NEUROMORPHIC
Computing and Engineering

PERSPECTIVE

Brain-inspired nanophotonic spike computing: challenges and prospects

Bruno Romeira[1,*], Ricardo Adão[1], Jana B Nieder[1], Qusay Al-Taai[2], Weikang Zhang[2,3], Robert H Hadfield[2], Edward Wasige[2], Matěj Hejda[3], Antonio Hurtado[3], Ekaterina Malysheva[4], Victor Dolores Calzadilla[4], João Lourenço[5], D Castro Alves[5], José M L Figueiredo[5], Ignacio Ortega-Piwonka[6,7], Julien Javaloyes[6,7], Stuart Edwards[8], J Iwan Davies[8], Folkert Horst[9] and Bert J Offrein[9]

[1] INL—International Iberian Nanotechnology Laboratory, Ultrafast Bio- and Nanophotonics, Av. Mestre José Veiga s/n, 4715-330 Braga, Portugal
[2] James Watt School of Engineering, University of Glasgow Rankine Buidling, Oakfield Avenue, Scotland, Glasgow, G12 8LT, United Kingdom
[3] Institute of Photonics, SUPA Dept of Physics, University of Strathclyde, 99 George Street G1 1RD Glasgow, United Kingdom
[4] Eindhoven Hendrik Casimir Institute, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands
[5] Centra-Ciências and Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[6] Institute of Applied Computing and Community Code (IAC-3). University of the Balearic Islands, Carretera de Valldemossa, km. 7.5, Palma 07122, Spain
[7] Department of Physics. University of the Balearic Islands, Carretera de Valldemossa, km. 7.5, Palma 07122, Spain
[8] IQE plc, Cardiff CF3 0LW, United Kingdom
[9] IBM Research—Zurich, 8803 Rüschlikon, Switzerland
[*] Author to whom any correspondence should be addressed.

E-mail: bruno.romeira@inl.int

Keywords: nanophotonics, resonant tunnelling diodes, nanoLEDs, nanolasers, neuromorphic computing, optical interconnects, spiking neural networks

OPEN ACCESS

RECEIVED
7 July 2022

REVISED
13 February 2023

ACCEPTED FOR PUBLICATION
16 June 2023

PUBLISHED
14 July 2023

Abstract

Nanophotonic spiking neural networks (SNNs) based on neuron-like excitable subwavelength (submicrometre) devices are of key importance for realizing brain-inspired, power-efficient artificial intelligence (AI) systems with high degree of parallelism and energy efficiency. Despite significant advances in neuromorphic photonics, compact and efficient nanophotonic elements for spiking signal emission and detection, as required for spike-based computation, remain largely unexplored. In this invited perspective, we outline the main challenges, early achievements, and opportunities toward a key-enabling photonic neuro-architecture using III–V/Si integrated spiking nodes based on nanoscale resonant tunnelling diodes (nanoRTDs) with folded negative differential resistance. We utilize nanoRTDs as nonlinear artificial neurons capable of spiking at high-speeds. We discuss the prospects for monolithic integration of nanoRTDs with nanoscale light-emitting diodes and nanolaser diodes, and nanophotodetectors to realize neuron emitter and receiver spiking nodes, respectively. Such layout would have a small footprint, fast operation, and low power consumption, all key requirements for efficient nano-optoelectronic spiking operation. We discuss how silicon photonics interconnects, integrated photorefractive interconnects, and 3D waveguide polymeric interconnections can be used for interconnecting the emitter-receiver spiking photonic neural nodes. Finally, using numerical simulations of artificial neuron models, we present spike-based spatio-temporal learning methods for applications in relevant AI-based functional tasks, such as image pattern recognition, edge detection, and SNNs for inference and learning. Future developments in neuromorphic spiking photonic nanocircuits, as outlined here, will significantly boost the processing and transmission capabilities of next-generation nanophotonic spike-based neuromorphic architectures for energy-efficient AI applications. This perspective paper is a result of the European Union funded research project ChipAI in the frame of the Horizon 2020 Future and Emerging Technologies Open programme.

1. Introduction

Artificial intelligence (AI) systems powered by concepts such as machine learning and deep neural networks (DNNs) are rapidly emerging [1]. However, the power budget required to run large scale deep learning algorithms on conventional computers is growing exponentially [2]. In fact, DNNs are a class of artificial neural networks (ANNs) that has achieved a widespread success due to improved computation, efficient algorithms, and the vast access to Big Data. ANNs are designed to mimic the brain's high-level organizational architecture, in which processing units (neurons) are stacked in layers and interconnected via adjustable weights (synapses). However, information in the brain is transmitted not only across multi-layers (spatial domain), but also in the form of nerve impulses (time domain), i.e. abrupt, short changes in potential commonly referred to as *spikes* that travel along nerve fibres and always have similar duration and amplitude. The spiking neural network (SNN) is a popular mathematical model that describes computation using temporally-encoded impulses [3]. A SNN processes information based on spike timing, and each artificial neuron is typically only active when it receives or emits spikes. In principle, such feature could significantly reduce the amount of energy required to run the neural network. Importantly, it allows for the encoding of information in both the spatial and temporal domains, making it capable of highly complex information processing functionalities that could approach the processing paradigm of the brain cortex. The first hardware implementations of SNNs were demonstrated using either conventional microelectronic chips (e.g. IBM's TrueNorth [4], and Intel's Loihi [5]) or memristors (see a comprehensive review in [6]). Despite their impressive results, these neuro-architectures are reaching the following bottlenecks: (i) fundamental quantum limits beyond which transistors cannot be miniaturized, (ii) speed limitation of $\sim$1 ns, and (iii) large synaptic energy of $\sim$1 pJ/spike, limited by electrical interconnects [5]. To overcome these limitations, a shift toward photonics-enabled neural networks is being pursued [7].

The use of photons, which have a non-interacting bosonic nature, is advantageous for low-loss and high-bandwidth information transmission. The use of light for transmission overcomes the parasitic constraints of electronics (Joule heating effect). Importantly, massive parallel connectivity can be achieved using methods such as wavelength division multiplexing [8], and the fan-in/-out bottleneck could be mitigated. As a result, photonic approaches for optical interconnects [9], matrix multiplication operations (tensor cores) [8, 10], and reservoir computation [11] appear promising architectures for neural networks. Nonetheless, there are still significant challenges for realizing neuro-architectures beyond deep learning using photonics. Among these are biomimicking the complexity and nonlinearities of neuronal dynamics (e.g. stochastic spiking, bursting [12]) and realizing optical memory (short- and long-term) functions for storing and accessing information and neurosynaptic weights [13]. Therefore, the next-generation of neuromorphic processors will almost certainly require synergies from a hybrid co-integration of electronics and photonics capable of: (i) implementing the neuron and memory nonlinear functions via optical-to-electronic-to-optical (O/E/O) transduction, and (ii) implement interconnectivity and weighted interconnects (synapses) taking advantage of all-optical photonics approaches. However, as evidenced by the plethora of components explored in recent years to realize photonic neurons, synapses and interconnects (figure 1), there is still no general agreement for a common photonic platform on which technologies could be combined to implement a large scale neuromorphic optical neural network and computing processor.

The impressive developments in the field of neuromorphic photonics over the last 10 years have recently been reviewed [7, 14], and will not be extensively discussed here. Figure 1 shows representative (but by no means exhaustive) components of optical emitters, panels (a)–(d) [15–18], and optical receivers, panels (j)–(m) [20–23], suited for neuron spike-based functions (and also for non-spiking functions). It is worth noting that, apart from a few attempts to miniaturize optical neurons using either micropillars, panel (c) [17], or photonic crystals, panel (d) [18], the majority of approaches thus far rely on bulky coherent laser sources with saturable absorbers (e.g. using distributed feedback (cavities), panel (a) [15]), or vertical-cavity surface-emitting lasers (VCSELs), panel (b) [16]. Indeed, scaling devices to the submicrometer range (100 times smaller than current components) for power-efficient solutions, specifically the laser, requires more advanced optical confinement methods, such as 2D or 3D photonic crystals or the use of plasmonic resonances (see a review in [24, 25]). Although such methods have been employed to fabricate nanoscale laser diode (nanoLD) and light-emitting diode (nanoLED) devices [26–30], operating such sources (and receivers) with brain-like functions, specifically the encoding and decoding of neuron-like excitable spiking signals, remains a significant challenge. Figure 1 displays examples of photonic synapses (panels (e)–(g)) which could be used to interface and weight the photonic spiking nodes, either using all-optical (panel (e) [19]) or electro-optical (panels (f) and (g) [31, 32]) methods. The approaches generally employ either non-volatile [19] or volatile [31, 32] memories as synaptic elements. One disadvantage, when compared to electronics, is that optical memories cannot usually be written and read at high frequencies. Finally, panels (h) and (i) present photonic components useful for realizing 3D optical couplers [33] and 3D photonic wires

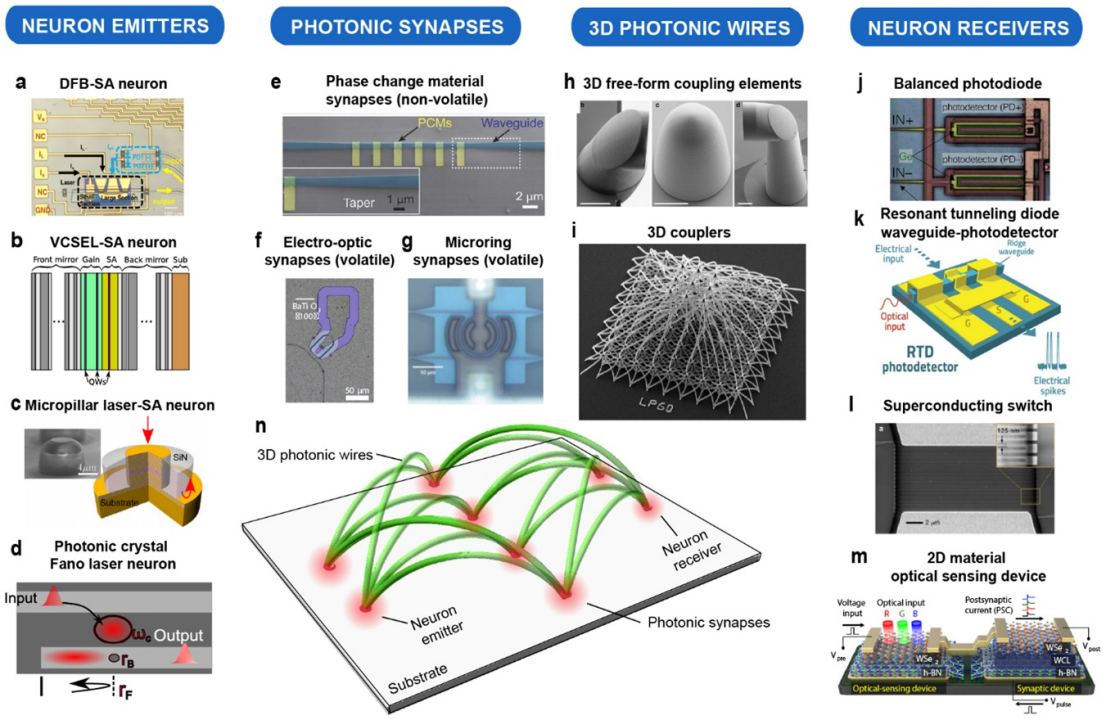

Figure 1. Representative implementations of neuron emitters (a)–(d), photonic synapses (e)–(f), 3D photonic wires and 3D couplers (h)–(i), and neuron receivers (j)–(m). Neuron optical emitters: (a) distributed feedback laser with a saturable absorber (© [2018] IEEE. Reprinted, with permission, from [15].), (b) vertical-cavity surface-emitting laser with a saturable absorber (Adapted with permission from [16] © The Optical Society.), (c) micropillar laser with a saturable absorber (Reprinted figure with permission from [17], Copyright (2014) by the American Physical Society.), (d) photonic crystal Fano laser neuron (Adapted with permission from [18] © The Optical Society.); Photonic synaptic weights: (e) non-volatile phase change material synapse (Reproduced from [19]. CC BY 4.0.), (f) volatile electro-optic synapse (Reproduced from [31], with permission from Springer Nature.), (g) volatile microring synapse (Reproduced from [32]. CC BY 4.0.); 3D optical couplers and wires using 3D microprinting: (h) 3D free-form coupling elements (Reproduced from [33], with permission from Springer Nature.), (i) dense arrangement of 3D couplers (Reprinted with permission from [34] © The Optical Society.); Neuron optical receivers: (j) balanced photodiodes (Reprinted figure with permission from [20], Copyright (2019) by the American Physical Society.), (k) resonant tunnelling diode waveguide-photodetector as a spiking receiver (Reprinted with permission from [21] © The Optical Society.), (l) superconducting thermal switch (Reproduced from [23], with permission from Springer Nature.), (m) 2D material-based optical sensing device (Reproduced from [22]. CC BY 4.0.); (n) Schematic representation of the concept of a photonic artificial neural network interconnecting neuron emitters, synapses ad neuron receivers by 3D photonic waveguides.

[34] using microscale 3D printing of polymer materials employing two-photon polymerization (TPP). This enables flexible 3D light coupling from and to neuron emitters and receivers, panel (n), complementary to the 2D waveguide interconnectivity.

