomedical Signal Processing and Control*, 112:108593, 2026.

[41] Nagy Mekhail, Robert M Levy, Timothy R Deer, Leonardo Kapural, Sean Li, Kasra Amirdelfan, Corey W Hunter, Steven M Rosen, Shrif J Costandi, Steven M Falowski, Abram H Burgher, Jason E Pope, Christopher A Gilmore, Farooq A Qureshi, Peter S Staats, James Scowcroft, Jonathan Carlson, Christopher K Kim, Michael I Yang, Thomas Stauss, Lawrence Poree, Dan Brounstein, Robert Gorman, Gerrit E. Gmel, Erin Hanson, Dean M. Karantonis, Abeer Khurram, Deidre Kiefer, Angela Leitner, Dave Mugan, Milan Obradovic, John Parker, Peter Single, e Nicole Soliday. Long-term safety and efficacy of closed-loop spinal cord stimulation to treat chronic back and leg pain (evoke): a double-blind, randomised, controlled trial. *The Lancet Neurology*, 19(2):123–134, 2020.

[42] Rina Zelmann, Angelique C. Paulk, Ishita Basu, Anish Sarma, Ali Yousefi, Britni Crocker, Emad Eskandar, Ziv Williams, G. Rees Cosgrove, Daniel S. Weisholtz, Darin D.

Dougherty, Wilson Truccolo, Alik S. Widge, e Sydney S. Cash. Closes: A platform for closed-loop intracranial stimulation in humans. *NeuroImage*, 223:117314, 2020.

[43] S. Nambi Narayanan e Sutha Subbian. Hh model based smart deep brain stimulator to detect, predict and control epilepsy using machine learning algorithm. *Journal of Neuroscience Methods*, 389:109825, 2023.

[44] Ramesh Perumal, Jenq-Wei Yang, Yu-Hsiu Kuo, Vincent Vigneron, Hsing-Hua Ho, Hugues Almorin, Chi-Fen Chuang, Yen-Chung Chang, Shih-Rung Yeh, e Hsin Chen. High-precision and efficient suppression of pathological brain activity in parkinsonian rats via a closed-loop deep brain stimulation approach. *Deep Brain Stimulation*, 9:7–17, 2025.

[45] Jiawei Liang, Weitong Liu, Jiang Wang, Bin Deng, Siyuan Chang, e Chen Liu. Nclsilico: A closed-loop neuromodulation platform in silico. *Biomedical Signal Processing and Control*, 90:105829, 2024.

[46] Joseph S. Neimat, Robert W. Bina, Steven C. Koenig, Emrecan Demirors, Raffaele Guida, Ryan Burke, Tommaso Melodia, e Jorge Jimenez. A novel closed-loop electrical brain stimulation device featuring wireless low-energy ultrasound power and communication. *Neuromodulation: Technology at the Neural Interface*, 28(3):455–463, 2025.

[47] Zhuojun Yu, Yangyang Wang, Peter J. Thomas, e Hillel J. Chiel. Tradeoffs in the energetic value of neuromodulation in a closed-loop neuromechanical system. *Journal of Theoretical Biology*, 604:112050, 2025.

[48] Olivier Oullier, Florian Roser, Paul Barbaste, e Xavier Vasques. Improving consciousness assessment through neuroadaptive artificial intelligence and quantum-enhanced brain-computer interfaces. *Clinical Neurology and Neurosurgery*, 266:109396, 2026.

[49] Jessica L. Myatt, Robert Toth, Timothy Denison, Isaac Grennan, Colin G. McNamara, Charlotte J. Stagg, e Andrew Sharott. Modulation of motor cortical theta and gamma oscillations using phase-targeted, closed-loop optogenetic stimulation of local excitatory and inhibitory neurons. *Brain Stimulation*, 19(2):103034, 2026.

[50] Wolf-Julian Neumann, Roee Gilron, Simon Little, e Gerd Tinkhauser. Adaptive deep brain stimulation: From experimental evidence toward practical implementation. *Movement Disorders*, 38(6):937–948, 2023.

[51] Farhan N. Rahman, Prabhkirat S. Bindra, Afra Nawar, H. Trask Crane, Jesus Antonio Sanchez-Perez, John A. Berkebile, Onur Selim Kilic, Chuoqi Chen, Vikram Abbaraju, Christopher J. Nichols, Jacob M. Cook, Samer Mabrouk, Asim H. Gazi, Jin-Oh Hahn, e Omer T. Inan. A wearable system enabling acute stress monitoring and closed-loop mitigation through transcutaneous median nerve stimulation. *Biosensors and Bioelectronics*, 301:118461, 2026.

[52] Weitong Liu, Bin Deng, Siyuan Chang, Fei Su, Shanshan Li, Jiang Wang, Jilei Li, Meirong Ma, Yi Yang, e Chen Liu. In-silico, adaptive closed-loop electrophysiology platform for real-time neuromodulation. *Biomedical Signal Processing and Control*, 112:108593, 2026.

[53] Mingeun Cho, Minhye Choo, Matthew Koh, Sunguk Hong, Junseung Mun, Juho Koo, ChangHwa Oh, e Sung-Min Park. Neuroprosthetic closed-loop strategy for sustained blood pressure reduction via simultaneous stimulation and recording from the spinal cord. *Neurotherapeutics*, 22(6):e00758, 2025.

