Low-Power, Small-Area, and High-Fidelity, Neuromorphic Send-on-Delta Autoencoder for Neural Compressed Telemetry in *High-Channel-Count* Intracortical Brain-Computer Interfaces

Aurora Pia Ghiardelli[(1)(2)], Yuming He[(1)(3)], Hua-Peng Liaw[(1)], Guangzhi Tang[(2)], Yao-Hong Liu[(1)]

[(1)]imec, Eindhoven, the Netherlands, [(2)] Maastricht University, Maastricht, the Netherlands, [(3)] University of Groningen, Groningen, the Netherlands

Intracortical brain-computer interfaces (iBCIs) are becoming increasingly important in neuroscience and neuroprosthetic applications. Advances in high-channel-count microelectrode arrays (MEAs) now enable simultaneous recording from hundreds to thousands of electrodes at high temporal resolution, but this scalability comes at the cost of dramatically increased data throughput, exceeding 100 Mbps in large-scale systems. Because clinically viable iBCIs require wireless transcutaneous telemetry, a fundamental mismatch arises between neural data generation and wireless transmission capability: low-power links provide insufficient throughput, whereas higher-rate solutions require substantially greater power and approach safety-limited budgets set by tissue-heating constraints. Consequently, energy-efficient on-chip compression is essential for next-generation implantable iBCIs, but practical compression methods must simultaneously satisfy three competing requirements: low hardware and computational cost under stringent power and silicon-area constraints, high signal fidelity to preserve waveform features required for downstream analyses such as spike sorting and behavioral decoding, and low latency for closed-loop applications. To address these challenges, we propose a multi-mode Send-on-Delta Autoencoder (SODA) framework for low-power and accurate neural signal compression in high-density iBCIs. SODA adopts a split architecture comprising an on-chip event-driven neuromorphic compressor and a remote decoder on an external device. The on-chip encoder combines delta event encoding and a spiking neural network (SNN) to exploit neural sparsity for efficient computation and low memory usage, while the remote RNN-based decoder enables accurate signal reconstruction. To preserve sparse action potentials at high compression ratios, the autoencoder is trained with a loss function that jointly optimizes the quality of spike reconstruction while minimizing the transmission rate. For efficient low-latency wireless transmission, we further integrate online processing with an event-based serializer (eSER) that packetizes SNN outputs. Experimental results show that SODA achieves a 98.5% reduction in memory access and a 99.6% reduction in MAC operations compared with state-of-the-art methods. These gains translate into an ultra-low power consumption of just 0.24 μW/channel and a compact silicon area of only 11,047 μm^2, while maintaining lower latency, up to 130$\times$ compression ratio, and accurate spike reconstruction. These results highlight the potential of neuromorphic, event-driven compression methods for low-power neural telemetry in next-generation implantable iBCI systems.