It is recognized that the integration of the wide-range of designs and materials options available to achieve both photonic and (known) neural functions is a challenge. As a result, several design choices must be made, including: (i) materials selection (silicon (Si), III–V, 2D materials, polymer or other), (ii) integration with (Si) electronics, (iii) wavelength operation, (iv) power output of the emitter and receiver sensitivity, (v) light in- and out-coupling strategy (vertical or horizontal plane), (vi) 2D or 3D interconnectivity, (vii) temperature operation (cryogenic or room-temperature), to name a few. As a result, selecting the components for co-integration in a single neuromorphic photonic platform still remains a remarkable challenge.

In this invited perspective, we discuss recent results and opportunities towards a photonic neuro-architecture based on neuromorphic III–V/Si nanoscale spiking neuron light-emitters and light-receivers, all of which share a common optoelectronic co-integration with either III-V gallium arsenide (GaAs), III-V indium phosphide (InP) or III–V/Si nanoscale resonant tunnelling diodes (nanoRTDs) [35]. Our approach can potentially solve the bottleneck associated with combining complex spatio-temporal neural functions in a single neuro-nanotechnology platform. The unique nanoRTD component discussed here as artificial neuron is based on a low-dimensional III–V double barrier quantum well (DBQW) semiconductor heterostructure. The resonant states of the DBQW, which are fine-tuned by the applied voltage across the nanoRTD device, filter the charge carrier flow, resulting in a non-monotonous N-shaped current–voltage (I–V) characteristic with a region of negative differential conductance (NDC) [35]. A wide-range of high-bandwidth applications take advantage of the on-chip electrical gain amplification

provided by the NDC region, including terahertz emitters and detectors (world's record oscillation of 1.92 THz at room temperature [36]) and photonic devices such as single-photon switches and photodetectors [37, 38]. Noteworthy, in the context of neuromorphic applications, the NDC property has been shown to enable functionalities analogous to biological neurons. The most remarkable properties include excitability (all-or-nothing spiking) and autaptic regenerative memory [21, 39]. However, in order to achieve neuron-like functions in the optical domain, integration with other optoelectronic components, such as lasers, light-emitting diodes (LEDs), and photodetectors is required since nanoRTDs are purely electronic devices. Yet, these optoelectronic components are typically bulky ($>100\ \mu m^2$) and power hungry (>1 pJ). To achieve power-efficient, cost-effective and small footprint solutions, efforts on miniaturization and integration (monolithic or hybrid) of such DBQW-based optoelectronic neuro-nanotechnology are required, while maintaining the unique neuron-like spiking properties.

Herein, we discuss the prospects, recent results, and relevant challenges of integrating nanoscale light sources and nanophotodetectors with nanoscale RTDs to realize novel components suited for emission and detection of neural-like optical spikes with a small footprint, fast speed, and low-power consumption. This perspective work is a result of the research project ChipAI [40]—Energy-efficient and high-bandwidth neuromorphic nanophotonic Chips for AI systems—in the frame of the Horizon 2020 Future and Emerging Technologies Open programme funded by the European Union. The paper is organized as follows. Section 2 introduces photonic SNNs, as well as the main applications of photonic SNNs, and the motivation towards nanoscale photonic SNNs. Section 3 describes the physical principles of nanoRTDs as fundamental building blocks of artificial neurons for SNNs, including their neuron-like nonlinear properties such as self-oscillations, excitability, bistability and bursting. In section 4, we briefly discuss the epitaxial growth of III–V DBQW structures on native III–V substrates and on germanium-on-silicon (Ge-on-Si) templates to enable future compatibility of this architecture with low-cost silicon substrates. In section 5, we discuss the integration of nanoRTDs with nanolight sources, specifically nanoLEDs and nanolasers, to realize spiking photonic neuron emitters, and the integration of nanoRTDs with photodetectors to realize spiking neuron receivers. Section 6 covers relevant strategies for interconnecting emitter-receiver spiking photonic neurons using (i) silicon photonics interconnects, (ii) integrated photorefractive interconnects, and (iii) polymer photonics 3D waveguide interconnections. In section 7, we present the first results on spike-based spatio-temporal learning methods for applications in AI-based functional systems using numerical simulations of artificial neuron models, and we illustrate simulation results of pattern image recognition, image edge detection, and SNNs for inference and learning tasks. Lastly, in section 8, we outline the main conclusions and future prospects of this nanophotonic neuro-nanotechnology.

2. Photonic SNNs

2.1. Comparison between DNNs and SNNs

ANNs represent an attempt to mimic the biological brain's complex architecture with its high degree of connectivity and parallelism, in order to achieve some of the brain's remarkable information processing capabilities. However, even state-of-the-art neural networks implemented in digital platforms represent a significantly simplified abstraction of the behaviour observed in biological neuronal networks. The architecture of artificial neurons and ANNs has over time evolved in what is usually considered as three major generations of neural networks (perceptrons, DNNs and SNNs) [41].

2.1.1. First generation—perceptrons

The output of a neuron is binary (0, 1) and is obtained by a simple thresholding of the weighted synaptic inputs.

2.1.2. Second generation—DNNs

Extensively used in deep learning, DNNs more closely mimic the high-level organizational (spatial) architecture of the brain where the processing units (neurons) are stacked in processing layers, with adjacent layers interconnected via adjustable weights. The output of a neuron can be a real number, obtained as a weighted synaptic input and transformed using a nonlinear activation function (such as the sigmoid or rectified linear unit functions). Back-propagation training method was a crucial development that allowed DNNs to proliferate [42]. In this case, the network is trained with a set of a labelled training data (unsupervised trainning can also be implemented in DNNs). The mismatch between network response and the label is used to determine the weight update that will minimize the network loss function. The artificial neurons used in DNNs use differentiable nonlinear activation functions.

Table 1. Qualitative comparison between DNNs and SNNs.

Features	DNNs	SNNs
Energy consumption (in dedicated hardware)	Continuous (high-energy consumption)	Only during firing events (low-energy consumption) [Continuous in digital SNN implementations]
Data processing	Frame-based (and synchronous)	Event-based (and asynchronous)
Activation function	Sigmoid, ReLU, etc.	Spike
Neuronal unit model complexity	Low	High
Short-term memory	Network level (e.g. recurrent connections)	Neuron and network levels (global-local learning)
Noise robustness	Low	High
Data processing	Spatial processing (frame-based)	Spatio-temporal processing (event-based)
Data sets	Static (typically)	Dynamic (typically)
Training	Mature approaches (differentiable activations)	More challenging (non-differentiable)
Applications	Non-event-based (e.g. image classification)	Event-based/temporal (e.g. time-series prediction, vision)

2.1.3. Third generation—SNNs

Despite the success of DNNs, the nonlinear dynamics of neurons in the human brain are more complex and operate in both spatial and temporal domains, wherein each neuron communicates with other neurons via the so-called action potentials (spikes). Nowadays, growing research efforts focus on extending methods derived from deep learning towards operation incorporating the temporal domain for information representation. Therefore, a paradigm shift in ANNs is the shift from DNNs to SNNs. Computation in SNNs is event-driven as in the biological brain and is performed on these events asynchronously. Therefore, each neuron in the network generates spike outputs in complex, time-resolved response to spatio-temporal inputs from upstream neurons, and remains quiescent otherwise. Table 1 shows a comparison of the key features of DNNs and SNNs. SNNs have already an in-built temporal short-memory at the level of individual neurons. In DNNs this can only be achieved by altering the network topolology (e.g. using recurrent connections). Therefore, SNNs are a particularly interesting candidate for processing natively temporal data with high parallelism due to their short-memory and spatio-temporal features.

2.1.4. Applications of SNNs

Nowadays, DNNs represent the key approach in the field of machine learning, and power a significant range of data processing algorithms. While these models are exceedingly powerful, they operate at massive scales, and require significant amount of resources, both in terms of time and energy (FLOPS), as well as in terms of input data [43]. Spiking (event-based) signalling enables computation with much higher degree of sparsity, and provides a natural, better fit for data represented natively in time. The spatio-temporal processing of data enables parallelizable computing, of relevance for the design of energy-efficient computers. The asynchronous nature of spiking neurons enables operation without need for clocking, and spikes represent a unique encoding scheme that is robust in amplitude, continuous in time and restorative as a result of spike firing excitability. Furthermore, use of SNNs unlocks new possibilities in terms of available learning rules, including biologically inspired approaches such as global-local learning and spike-timing-dependent plasticity [13, 44]. Finally, these networks provide a perfect match for neuromorphic hardware, such as brain-inspired event-based vision sensory systems [45].

2.2. Spiking neurons

Signalling in neurons is typically enabled by a wide range of nonlinear dynamical phenomena including pulsing, excitable spikes, bursting and oscillations. Spiking allows for temporal analogue-like encoding of information which is highly resilient to noise and extremely efficient, consuming only approximately 10^4 adenosine triphosphate (ATP) molecules to transmit a bit at a chemical synapse [46], corresponding to around 1 fJ bit^{-1} at 32 bit s^{-1} (i.e. several orders of magnitude lower than existing neuromorphic technologies). Each neuron in the human neocortex receives input spikes from approximately 10^4 other neurons, with each neuron spiking at a sparse rate between 0.1 and 100 Hz. This parallelism and sparse activity combined with temporal integration property makes the brain a formidable power-efficient and error-tolerant computing natural machine. This represents a very powerful encoding system when many

connections (10^{14} synapses in the human brain) are established between individual neurons, and hardware solutions biomimicking neurons are being developed in electronic and photonic-based SNN platforms.

2.3. Electronic SNNs

There is nowadays a wide range of both industrial and academic research activities aiming at realizing neuromorphic hardware. Among the industrial leaders are Cerebras with their WSE-2 processor, BrainChip with their Akida neural processor, IBM with their TrueNorth chip and Intel with their recently released second generation of Loihi self-learning chip. In Europe, the FACETS program demonstrated a chip with 200 000 neurons and 50 million synaptic connections which led to the BrainScaleS project and to the neuromorphic computing platform in 'The Human Brain Project' (HBP). Additionally, SpiNNaker, an SNN architecture, using a massively parallel, manycore supercomputer architecture designed by the Advanced Processor Technologies Research Group, University of Manchester, is being used as one component of the neuromorphic computing platform for HBP. Thus far, these approaches rely mostly on conventional complementary metal-oxide-semiconductor (CMOS) transistors. Alternative solutions include in-memory computing using resistive switching devices (e.g. memristors, phase change), and spintronics—see a review in [47]—in an attempt to emulate the action-potential functionalities of the brain using beyond Moore's technologies.

2.4. Photonic SNNs using optical neurons

Optics and optoelectronics are among the key prospective technologies that will allow to overcome both interconnect energy and bandwidth density constraints of neuromorphic electronic approaches. Using photonics, light-speed neuron-like spikes can be achieved in semiconductor lasers namely using commercially available VCSELs [48, 49], and other optoelectronic platforms (review in [50]). To decrease spatial footprint, approaches for optical neural networks are being tested using either mature III–V/Si photonic integrated platforms or fibre optics-based reservoir computing, therefore allowing for the implementation of coherent light-based deep learning, reservoir computing, and photonic accelerators (tensor cores) (review in [7]). Currently, many of these approaches require off-chip coherent laser sources, detectors, and optical amplifiers, and additional pre- and post-processing steps, hence imposing a bottleneck for compact and efficient solutions. Importantly, several of the photonic platforms are lacking suitable optical neurons for use in SNN architectures.

In the past few years, there have been various initiatives aiming at realizing all-optical and optoelectronic SNNs. Here we discuss a few representative examples. In 2019, a photonic integrated neuromorphic computing platform based around phase-change materials (PCMs) and devices on silicon photonics was realized, figure 2(a), providing an hardware with the basic integrate-and-fire functionality of neurons and the plastic weighting operation of synapses [51]. The architecture required more than 100 optical elements occupying a mm^2 footprint to emulate four neurons and sixty synapses. Still, biological neurons exhibit more complex spike-based functions. In 2021, artificial optical neurons were implemented with anti-symmetrically coupled degenerate optical parametric oscillators (DOPOs) [52]. The nonlinearity and phase bistability of the DOPOs were used to generate two spiking modes of class-I (saddle-node bifurcation) and class-II (Andronov–Hopf bifurcation) neurons that had been originally classified by A.L. Hodgkin. Using these optical neurons, an experimental setup of a DOPO-based neural network was shown using a 240-node ANN composed of 480 DOPOs with antisymmetric couplings and based on time-domain multiplexing in a 1 km fibre-ring cavity, figure 2(b). This architecture required several bulky components such as modulators and optical amplifiers. Energy efficient and scalable photonic-based approaches demand at least a 100-fold reduction in the size of these components that can perform complex spiking signalling functions while using significantly less energy. A brain-inspired neuron architecture based on the Izhikevich model and implemented in nano-optoelectronic hardware has been suggested in 2022 [53]. The optoelectronic neurons consist of two photodetectors for excitatory and inhibitory optical spiking inputs, electrical transistors' circuits providing spiking nonlinearity, and a laser for optical spiking outputs, but the demonstration of a photonic SNN is yet to be shown. In perspective, active optical spiking components integrated on-chip in a practical scalable SNN architecture are still at an early stage of development.

2.5. Towards photonic SNNs using nanoscale optical neurons

2.5.1. Miniaturized light spiking sources

Subwavelength nanolaser sources emitting pulses: In order to reduce the size of a laser to the submicron scale, metallic cavity semiconductor and plasmonic light sources have shown the highest potential. Several groups already achieved sub-μm nanolasers with ultralow threshold currents (<10 μA), enabling energy-efficient lasers—see review in [24, 25]. Yet, these do not feature brain-like nonlinear neurosynaptic functions and this remains largely unexplored.