[54] Junseung Mun, Jiho Lee, e Sung-Min Park. Real-time closed-loop brainstem stimulation modality for enhancing temporal blood pressure reduction. *Brain Stimulation*, 17(4):826–835, 2024.

[55] Paolo Botta, Akira Fushiki, Ana Mafalda Vicente, Luke A. Hammond, Alice C. Mosberger, Charles R. Gerfen, Darcy Peterka, e Rui M. Costa. An amygdala circuit mediates experience-dependent momentary arrests during exploration. *Cell*, 183(3):605–619.e22, 2020.

[56] Shunchang Wu, Qiang Sun, Zhenhui Liu, e Qi Yang. A biologically inspired control architecture translating motoneuron dynamics into neuroprosthetic systems. *Results in Engineering*, página 110179, 2026.

[57] Rafał Bogdan Drobot. Nocturnal closed-loop transcutaneous tibial nerve stimulation: a cardio-urologic hypothesis to interrupt the bladder–heart arousal axis and extend life in older adults. *Medical Hypotheses*, 208:111902, 2026.

[58] Michal Ramot e Alex Martin. Closed-loop neuromodulation for studying spontaneous activity and causality. *Trends in Cognitive Sciences*, 26(4):290–299, 2022.

[59] Adriaan J Taal, Ilke Uguz, Sabina Hillebrandt, Chang-Ki Moon, Victoria Andino-Pavlovsky, Jaebin Choi, Changmin Keum, Karl Deisseroth, Malte C Gather, e Kenneth L Shepard. Optogenetic stimulation probes with single-neuron resolution based on organic leds monolithically integrated on cmos. *Nature Electronics*, 6(9):669–679, 2023.

[60] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee micro*, 38(1):82–99, 2018.

[61] Daisuke Ishikawa, Nobuyoshi Matsumoto, Tetsuya Sakaguchi, Norio Matsuki, e Yuji Ikegaya. Operant conditioning of synaptic and spiking activity patterns in single hippocampal neurons. *Journal of Neuroscience*, 34(14):5044–5053, 2014.

[62] Sa Zhou, Yang Liu, Adam Turnbull, Cristiano Tapparello, Ehsan Adeli, e F. Vankee Lin. Personalized cognitive enhancement for older adults: An aging-friendly closed-loop human-machine interface framework. *Ageing Research Reviews*, 112:102877, 2025.

[63] Tara L Skarpaas, Beata Jarosiewicz, e Martha J Morrell. Brain-responsive neurostimulation for epilepsy (rns® system). *Epilepsy research*, 153:68–70, 2019.

[64] Joohi Jimenez-Shahed. Device profile of the percept pc deep brain stimulation system for the treatment of parkinson's disease and other movement disorders. *Expert Review of Medical Devices*, 18(4):319–332, 2021.

[65] Mattia Arlotti, Sara Marceglia, Guglielmo Foffani, Jens Volkmann, Andres M Lozano, Elena Moro, Filippo Cogiamanian, Marco Prenassi, Tommaso Bocci, Francesca Cortese, Paolo Rampini, Sergio Barbieri, e Alberto Priori. Eight-hour adaptive deep brain stimulation in patients with parkinson disease. *Neurology*, 90(11):e971–e976, 2018.

[66] Bin Deng, Guohang Xiu, Weitong Liu, Jiang Wang, e Chen Liu. Pid neural network-based multi-channel feedback closed-loop control for deep brain stimulation. *Biomedical Signal Processing and Control*, 110:108133, 2025.

[67] Colin Goudelocke, Jennifer Sobol, Denise Poulos, Ekene Enemchukwu, Stanley Zaslau, e Rohit Dhir. A multicenter study evaluating the frequency of use and efficacy of a novel closed-loop wearable tibial neuromodulation system for overactive bladder and urgency urinary incontinence (freeoab). *Urology*, 183:63–69, 2024.

[68] Hanbin Cho, Jeffrey Ojemann, e Jeffrey Herron. Open mind neuromodulation interface for the cortec brain interchange (omni-bic): an investigational distributed research platform for next-generation clinical neuromodulation research. Em *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, páginas 1–6, 2023.

[69] Amy S Chuong, Mitra L Miri, Volker Busskamp, Gillian AC Matthews, Leah C Acker, Andreas T Sørensen, Andrew Young, Nathan C Klapoetke, Mike A Henninger, Suhasa B Kodandaramaiah, et al. Noninvasive optical inhibition with a red-shifted microbial rhodopsin. *Nature neuroscience*, 17(8):1123–1129, 2014.

[70] Marius Pachitariu, Shashwat Sridhar, Jacob Pennington, e Carsen Stringer. Spike sorting with kilosort4. *Nature methods*, 21(5):914–921, 2024.

[71] Tae Kyun Kim. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540–546, 2015.

[72] Xueyi Wang. A fast exact k-nearest neighbors algorithm for high dimensional search using k-means clustering and triangle inequality. Em *The 2011 international joint conference on neural networks*, páginas 1293–1299. IEEE, 2011.

[73] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, e Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[74] G Ayorkor Mills-Tettey, Anthony Stentz, e M Bernardine Dias. The dynamic hungarian algorithm for the assignment problem with changing costs. *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27*, 7, 2007.