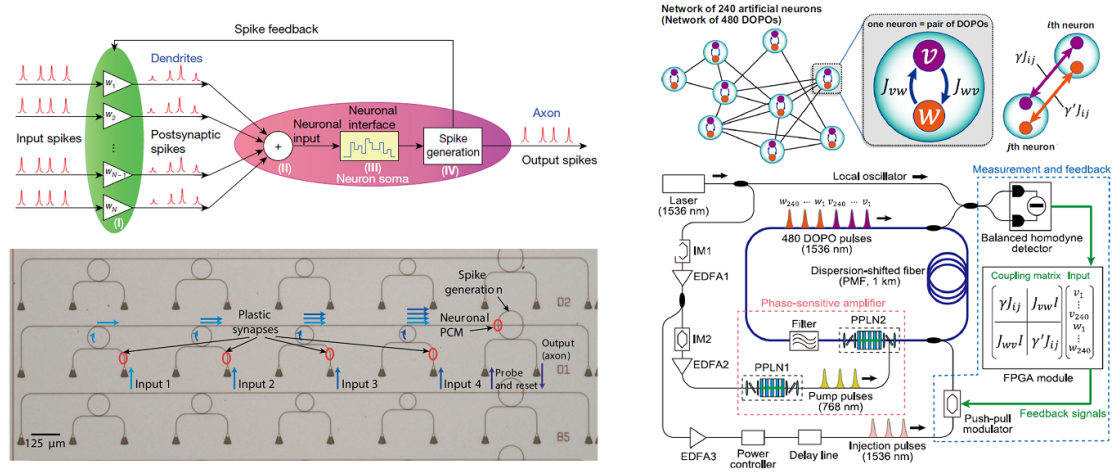

Figure 2. Representative implementations of photonic spiking neural networks using all-optical neuromorphic spiking neurons hardware. (a) (*top*) Schematic of an all-optical spiking neural network circuit consisting of several pre-synaptic input neurons and one post-synaptic output neuron connected via phase change material synapses; (*bottom*) optical micrograph of fabricated optical neurons showing four input ports (Reproduced from [51], with permission from Springer Nature.). (b) Experimental setup of a degenerate optical parametric oscillator (DOPO)-based neural network. (*top*) A 240-node artificial neural network composed of 480 DOPOs; (*bottom*) schematic diagram based on time-domain multiplexing in a 1 km fibre-ring cavity (Reproduced from [52]. CC BY 4.0.).

Subwavelength nanoLED sources emitting pulses: For short distance on-chip brain-inspired communications, incoherent nanoLED sources may represent an alternative to nanolasers since they require neither high-Q cavities, nor a lasing threshold. However, LEDs are simple linear devices not designed to achieve nonlinear self-pulsating signals and new approaches are needed to achieve nonlinear nanoLEDs for the emission of neuron-like spiking signals.

2.5.2. Miniaturized light spiking receivers

Receiving/decoding low-photon spikes using photodetectors: In d.c. operation, light sources emit in the low photon count regime ($\ll 10^4$), and detecting modulated light (e.g. spike pulses) operating at multi-GHz speeds is limited by shot noise level ($\sim$20 photons). To overcome this limitation, the use of superconducting-nanowire single-photon neuron-like detectors has been proposed [23], and impressive advances are being made in integrating full neuromorphic circuits based on superconducting electronics [54, 55], but these operate at cryogenic temperatures ($<$10 K) with attendant expenses and limitations [56]. Hence, room-temperature nonlinear highly-sensitive detection solutions capable of receiving low-photon spiking signals are of crucial importance.

2.5.3. Nanophotonic integrated circuit platform for SNNs

Considering the challenges of integrating nanoscale artificial neuron light sources and nanophotodetectors in a scalable SNN platform, in sections 3–7 of this perspective, we discuss the recent results and prospects, of a novel platform, figure 3, using innovative III–V/Si materials, panel (a) (sections 3 and 4), specifically DBQW nanostructures providing unique neuron-like spike-emitting-receiving properties (panel (b), section 3). These materials are used to design nanodevices consisting of light spiking sources and detectors with sub-λ confinement (nanoLEDs, nanolasers) and nano-photodetectors (nanoPDs), panel (b) (section 5). In sections 6 and 7 of this perspective, we discuss the strategies to use these nanophotonic neurons for energy-efficient and high-bandwidth photonic interconnected synaptic links, panels (c) and (d), for applications in artificial SNNs.

3. Spiking neurons

3.1. NanoRTD neurons

Neurons exhibit excitability [58], the dynamical property that is essential for biologically inspired AI. Seeking an alternative architecture that supports spikes as information carriers, here we look at DBQW-based resonant tunnelling diode (RTD) circuits as excitable neuromorphic spike generators [59]. Typically, the active DBQW layer consists of a thin 10 nm of either aluminium arsenide-gallium arsenide-aluminium

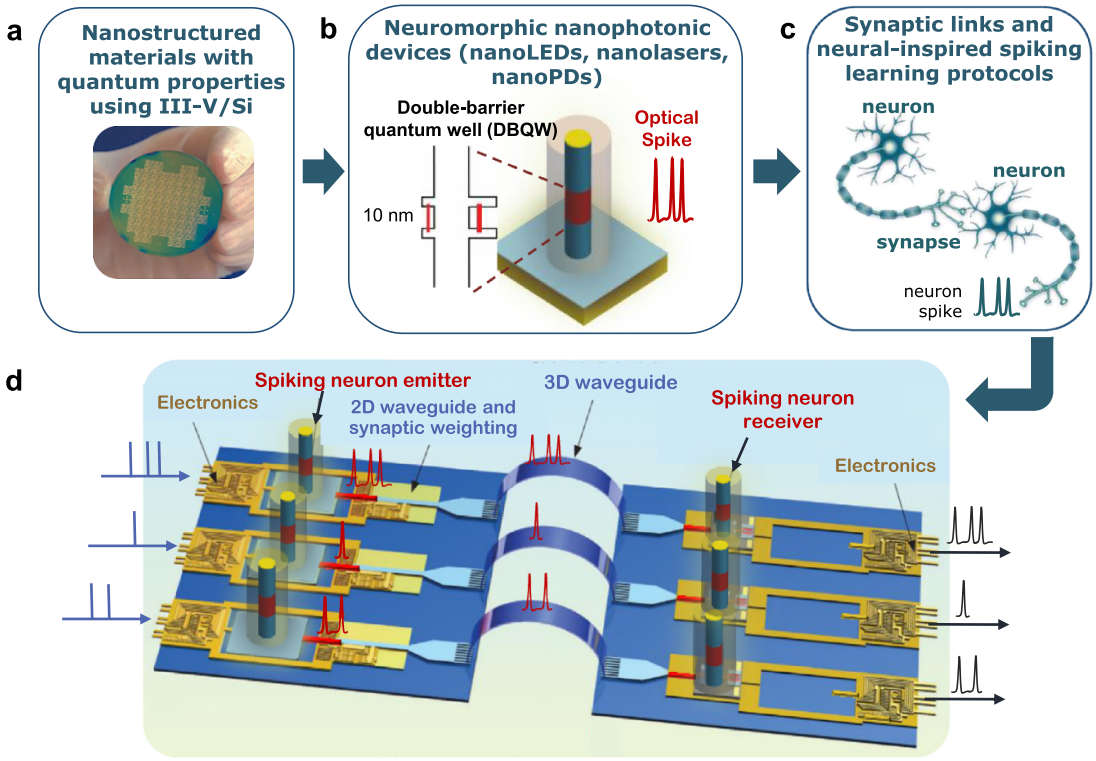

Figure 3. Implementation of a nanophotonic integrated circuit for spiking neural networks (SNNs) based on double barrier quantum well (DBQW) nanophotonic devices and interconnected (2D and 3D) synaptic links. The platform uses (a) innovative III–V/Si materials to implement (b) light spiking sources and detectors with subwavelength (sub-λ) confinement. This allows for the implementation of energy-efficient and high-bandwidth photonic interconnected neuron-inspired synaptic links, (c)-(d). The schematic in (d) only shows one-to-one device feed-forward interconnections (scheme in (d) © [2012] IEEE. Reprinted, with permission, from [57].).

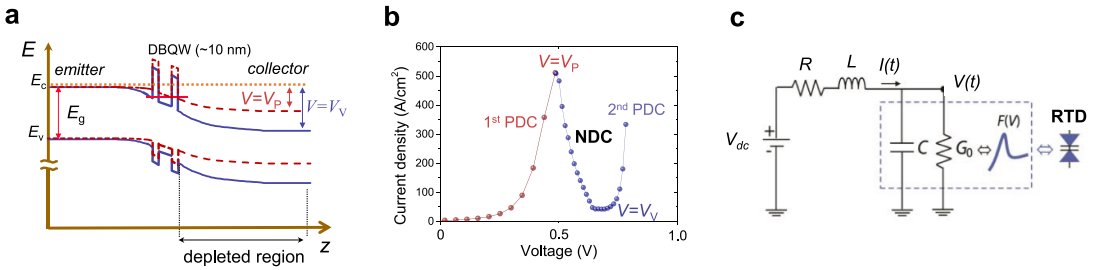

Figure 4. Overview of the key physical properties of nanoscale resonant tunnelling diodes. (a) Electronic band diagram showing the conduction and valence bands, E_c and E_g, respectively, for an applied voltage across the emitter and collector that corresponds to the peak, V_p, and valley, V_v, resonant conditions. The bandgap of the semiconductor is represented by the parameter E_g. (b) Example of a simulated current–voltage characteristic of an RTD device with positive differential conductance (PDC) and negative differential conductance (NDC) regions. (c) Circuit schematic of an RTD biased with a d.c. voltage source, V_{dc}. A nonlinear voltage-controlled current source, $F(V)$, in parallel with the equivalent capacitance, C, are used to model the circuit. The other circuit parameters include a series inductance, L, and a resistance, R.

arsenide (AlAs/GaAs/AlAs) or aluminium arsenide-indium gallium arsenide-aluminium arsenide (AlAs/InGaAs/AlAs) semiconductor heterostructure. This heterolayer is inserted in the epitaxial growth direction of the III–V semiconductor layer stack which enables 1D carrier confinement and filtering of carriers (e.g. electrons) via the resonant tunnelling effect across the quantum well's quasi-bound states (resonances), as shown in figure 4(a) [35].

NanoRTDs can be combined with photonics (e.g. LEDs, lasers and photodetectors) to process and transmit optical signals [35, 39, 60]. The DBQW nanometric size enables circuit high-speed. The resonance in the quantum well (which controls the current flow via resonant tunnelling effect) can be fine-tuned by applying a voltage across the emitter and collector sides. This resonance control provides an energy passband-like filtering effect for carriers. As a result, a current–voltage (I–V) characteristic with a pronounced NDC is achieved—also called negative differential resistance, figure 4(b). The theoretical cut-off

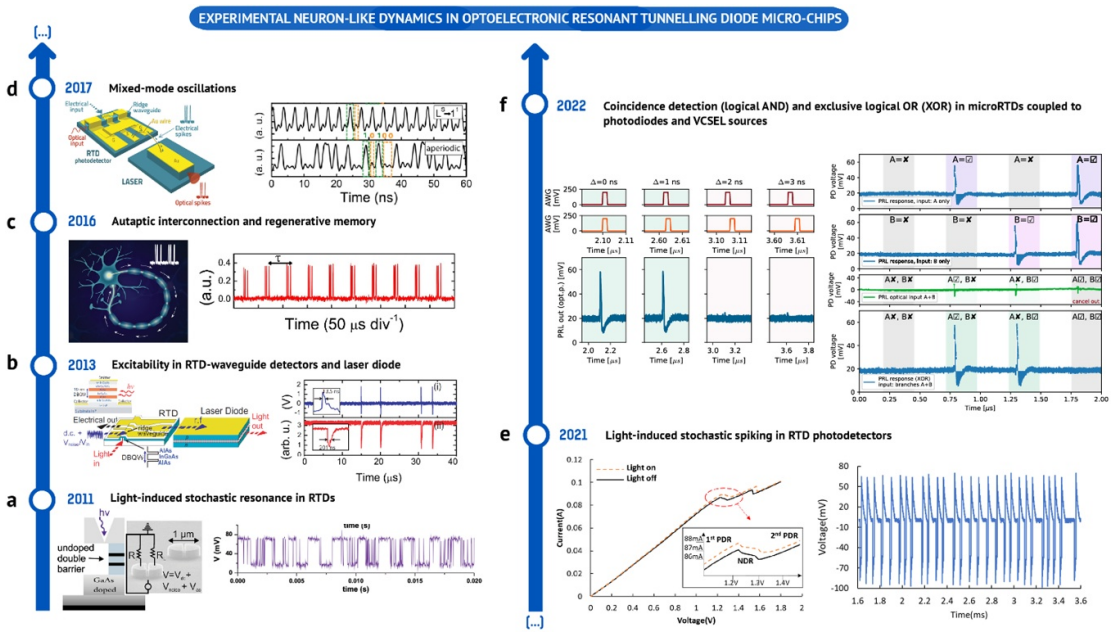

Figure 5. Examples of experimental neuron-like dynamical functions and tasks demonstrated in optoelectronic microscale RTD-based chips over the last decade. These include: (a) light-induced stochastic resonances (Reprinted from [71], with the permission of AIP Publishing.), (b) electrical and optical neuron-like spiking and bursting generation (Reprinted with permission from [21] © The Optical Society.), (c) autaptic (self-feedback) photonic neuron regenerative memory (Reproduced from [39]. CC BY 4.0.), (d) optical mixed-mode oscillations (Reproduced from [60]. CC BY 4.0.), (e) light-induced stochastic spiking demonstrated in RTD photodiodes (Reproduced from [72]. CC BY 4.0.), and (f) demonstration of in-device coincidence detection (logical AND), (right), and exclusive logical OR (XOR) tasks (left) (Reproduced from [73]. CC BY 4.0.).

frequency of a single DBQW is limited by the tunnelling escape time in the quantum well [35]. Notably, the resonant tunnelling effect and NDC characteristic, when combined with a suitable circuit design, figure 4(c), have been exploited in a wide range of nonlinear applications, such as THz oscillators, emitters and detectors for imaging [61], and wireless communications beyond 5 G [36].

Importantly, the DBQW can be epitaxially integrated with other active materials (either based on III–V or III-N), and RTD-based photodetectors and light sources have been demonstrated. Notable examples include single-photon switches and detectors [37, 38], near-infrared photodetectors for optical communications [62], mid-infrared detectors for sensing [63], III-nitride LED sources [64], III–V unipolar (n-type) bistable light-emitting RTDs [65], bipolar (p–n-type) RTD-based LEDs [66–68] and RTD-lasers [69, 70]. These works open unique opportunities to combine the nonlinear electrical properties of RTDs with the optical characteristics of photonic devices to create novel architectures of interest for emergent neuromorphic optical computing systems.

For neuron-like computation, early works evoked RTD-based devices as nanoelectronic candidates for cellular neural networks as a form of threshold logical gates [74]. In the context of photonics, reports on optoelectronic RTDs integrated with photo-sensitive absorption layers and lasers showed potential for bio-neural functionalities. Figure 5 summarizes the main experimental works reported in the last decade that demonstrate relevant nonlinear neural dynamic functions of optoelectronic microRTDs. The most remarkable examples include the use of bistable RTD photodetectors for light-induced stochastic resonance phenomena (figure 5(a) [71]), excitability in photosensitive RTD-laser hybrid integrated circuit devices for spiking and bursting (figure 5(b) [21]), autaptic (self-feedback) neuron-like signalling (figure 5(c) [39]), and mixed-mode neuron-like oscillations (figure 5(d) [60]). More recently, light-induced stochastic spiking was demonstrated in RTD photodiodes (figure 5(e) [72]), and in-device coincidence detection (logical AND), and exclusive logical OR (XOR) tasks were demonstrated using an opto-electro-optical (O/E/O) artificial neuron built with an RTD coupled to a photodetector as a receiver and a vertical cavity surface emitting laser as a transmitter (figure 5(f) [73]). In what follows, we present the key nonlinear dynamic functions of neuromorphic RTDs biomimicking neurons.

3.2. Neural dynamics in nanoscale RTDs

Figure 4(c) depicts the equivalent circuitry of a DBQW RTD driven by a d.c. bias supplied by an external source of voltage. The dynamic properties of this nonlinear system are described by the following differential equations,

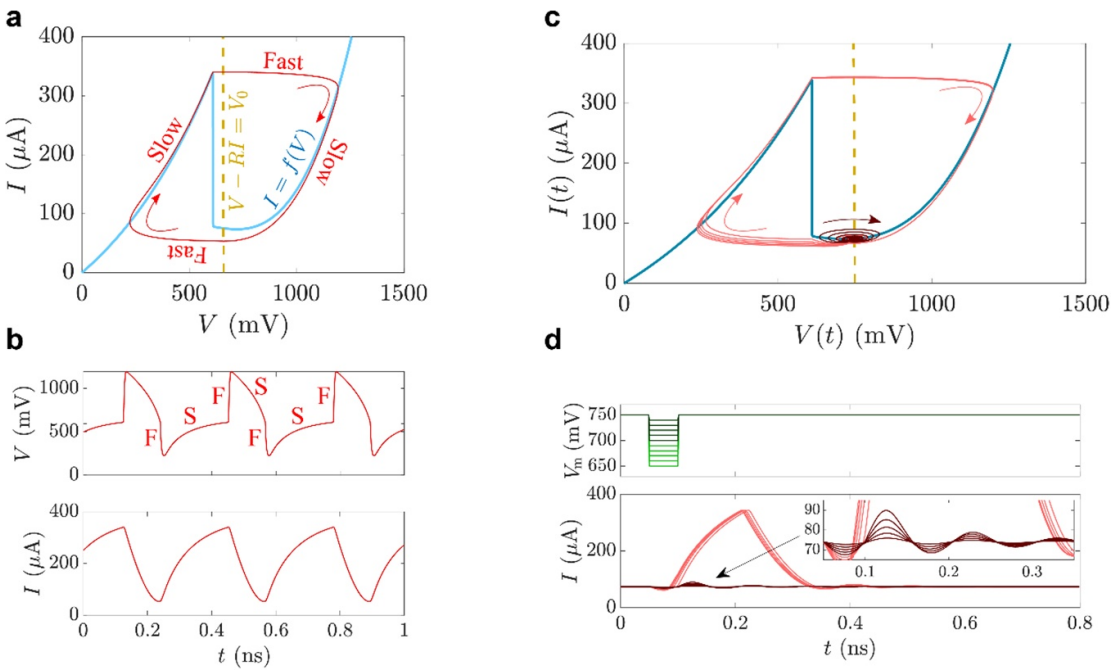

Figure 6. Simulation of equations (1)–(3) with $V_m(t) = V_0 = 650\,\text{mV}$ and $\sigma = 0$. The system's response, I–V characteristic and circuit's load line are coloured in red, blue and yellow, respectively. (a), (b) The system exhibits self-oscillations with stages of slow and fast dynamics: (a) Phase space; (b) Voltage and current over time. (c), (d) Simulations of equations (1)–(3) with $V_0 = 750$ mV and $\sigma = 0$, driven by 50 ps long input negative square voltage pulses of varying amplitudes. (c) System's response to input pulses on the phase space. (d) On top is shown the voltage input pulse profiles. The system's current response to each input pulse is shown at the bottom. Sub (supra) threshold pulses and their responses are coloured in dark (light) shades of green and red, respectively. (Reprinted (figure) with permission from [59], Copyright (2021) by the American Physical Society.).

$$C\frac{\mathrm{d}V}{\mathrm{d}t} = I - F(V) + \sigma\xi(t) \tag{1}$$

$$L\frac{\mathrm{d}I}{\mathrm{d}t} = V_m(t) - V - RI. \tag{2}$$

Here, V, I are the voltage and current across the RTD, respectively. The parameters R, C, L represent the circuit's intrinsic resistance, capacitance and inductance, respectively. Unless otherwise specified, their values for the remainder of this Section are $R = 10\,\Omega$, $C = 2\,\text{fF}$, $L = 126\,\text{nH}$. The parameter $V_m(t)$ represents the external driving voltage (which includes both a d.c. and a time-varying signal). The intrinsic fluctuations in the system are accounted for by a time-uncorrelated white noise function, $\xi(t)$, and a noise intensity, σ. The function $F(V)$ is the RTD nonlinear I–V characteristic. As previously described, incident electrons may resonate with the confinement energy levels in the quantum well due to the DBQW structure (figure 4(a)), depending on their Fermi energy levels. As a result, the I–V characteristic is locally maximised and exhibits one or more regions of NDC between regions of positive differential conductance (PDC). Schulman *et al* provides an expression for $F(V)$ derived from the Fermi–Dirac statistics [75],

$$F(V) = a\ln\left(\frac{1+e^{\frac{(b-c+n_1V)q}{kT}}}{1+e^{\frac{(b-c-n_1V)q}{kT}}}\right)\left[\frac{\pi}{2} + \tan^{-1}\left(\frac{c-n_1V}{d}\right)\right] + h\left(e^{n_2qV/kT} - 1\right) \tag{3}$$

where q is the electron charge, κ is the Boltzmann constant and T is the temperature. The remaining a, b, c, d, n_1, n_2, h parameters tune the shape of the I–V curve and can be, for instance, determined by fitting the experimental data. Figure 6(a) illustrates an example of an I–V characteristic with an NDC region embedded between two PDC regions. The limiting peak and valley points are, $V_V = 609.63$ mV and, $V_P = 720.65$ mV, respectively. In what follows, using equations (1)–(3) we summarize some of the nonlinear dynamical regimes found in nanoscale RTDs and of interest for neuromorphic spike-based processing and computation.

3.2.1. Bistability and slow-fast oscillation dynamics

It is well known that an RTD subjected to a d.c. voltage source (i.e. $V_m(t) = V_0$) exhibits a fixed output when biased in the PDC region, and self-oscillations when biased in the NDC region [76]. These correspond to a

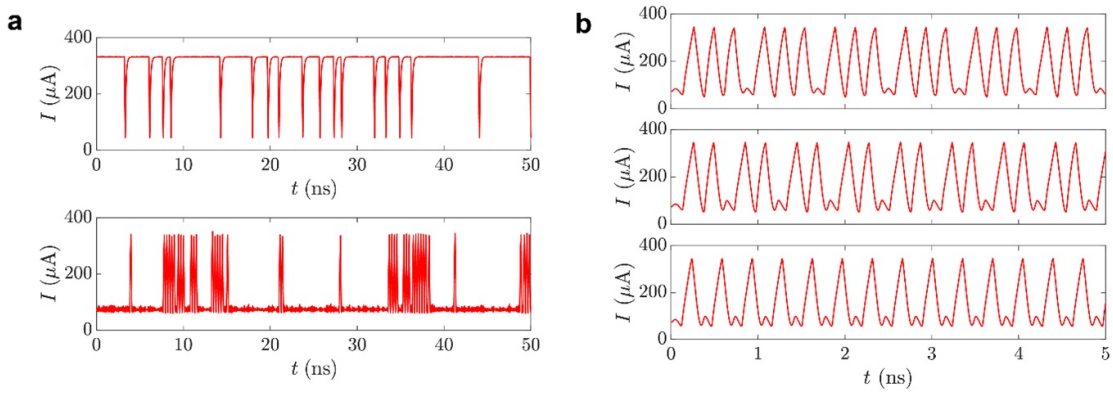

Figure 7. Simulated sochastic spiking and bursting, and mixed-mode oscillations. (a) Simulation of spikes fired randomly using equations (1)–(3). The bias was set close to the NDC region and the noise intensity was non-zero. Top: $V_0 = 605$ mV, $\sigma = 4.52 \times 10^{-12}$ A s$^{1/2}$; Bottom: $V_0 = 725$ mV, $\sigma = 8.03 \times 10^{-12}$ A s$^{1/2}$. (b) Simulation of mixed-mode oscillations using equations (1)–(3). The system was driven with an input sinusoidal voltage, $V_m(t) = V_0 + V_{\text{amp}} \sin(\omega t)$, where $V_0 = 730$ mV, $V_{\text{amp}} = 100$ mV; Top: $\omega/2\pi = 30.9$ GHz; Middle: $\omega/2\pi = 31.7$ GHz; Bottom: $\omega/2\pi = 36.3$ GHz.

fixed-point attractor and a limit cycle, respectively. Depending on the system parameters and the *I*–*V* characteristic, there may also be a narrow bias range near the peak and valley points where both types of solutions coexist, resulting in *bistability* [59, 60, 77]. If the stiffness coefficient, $\mu = \sqrt{C/L}$, of the circuit is much smaller than its critical conductance, defined as the absolute value of the minimal differential conductance in the NDC region, the self-oscillations regime exhibits two slow stages and two fast stages, as shown in figure 6(a) [59, 60, 77]. During the slow stages, the orbit remains close to the *I*–*V* characteristic in the PDC regions until it reaches either its peak or valley, at which point it quickly leaps to the other PDC region. For the values of C, L chosen above and for the parameter $\sigma = 1.26 \times 10^{-4}\,\Omega^{-1}$, self-oscillations with slow-fast dynamics are obtained, as illustrated in figure 6(b).

3.2.2. Excitability

A slow-fast RTD is demonstrated as an excitable spike generator when biased in either PDC regions, but proximal to the NDC region, either close to the peak or the valley of the *I*–*V* characteristic, and introducing a perturbation, e.g. a square voltage pulse. This is equivalent to displacing the load line for a fraction of time. If the pulse amplitude/width is sufficiently large, the load line enters the NDC region and the RTD undergoes a series of transitions, figure 6(c). The system then returns to the fixed point attractor in the PDC region. In response to sub-threshold perturbations, the system quickly returns to the fixed point. Figure 6(d) shows responses of an RTD biased near the *I*–*V* curve valley and subjected to negative square voltage pulses (thus the load line is displaced to the left in figure 6(c)). Indeed, as shown in figure 6(d), there is a pulse amplitude threshold of 50 mV that distinguishes excitable pulses from negligible responses (the so-called *all-or-nothing* response [12]). Similarly, when biased close to the *I*–*V* curve peak and subjected to positive square voltage pulses, the RTD exhibits excitable pulses [59].

3.2.3. Stochastic spiking and bursting

An RTD may fire spikes spontaneously if biased in the proximity of the NDC region and subjected to a non-zero level of noise (σ). As noise is permanently perturbing the circuit, the excitability threshold may be exceeded, triggering the excitable response. The higher the noise level, the more likely the latter will occur and the more frequent spikes will be fired [21, 59]. When the RTD is biased close to the peak of the *I*–*V* characteristic, isolated spikes seem to appear following an exponential statistic that depend on the noise level, figure 7(a), top panel. However, when the bias is set close to the valley, the spikes appear agglomerated in *bursts*, as shown in figure 7(a), bottom panel. Experiments have also revealed stochastic bursting [21]. The causes of this phenomenon are elucidated in [59]. It is related to the wide basin-like profile of the *I*–*V* curve in the valley. The latter leads to a bistability regime between the periodic oscillating solution of the RTD and the quiescent steady state in the valley. When an excitable orbit is triggered by noise, the system after firing the excitable response may remain trapped for some time within the attractor of the limit cycle, which appears as a bursting response that consists of an integer number of peaks.

3.2.4. Mixed-mode oscillations

When an RTD is subjected to an a.c. voltage signal, it may exhibit mixed-mode oscillations (MMOs), which are oscillations with various amplitudes that occur in sequence. This is due to the a.c. modulation which then

Figure 8. (a) AIXTRON G4 MOVPE Planetary Tool (courtesy of AIXTRON SE). (b) EpiCurve *in-situ* monitoring tool (courtesy of LayTec AG).

drives the nonlinear dynamical system alternately into regions in the space of parameters where it responds with self-oscillations and fixed point with relaxation-oscillations decay or, alternatively, self-oscillations with different sizes [12, 60]. The pattern of small and large oscillations is determined by the amplitude and frequency of the driving signal. Figure 7(b) shows examples of MMO responses to sinusoidal signals from the RTD. Depending on the frequency, 3–1, 2–1 and 1–1 patterns (from top to bottom, respectively, in panel (b)) of small and large amplitude oscillations are seen. Experiments have shown that MMOs have potential applications in high-frequency transmission of binary code [60].

4. III–V/Si DBQW structures for nanophotonic neurons using InP, GaAs, and Ge-on-Si substrate materials

The DBQW-based epi-structure designs of the nanophotonic neurons are discussed in section 5. The epi-materials for the optical neuron nanodevices are obtained using homo-epitaxial growth on InP and GaAs, and also hetero-epitaxial growth on Ge-on-Si substrate materials to enable low-cost substrate solutions. The structures produced included nanoLEDs, nanolasers, nanophotodetectors, RTDs and integrated structures of these. In what follows, we briefly describe the growth of these materials, which are fully compatible with existing semiconductor foundries of III–Vs, thus enabling a scalable neuromorphic semiconductor nanotechnology.

Growth of the III–V elements are performed by MOVPE in a semiconductor foundry (IQE.plc), using an Aixtron G4 epitaxial reactor, as shown in figure 8(a), using metalorganic trimethyl- precursors of gallium, indium and aluminium and using Group V hydrides, arsine and phosphine. Deposition is performed as required on substrate diameters from 76 mm for InP structures to 150 mm for GaAs and GaAs/Si. Standard growth conditions are used for III–V material deposition, where reactant concentration control, by reactor *in-situ* monitoring (see figure 8(b)), and a suite of ex-situ characterisation tools, were all enabled to ensure accurate control and conformance to the epitaxial specifications targeted.

Development of the Ge-on-Si process is performed on an ASM epsilon CVD tool in a semiconductor foundry (IQE.plc), enabling a germanium surface whose lattice parameter was close to that of GaAs. The lattice mismatch between silicon and germanium results in high levels of defects that propagate from the interface to the surface—these are measured as the threading dislocation density. As an example, the main challenge of the epitaxial growth of III–V GaAs on Ge-on-Si templates (specifically DBQW nanostructures with reproducible thin layers) is to improve the growth conditions in order to minimize surface defects on the germanium, prior to the overgrowth of the GaAs element.

5. Nanophotonic neuron emitter and receiver spiking nodes

5.1. NanoLED neuron emitter: nanoRTD-LED

We introduce a nanoscale nanophotonic neuron circuit formed by a DBQW nanostructure integrated into a subwavelength metal-dielectric cavity nanolight-emitting diode (nanoLED). The key neuron-like functions of this nanoRTD-LED emitter architecture was first proposed in [78]. In this Section, we summarize the main properties of the nanoLED neuron emitters, as well as recent advances reported elsewhere in realizing

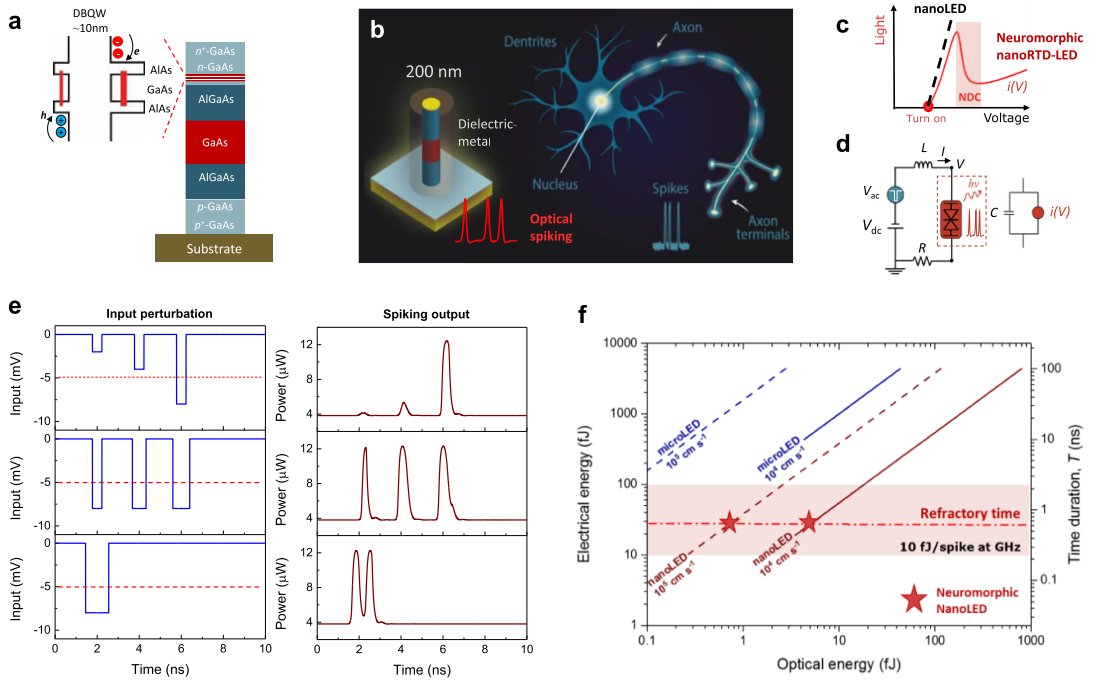

Figure 9. Static and dynamic properties of nanoLED neuron emitters. (a) Semiconductor epitaxial heterostructure schematic consisting of a DBQW nanostructure and a AlGaAs/GaAs/AlGaAs light-emitting diode (LED) heterostructure. In the inset is shown the 10 nm thick DBQW AlAs/GaAs/AlAs nanostructure inserted in the n-type section of the LED stack. (b) Schematic of the nanophotonic artificial spiking neuron in a nanopillar metal-dielectric architecture. Also shown is the schematic of a biological spiking neuron. (c) Comparison of the light-voltage (L–V) curves between a standard nanoLED (dashed black line) and a neuromorphic nanoRTD-LED (solid red line). (d) Schematic of the nanoRTD-LED circuit modeled by a nonlinear voltage-controlled current source, $i(V)$, in parallel with the equivalent capacitance, C. (e) Optical spiking signatures of the nanoRTD-LED (*right*) when perturbed by various electrical (sub-10 mV) pulsed signals (*left*). (f) An analysis of the performance of standardl and neuromorphic LEDs that demonstrates how electrical energy varies with optical energy per spike. The diagonal lines are the values for the micro- and nanoLEDs operating in a non-spiking regime (i.e. pulsed current-driven scheme). The dashed lines represent the performance for poor passivated LED devices while the solid lines represent improved surface passivated LED devices (see discussion regarding passivated nanopillars in figure 10). The dashed-dot horizontal line intersecting the diagonal traces indicates the refractory time, corresponding to the electrical and optical energy per spike (stars) of the neuromorphic nanoRTD-LED. (Reproduced from [78]. CC BY 4.0.).

and improving such nanoemitters, namely increasing their output power enabled by light extraction [79], and surface passivation [80] methods.

The design of the nanoLED neuron emitter takes advantage of the unique physical properties of DBQW nanostructures, as discussed in section 3. The nanoLED neuron emitter consists of an n-type DBQW (AlAs/GaAs/AlAs material system) monolithic integrated with an inverted p–i–n-type GaAs/AlGaAs nanopillar LED and is shown schematically in figure 9(a) (although other III–V material systems could be considered, as in the case of the nanolaser discussed in section 5.2). The DBQW enables control of the electron injection into the nanopillar active region of the LED. The resulting nanoLED has an NDC characteristic, figure 9(c), that controls the all-or-nothing optical spiking response, figure 9(e), of the nanoLED. In contrast to the light-voltage (L–V) characteristic (black dashed line in figure 9(c)) of typical LEDs, this results in a voltage-controlled NDC in the L–V curve (red solid line in figure 9(c)). Due to the Purcell enhancement of the spontaneous emission [81], the dielectric-metal nanocavity of the nanopillar, figure 9(b), offers strong light–matter interaction at the nanoscale, leading to faster and more efficient light emission [82, 83], a key feature for achieving high-bandwidth optical spiking.

The dynamic model used to analyse the optical spiking dynamic properties of the nanoRTD-LED neuron emitter was first reported in [78], and combines the differential equations of the optoelectronic lumped circuit, figure 9(d), which take into account the nonlinear voltage-controlled current characteristic of the nanoRTD, figure 9(c), and the rate equations of the nanopillar LED. Figure 9(e) shows an example of numerically simulated optical spikes at multi-gigahertz speeds induced by an input perturbation. It shows some of the characteristics of excitable systems such as all-or-nothing response and refractory time. The all-or-nothing optical spike signals are activated using extremely low (sub-10 mV) electrical pulse signals. This results in a nanoLED neuron emitter with ultralow switching energy (here in the range of 10–100 fJ per emitted spike). Figure 9(f) compares the operation of the neuromorphic nanoRTD-LED with a typical current-driven pulsed (non-spiking) nanoLED source in terms of electrical and optical energy per spike. We

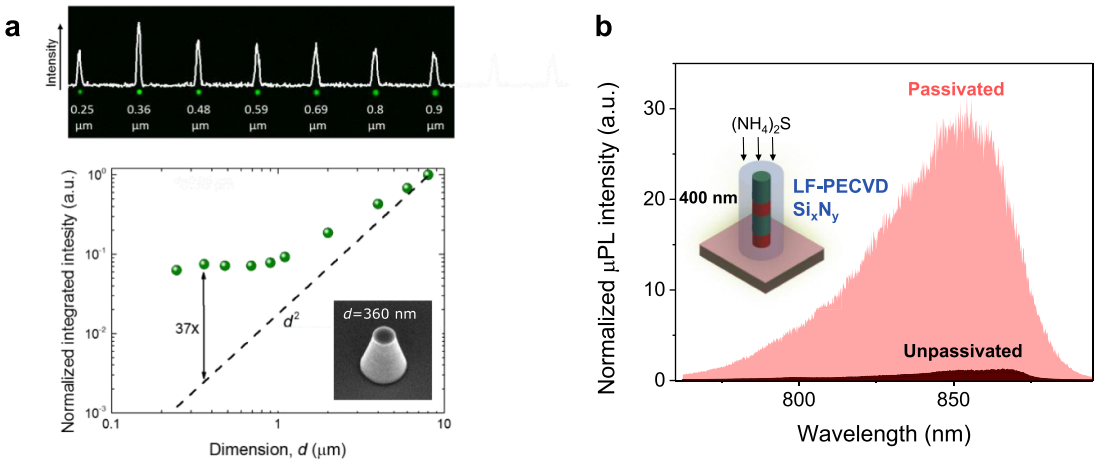

Figure 10. Results from experiments demonstrating quantum efficiency improvements in nanopillar LEDs using surface passivation and light extraction methods. (a) Microphotoluminescence results for nanopillars ranging from 0.25 μm to 0.9 μm, along with confocal microscopy images and intensity profiles are shown on top. The bottom of the image displays the integrated intensity plot as a function of diameter. The dashed line indicates a d^2 dependence (Reprinted with permission from [79] © The Optical Society.). (b) Experimental microphotoluminescence results showing the spectra from a single nanopillar with a width of around 400 nm for a unpassivated sample (red dark colour) and a passivated sample (pink light colour). The passivation treatment used ammonium sulphide followed by Si_xN_y coating deposited by low-frequency PECVD. A schematic of the best treatment procedure is shown inset, displaying the pillar coated with a Si_xN_y dielectric (Reproduced from [80]. CC BY 4.0.).

consider two case scenarios: (i) with and (ii) without an improved surface passivation method. A typical microLED (blue traces) operates with electrical energy well above 100 fJ/spike and displays modulation bandwidths well below 1 GHz, limited by the spontaneous recombination process. Instead, Purcell-enhanced nanoLEDs (red traces) are particularly well-suited for operation at multi-gigahertz speeds in the 10–100 fJ/spike range, especially when their surfaces have been passivated to reduce non-radiative effects in nanopillars (see also figure 10). We observe there is a clear trade-off between the electrical and optical energy per pulse produced in conventional LEDs. However, for the neuromorphic nanoRTD-LED (star symbols) the electrical and optical energy per emitted spike is given by the intersection of the dashed-dot horizontal red line and the diagonal traces (standard nanoLEDs) (assuming a refractory time of ∼650 ps for the examined spiking nanoRTD-LED). We observe the energy per spike is nearly constant and independent of the incoming modulating frequency signal, in contrast to standard pulsed current modulation schemes. We note that the all-or-nothing response, which activates the spiking response with a constant refractory time for a given input above a certain threshold, is the cause of this important characteristic of the spiking nanoRTD-LED. Finally, since RTD-based photodetectors have already been demonstrated [72, 84], the photosensitive characteristics of DBQW structures could be used to create optically activated neuromorphic nanoLEDs, and the neuron receiver architecture is covered in more detail in sections 5.3 and 5.4.

Despite the promising spiking properties of nanoLED neuron emitters, experimentally demonstrating power-efficient III–V nanoLEDs with reasonable large optical output, and with such neuromorphic (spiking) properties, remains a significant challenge. In recent years, advances in the quantum efficiency of nanoLEDs have been made using several III–V materials. The miniaturization approach relies on the use of cavities such as photonic crystals [26], metal-dielectric [27] or plasmonics [29], allowing the realization of wavelength and subwavelength scale devices. These advancements are creating expectations that nanoLEDs will be both efficient and fast, outperforming nanolasers [85, 86]. However, to date the external quantum efficiency (EQE) of III–V nanoLEDs at room-temperature remains limited to values below 1%, resulting in ultralow output powers (in the nW or even pW range) [26, 27, 29], and impractical optical systems. Taking the example of III–V nanopillars considered in our neurons, and neglecting losses related to metallic structures in metal-dielectric or plasmonic nanocavites, the main reasons for the extremely low EQEs are two-fold. Firstly, efficiently coupling the light output to a nanowaveguide [27], or a plasmonic waveguide [29], remains a challenge when the area of the light source is reduced to the deep sub-μm^2 range. Therefore, new light extraction strategies are of key importance. Secondly, as the surface-to-volume ratio increases significantly at these small scales, non-radiative effects in III–V materials, specifically surface-related properties, become more important. In what follows, we provide a concise review of the results recently reported on light extraction and surface passivation in GaAs-based light-emitting materials integrated with DBQW nanostructures (the key architecture of the nanoLED neuron emitter).

To begin, we will review recent developments on light extraction in nanostructured GaAs-based LEDs. Despite extensive research in out-coupling methods employing 2D photonic crystals [26], and nanowaveguides integrated with grating couplers [27], these approaches are difficult to implement when the size of the light-emitting structures is drastically reduced to the deep-subwavelength ($\ll\lambda^3$) scale. In a recent work, a strong enhanced signal in vertical-emitting undoped AlGaAs/GaAs/AlGaAs tapered pillars in a GaAs substrate has been demonstrated [79], when the emitting nominal area is decreased to the sub-μm scale. We note that identical GaAs-based materials were used, as previous discussed in the architecture of the neuromorphic nanoLEDs. The work used pillars ranging from 200 nm to 8 μm lateral width. These were fabricated using e-beam lithography and dry etching techniques, and characterized using a micro-photoluminescence (PL) microscope with $\lambda = 561$ nm laser excitation. Figure 10(a) shows confocal images examples (top) of the emission for optically pumped nanopillars (the respective intensity profiles are shown inset). As shown in the bottom plot, presenting the integrated intensity as a function of diameter, for the case of micropillars, the light emission is reduced as the diameter decreases following a typical scaling law, d^2, of planar LEDs. However, as d is reduced from 4 μm to 0.2 μm sizes, particularly in the range of 300 nm $< d <$ 400 nm, although the nominal emission area is reduced by a factor of more than 100, the intensity is reduced only by ~10 times. For example, the emitting intensity peaks for pillars with $d = 360$ nm and the respective integrated intensity is comparable to pillars with $d \sim 1$ μm. This strongly deviates from the d^2 dependence observed for micropillars, resulting in a 37-fold enhancement of emission. This striking effect has a large effect on the EQE. The vertical-emitting nanopillar-based LEDs have the potential to achieve notable large EQE ~45%, whereas the efficiency of μm-pillar planar LEDs, without further methods, are limited to EQE ~2%. These results offer a versatile method of light-outcoupling in nanostructures with prospects to improve the performance of interconnected nanoLED devices which are of key importance for efficient photonic neural network architectures.

Lastly, we review the non-radiative effects in nanopillar structures for the same GaAs-based materials previously described. Typically, large densities of electronically active defects can be found on the surfaces of GaAs-based compounds and at their interfaces with dielectrics, namely because of dry etching during the patterning of the nanopillars. Because of the large surface-to-volume ratio, this is especially important when the devices are scaled to the deep subwavelength size ($\ll 1$ μm). Passivation of GaAs with silicon nitride (Si_xN_y) has been proposed to improve GaAs electrical performance [87]. Recently, it has been demonstrated that surface passivation using a combination of ammonium sulphide chemical treatment followed by encapsulation with a thin layer of Si_xN_y (80 nm) deposited by low-frequency plasma enhanced chemical vapour deposition (LF-PECVD) can result in improved optical properties of nanopillar LEDs [80]. The results show that passivated GaAs nanopillar surfaces are robust, stable, and long-lasting. A micro-PL setup was used to collect the PL of fabricated nanopillars. Figure 10(b) depicts the passivated and unpassivated PL spectra of a representative pillar (width of $d = 400$ nm). When compared to the unpassivated pillar, the Si_xN_y coated sample exhibits a 29-fold increase in the PL integrated intensity. This improvement was measured for all nanopillar sizes (not shown) and is due to hydrogen species attack on the GaAs surface, which reduces the density of surface states significantly, as reported in the LF-PECVD process. These findings, when combined with efficient light extraction methods, could lead to significant EQE improvements in nanostructures, which is critical for the miniaturization of neuromorphic nanoLEDs. Further, since the DBQW structures reported here use Al layers for the barriers, the passivation treatment is of key importance for both optical and electrical properties of the nanodevices. Other challenges to overcome in order to realize fully functional nanoLEDs include the fabrication of low-resistance ohmic contacts, and the demonstration of low-loss metal-dielectric subwavelength cavities that could offer a significant light–matter interaction.

5.2. Nanolaser neuron emitter: nanoRTD-LD

The monolithic integration of a nanolaser with an RTD would enable electro-optical spiking functionality with a small footprint and low power operation. This, however, poses several challenges. Firstly, to achieve lasing, in comparison to the nanoLED case analyzed in the previous Section, several additional aspects needed to be addressed: (i) achieving the appropriate resonant wavelength, (ii) having enough gain while minimizing losses in the cavity. Achieving resonant wavelength is not trivial, as the nanocavity sizes are of the order of magnitude of the wavelength. When using high-resolution lithography process (e.g. electron-beam lithography), innaccuracies of a few tens of nanometers are to be expected, which results in a large resonant wavelength detuning (e.g. 30 nm offset in the cavity size translates into 100 nm resonant wavelength of a cavity for the case of an InGaAs active region). A significant offset between the resonant wavelength and the material gain can cause a limited modal gain.

High-reflectivity mirrors are essential to maintain moderate mirror losses and keep the lasing threshold within reach. In principle, metal mirrors can provide high reflectivity close to the unity [88], however the net reflectivity back into the resonant mode is highly dependent on the verticality of the reflecting facets, which

are defined with reactive-ion-etching methods. A small sidewall angle of 5° can reduce net reflectivity to 83% [88]. Another important aspect is that lasers can only be used in planar photonic integrated circuits if they are waveguide-coupled. This brings an additional optical loss mechanism into the cavity, which can compromise the cavity quality factor and prevent lasing. Therefore, controlling the waveguide coupling strength is critical. Weak and controlled waveguide coupling of similar structures can be achieved through evanescent coupling [27].

Regarding electrical-related challenges, surface recombination and ohmic contact losses are the most relevant. Non-radiative surface recombination (discussed also in section 5.1) limits the amount of carriers contributing to optical gain, and plays an important factor in nanocavities due to their high surface-to-volume ratio. Adequate surface passivation methods are required to address this. Moreover, nanocavities have a small horizontal cross section and therefore suffer from high ohmic contact resistance. An alternative device configuration for reduced contact resistance is the incorporation of a tunnel junction to enable both sides of the diode with an n-type contact [89]. Excessive contact resistance leads to pronounced Joule heating, which red-shifts the gain curve producing a thermal roll-off. A strategy for heat sinking is therefore very important.

In addition to the fabrication of the nanolaser device, the integration of a DBQW to provide nonlinearity should be considered. In this regard, if the position of the DBQW is misplaced in the laser diode layer stack, it may induce additional electrical or optical losses. In our approach, the DBQW-based RTD is embedded in the InGaAs contact layer to keep the electrical and optical functions as separate as possible. Additional RTD-related design considerations include matching the electrical nonlinearity with the laser operational regime. The main issue here is that RTDs are voltage-driven devices, whereas lasers are current driven devices. For a laser to operate in the spiking mode, the NDC region (i.e. peak and valley currents) must be located between the lasing current threshold and the thermal roll-off of the device. Finally, as discussed in section 3.2, controlling capacitance in RTDs is important for tuning their dynamic characteristics. It is therefore ideal that the device design includes a mechanism for adjusting the capacitance, so that the correct electrical regime can be targeted.

Several approaches for integrating a laser with RTDs have previously been reported. A hybrid integration strategy [21] simplifies the epitaxial growth and processing of components; however, the footprint is compromised, and wired connections may introduce additional inductance to the experimental circuit while sacrificing operation speed. Monolithic integration optimizes the device's footprint and minimizes external electrical parasitics, and has been demonstrated experimentally with quantum well lasers on GaAs [70] and InP [69]. However, monolithic devices integrating an RTD with MQW lasers typically have thousands of μm^2 active area, which leads to limited operation speeds compromised by the large capacitance values.

The integration of an RTD with a nanoscale optical device can allow for high-speed operation and a remarkable integration density of interest for compact neuromorphic systems. Various nanolasers have been demonstrated in recent years. The main concepts rely on the use of metal-cladding cavities [90], or photonic-crystal confinement [91]. Such nanoscale light sources have shown interesting performance in terms of efficiency and speed, and therefore are promising for applications that require high integration density and low power operations. A nanolaser with nonlinear performance will broaden the potential application domains, especially in the neuromorphic field.

Realising the spiking nanolaser functionality on an integrated platform requires significant design freedom. An InP-based platform is advantageous due to light-emitting functionality in the infrared region, and a variety of demonstrated passive devices. InP membranes on silicon (IMOS) allow for a nanophotonic platform with active and passive functionalities, as well as the potential for future integration with electronics [92]. IMOS also allows for double-side processing, which adds a degree of flexibility to fabrication processes.

Here we discuss our proposed design of a nanolaser-RTD. A metal-cavity nanolaser structure (figure 11) is created by a semiconductor nanopillar encapsulated in a dielectric-metal shell. The epitaxial layer stack of this nanopillar forms a p–i–n diode, allowing for radiative recombinations to take place in the intrinsic InGaAs layer. Electrical injection is carried through highly doped InGaAs contact layers. Double-side processing allows direct light extraction through an optical window. The position of the DBQW structure within the n-InGaAs contact layer is selected so that it does not overlap with the optical mode.

The fabrication of electrical contacts on both sides of the InP membrane provides capacitance tuning functionality, which is critical for achieving RTD-based excitability [59]. We fabricated and tested RTD devices with a layer stack that included an AlAs/InGaAs DBQW structure embedded in a highly doped n-InGaAs layer (inset of figure 12). As shown in figure 12, micro-sized devices of various radii exhibit consistent peak current densities of around 30 kA cm^{-2} and peak-to-valley current ratios of up to 13. For the design of nanolaser devices, it is important that such peak current density value exceeds the lasing threshold.

Scaling down photonic devices on chip would open the possibility for large-scale neuromorphic photonics with low energy consumption and high integration density, provided that fabrication tolerances

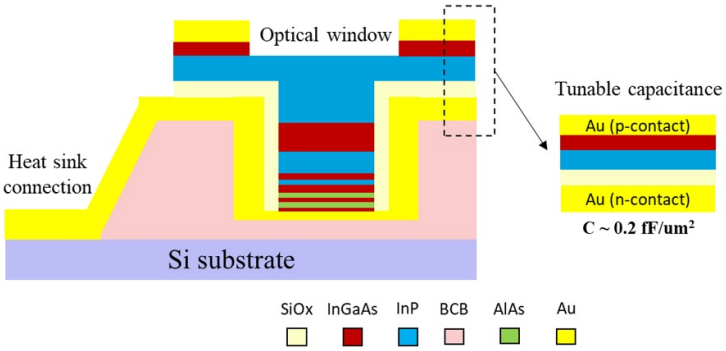

Figure 11. Flipped cavity nanolaser with RTD device on InP membrane on silicon (IMOS) platform. Inset on the right is shown the layer stack of contacts-related capacitor structure.

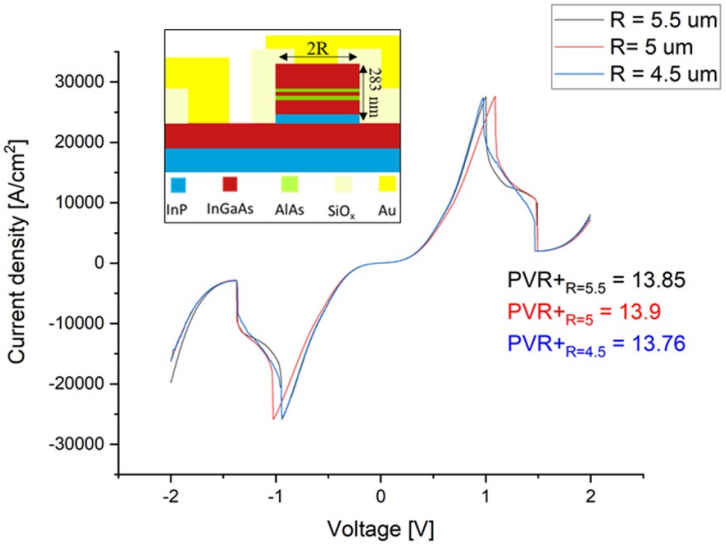

Figure 12. Experimental current–voltage characteristics of AlAs/InGaAs microRTD devices on the InP platform with circular cross section and different radii (R). The respective peak-to-valley (current) ratio (PVR) for each R is displayed.

are well controlled. For example, surface imperfections in the active medium cause additional surface recombination losses, effectively increasing threshold current density and thereby compromising the device efficiency. This is particularly critical in nanolasers because of the high surface-to-volume ratio of the active medium. This effect can prevent a nanolaser device from reaching the lasing threshold. To achieve lasing in nanoscale structures, efficient passivation of surface recombination effects should be implemented. In recent experiments (figure 13), we have demonstrated that combining ammonium sulphide treatment, SiO_x encapsulation, and high-temperature post-annealing improves surface recombination velocity by a factor of 19.

Other fabrication non-idealities that can sacrifice the cavity quality factor include non-vertical pillar slopes, which introduce additional reflection losses. Such mirror losses become more dominant in smaller devices, significantly increasing the threshold gain required to overcome the losses. Finite-difference time domain (FDTD) simulations (figure 14(a)) of a 1 μm long nanolaser cavity show that for a slope angle of 5° the threshold gain increases by 3.5 times when compared to vertical sidewalls. Recently, our etched nanolaser pillar tests (figure 14(b)) revealed high verticality slopes (∼1.6°), which can enable a low-threshold in experimental nanolasers.

Finally, it is important to implement an appropriate thermal management strategy to prevent nanolaser overheating, which causes thermal roll-off at lower current levels (before lasing). In the IMOS platform, since the polymer bonding layer introduces thermal isolation between the photonic membrane and the Si-substrate, we propose connecting the devices to the substrate through thermal vias (figure 11) for efficient

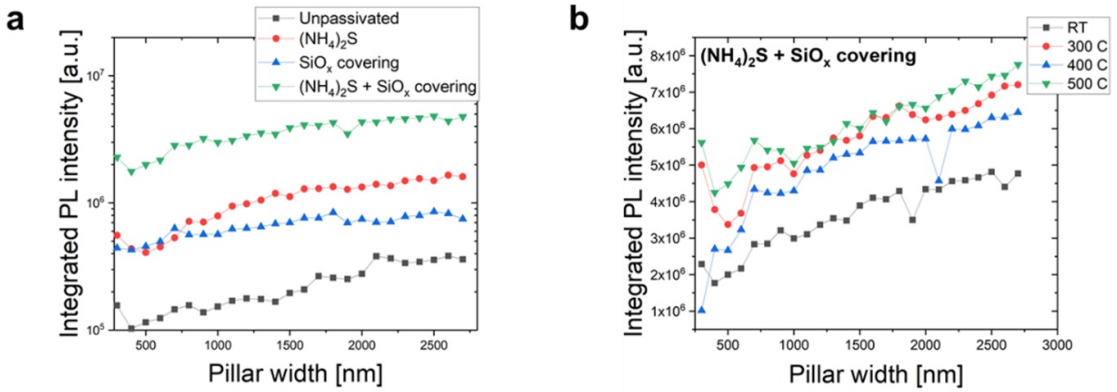

Figure 13. Photoluminescence intensity measurements for passivated InGaAs nanopillars. (a) Encapsulation effect. (b) Annealing effect (Reprinted with permission from [93] © The Optical Society.).

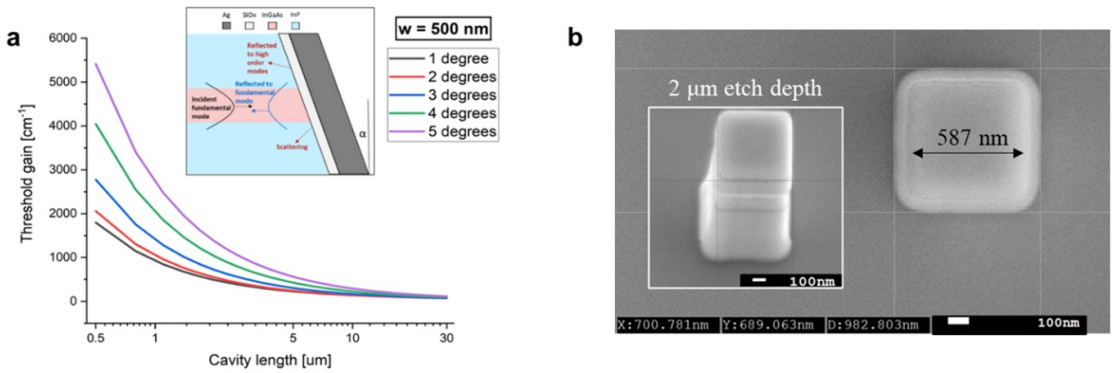

Figure 14. (a) Effect of etching slopes on the threshold gain of metal-cavity nanolasers for different sidewall angles. The inset shows a schematic of the sloped facets. (b) Scanning electron microscope (SEM) images of etched nanolaser pillars with high verticality slopes. The calculated etching slope angle is ∼1.6°.

heat sinking. Solving all the aforementioned aspects in RTD-based nanolasers is of central relevance to achieve the next-generation of high-performance nanolaser neuron spiking emitters.

5.3. Nanophotodetector RTD neuron receiver: nanoRTD-PD

The RTD-based nanophotodetector neuron receiver consists of either micro- or nanopillars GaAs/AlGaAs (for operation at 850 nm) or InGaAs/InP (operation at 1310 nm or 1550 nm) epilayer stacks with light absorbing layers surrounding the DBQW. The layout of the baseline nanoRTD-PD for detection at 1310 nm or 1550 nm is illustrated in figure 15. It consists of an AlAs/InGaAs/AlAs DBQW surrounded by a thick layer of a light absorbing material in the collector side of the RTD (typically the same material as the quantum well of the DBQW). The contact layers comprise of a higher band-gap material, such as InP. The device is considered forward biased, with current flowing from the top contact to the bottom contact.

The spike detection principle in the nanoRTD-PD, schematically illustrated in figure 15(c), is as follows. When d.c. biased near the peak of first PDC, and in the absence of light, the voltage and current flowing through the nanoRTD-PD remain constant. In the presence of an optical spike (triggering event), the electron–hole pairs generated by the absorption of the optical spike are separated by the built-in electric field induced across the collector by the d.c. applied voltage. The electrons drift towards the collector contact, thus adding to the circuit's external current, while the photo-generated holes moving towards the emitter contact are trapped by the DBQW collector barrier, accumulating in the the collector region. This leads to an enhancement of the electric field across the DBQW and the collector depleted region, allowing the resonance in the quantum well to be reached at a lower bias voltage than under dark conditions. As a result, the circuit load line 'jumps' momentarily to the NDC region of *I–V*, leading to the generation of a voltage spike corresponding to the valley-to-peak voltage difference (section 3.2). This can be interpreted as an overall shift to lower voltage of the nanoRTD-PD *I–V* curve under illumination, figure 15(b). This enables to

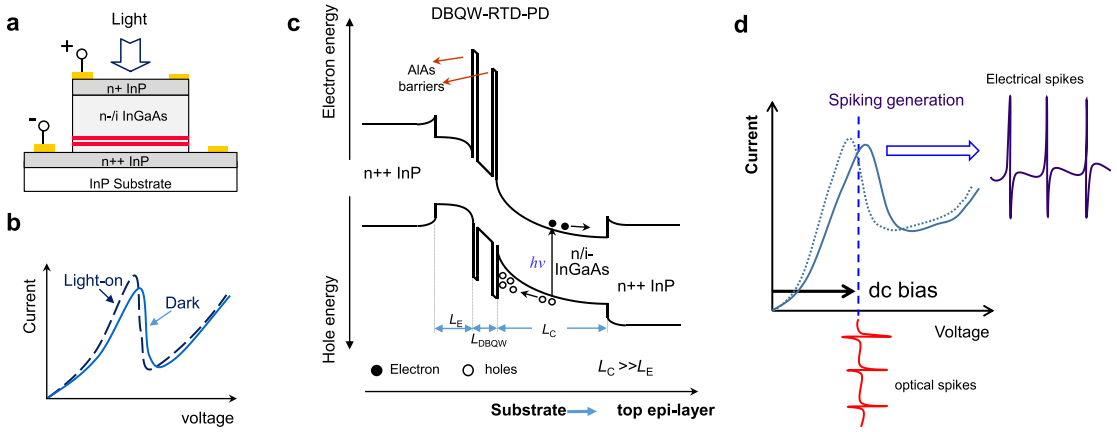

Figure 15. (a) Physical layout and layer stack of a microRTD-PD with a light absorbing region on the collector side; (b) Effect of illumination in the *I–V* characteristic. (c) Schematic representation of the RTD-PD conduction and valence bands' profile under bias, and of the principle of operation of the RTD-PD, showing the effect of light absorption: accumulation of holes in the collector region adjacent to the DBQW and electrons drifting towards the collector contact. (d) RTD-PD operation as an optical-electrical neuron-like spike transducer.

achieve optical-to-electrical spike conversion, figure 15(d). The bandwidth of the nanoRTD-PD excitable photodetector is limited by the pulse duration and dead time. These properties are mainly determined by the circuit's intrinsic PDCs, NDC, capacitance and inductance. Spiking nodes with electrical and optical inputs and outputs can be realized by combining nanoRTD-PD neuron receivers with neuron nanolight emitters (previous Sections).

Several loss mechanisms can influence the optical response of nanoRTD-PD devices, including a non-ideal photodetection quantum efficiency due to the limited absorption through the finite absorption layers, optical reflection of the incident light at semiconductor surface, losses due to fibre-chip coupling, losses in internal contact layers, and carrier recombination before the carriers reach the contacts. Moreover, due to bandwidth limits such as RC time, transit time, and carrier diffusion, the photocurrent generated at very high frequency (GHz band) may be smaller than the stationary state (low frequency operation).

5.4. Nanofabrication and spike characterization of nanoRTD-PD neuron receivers

Several epilayer stacks have been investigated for the fabrication of nanoRTD-PDs with the goal of developing excitable neuromorphic light-induced spike generators with low power consumption. The fabrication of nanometre-sized nanoRTD-PDs with suitable epitaxial layer structures, figure 16(a), to meet the expected low-power consumption requirements, as well as the realization of device architectures that provide efficient optical absorption, present challenges. A fabrication process for nano-sized InP RTDs has been successfully developed. Figure 16 shows in panels (b) and (c) scanning electron microscope (SEM) images of the various stages of the nanofabrication process for a representative nanoRTD-PD device consisting of a nanopillar with a diameter of 500 nm. In panel (d), the *I–V* characteristics of nanopillar devices with diameters of 500 nm, 700 nm and 800 nm are shown. Clear signatures of NDC regions are achieved. The peak currents of the fabricated devices are in the range of a few hundred μA allowing for an energy-efficient solution. Nanopillars with optical windows for efficient light in-coupling represent one of the ongoing challenges in operating nanoRTD-PDs as neuron optical receivers. In what follows, we discuss the observation of electrical spikes using nanopillar devices, as well as the observation of light-induced spikes using already established micropillar RTD-PD devices with light in-coupled from the top open mesa windows, as illustrated in figure 15(a).

Experiments to generate spikes with RTD-PDs featuring optical windows on top of the pillar mesa have been reported recently [72]. Figures 17(a) and (b) show the electrical and optical experimental setups, respectively, used to investigate the generation of spike signals in RTD-PDs. Firstly, for electrical spike triggering, figure 17(a), the devices were driven by an electrical pseudo-random bit sequence with an amplitude of 1 V and a frequency of 500 MHz, and biased at −0.59 V, i.e. near to the device's peak voltage. The nature of the spikes is similar to that seen in figure 17(c), which depicts the periodic generation of electrical spikes. Noteworthy, similar spiking phenomena can be activated by optical modulated signals, figures 17(b) and (d) [72]. The injected light plays the role of external perturbations that drive the RTD to generate stochastic spiking. As shown in figure 17(d), excitable phenomena (here stochastic spiking), can be elicited by carefully controlling either the bias voltage or the intensity of the optical perturbation signal. We

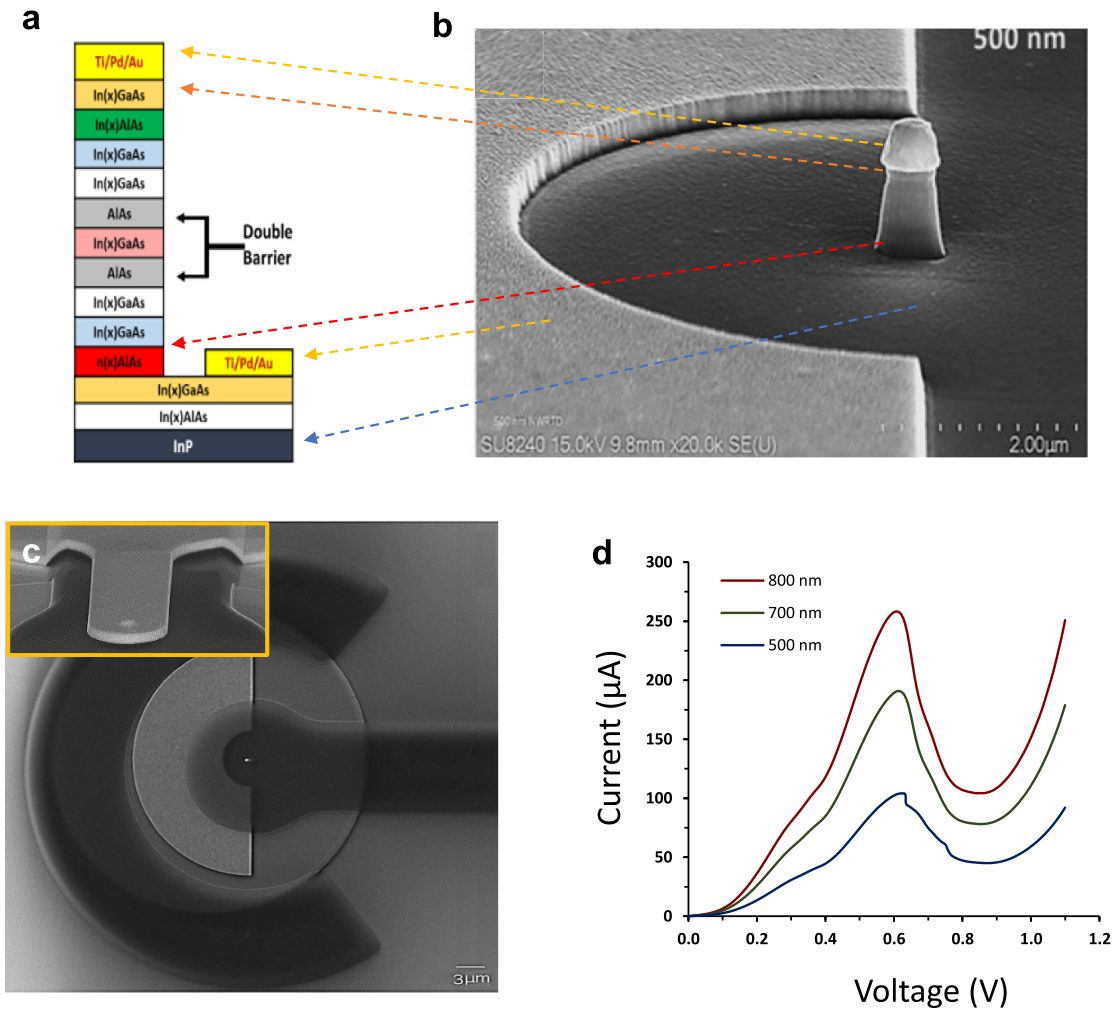

Figure 16. Fabrication of nanopillar RTD-PD devices. (a) Semiconductor epilayer stack structure. (b) Scanning electron microscope (SEM) image of a 500 nm diameter InP-based RTD-PD nanopillar device. The top metal contact is also shown. (Reprinted, with permission, from [94]. © [2022] IEEE.) (c) SEM image of a 500 nm nanopillar after the etch back process using BCB to open the top and bottom contacts. The top inset shows the bond top pad covering the nanopillar. (Reprinted, with permission, from [94]. © [2022] IEEE.) (d) I–V characteristics of InP-based RTD nanopillars with 500 nm, 700 nm and 800 nm pillar diameters showing clear regions of NDC. (Reprinted, with permission, from [94]. © [2022] IEEE.)

notice the spikes are fired using a rather low power input (∼90 μW). In this example, the pulse duration of each spike is about 12.5 μs, with a typical frequency between spikes of 12 kHz (i.e. the refractory response), which is dependent on the applied voltage and the intrinsic properties of the RTD-PD oscillator circuit.

We observe that by carefully designing the circuit (decreasing the capacitance and inductance values), the oscillation frequency of the spike firing may be tuned from kHz up to GHz frequencies. Using a micropillar RTD-PD, figure 18 illustrates an example of deterministic firing of high-speed electrical spikes (blue traces) triggered by two consecutive 1 ns optical pulses (red trace) separated by 5 ns (panel (a)) and 2.45 ns (panel (c)). The separation threshold (i.e. 2.45 ns) in this example can be regarded as the RTD-PD circuit refractory time.

These studies demonstrate the versatility of nanoRTD-PDs in performing multiple spike-based firing functions (here oscillatory, stochastic and deterministic spiking) by adjusting a few operating parameters (e.g. bias voltage, amplitude or frequency of the incoming signal, etc.). As a result, we anticipate nanoRTD-PDs to function in a variety of neuromorphic spiking dynamic behaviours that are entirely compatible with optical signal transmission systems and do not require any additional electronic or optoelectronic components.

5.5. Summary of nanophotonic spiking neurons

In the previous Sections, we discussed the monolithic integration of nanoRTDs with nanoLEDs, nanoLDs, and nanoPDs to realize neuron emitter and receiver spiking nodes. Table 2 presents a comparison between the nanophotonic neurons presented in this work in terms of speed, energy consumption, and type of neuron model emulated, with four representative analog photonic/optoelectronic solutions

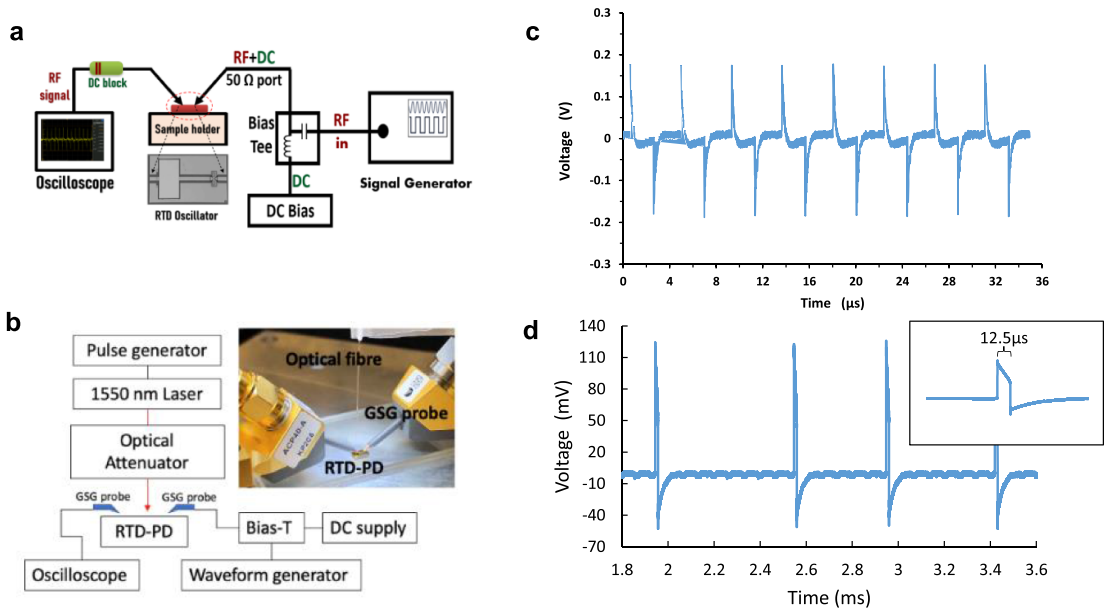

Figure 17. Experimental set-up for characterization of spike generation in micro- and nanopillar RTD-PDs using (a) electrical signal, and (b) optical signal injection. (c) Experimental spike outputs from a 500 nm diameternanopillar RTD device. (d) Stochastic spiking generation from a micropillar RTD-PD induced by light perturbation. The frequency of the light pulse train was set as 15 kHz, with a duty cycle of 30%. The average optical power was kept constant at 87.6 μW. The RTD-PD was biased at the 1st PDC region, very close to its NDC region. The inset indicates the width of the spike. The width of the spikes is 12.5 μs. Panels (a) and (c) © [2022] IEEE. Reprinted, with permission, from [94]. Panels (b) and (d) Reproduced from [72]. CC BY 4.0.

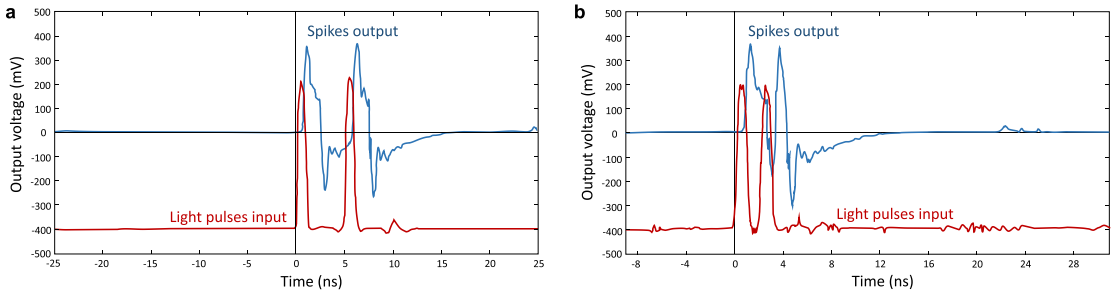

Figure 18. Example of deterministic firing of high-speed spikes from a micropillar RTD-PD device (active area of 15 μm × 15 μm). Spike output (blue trace) triggered by two consecutive 1 ns optical pulses (red trace) with identical amplitude and separated by (a) 5 ns and (b) 2.45 ns.

(superconducting Josephson junctions [23], phase change materials [51], excitable micropillar lasers [17], and Izhikevich-inspired photonic neuron [53]), and two analog electronic neuromorphic architectures (TrueNorth [4], and Loihi [5]). Although by no means exhaustive, the solutions chosen from the literature relate to representative artificial neurons capable of spike- (pulse) based signalling. Since not all of the solutions addressed in this perspective have been shown experimentally, the numbers in table 2 comprise the best projected values taking into account the designed performance based on physical models of the nanophotonic neurons. We conclude the nanophotonic spiking architecture discussed here has the advantage of offering a wide range of spatio-temporal neuron-like complex functions, which are difficult to realize using optics-based approaches, while providing energy consumption and spike event timescales of comparable performance to some of the best reported photonic and optoelectronic approaches. These neuron-like functions are similar to those that are many times only possible using complex circuit design in electronics-based neuron models, while providing higher speed potential (>GHz) and consuming much lower energy per event. Potential limitations of the proposed nanophotonic neuron architecture for scalable SNNs include the idle state voltage and current required to drive the nanophotonic neurons. Specifically, in the nanoLED and nanoLDs this limits the energy consumption to values ranging from 10 to 100 fJ per event fired, due to the diode turn-on voltage. However, alternative circuit design solutions, such as on-off pulse driving of the nanolight sources, could be implemented to minimize power consumption.

Table 2. Comparison of electronic, photonic and optoelectronic neuron architectures (including this work) capable of spike-based signalling in terms of speed, energy consumption, and type of neuron model emulated.

Nanophotonic neuron architecture (this work)	Energy/event (J)	Pulse/spike event timescales	Neuron-like functions
nanoRTD-LED [78]	$\sim 10^{-14}$	<1 ns	Spike firing (excitable model), stochastic spiking and bursting, self-oscillations, mixed-mode-oscillations, resonate-and-fire, bistability, autaptic (++)
nanoRTD-LD [95]	$\sim 10^{-13}$	~100 ps	
nanoRTD-PD	$\sim 10^{-15}$	<100 ps (~2 ns[a])	
Photonic and optoelectronic neuron architectures			
Superconducting Josephson junctions [23]	$> 10^{-14}$	300 ps–15 ns (turn-on and turn-off)	Integrate-and-fire model
Phase change materials [51]	$\sim 10^{-12}$	500 ps–1.5 ns (read and write)	Integrate-and-fire model
Micropillar laser [17]	$> 10^{-14}$	~200 ps	Spike firing (excitable model)
Izhikevich-inspired photonic neuron [53]	$\sim 10^{-13}$	~100 ps	Izhikevich model (aided by electronic transistors)
Electronic neuron architectures			
TrueNorth [4]	$> 10^{-12}$	$< 10^{-3}$ s	Leaky-integrate-and-fire model. Izhikevich model (when combining 1–3 neurons)
Loihi [5]	$> 10^{-12}$	$< 10^{-3}$ s	Leaky-integrate-and-fire model

[a] Shown experimentally in figure 18. ++ Discussed in section 3.

Note: electronic neuron architectures integrated in a very large scale SNN.

6. Photonic synaptic interconnects

Optical interconnects for neuromorphic photonic chips require interconnectivity in highly branched networks connecting the different computational elements of a neural network consisting of RTD-based nanolight sources and nanophotodetectors. Active weights are required in the training phase of a neural network, when the weights in the network are adjusted to minimize the classification error after the evaluation of each batch of training samples. Also, in a neural network system that is designed to execute pre-trained neural networks, active weights make it possible to reconfigure the system and start processing a completely different neural network. In this Section, we discuss potential interconnects for on-chip and inter-chip connections comprising of actively weighted optical synaptic interconnects based on silicon photonics and photorefractive materials (section 6.1), and polymer-based materials for 3D optical waveguide connections (section 6.2).

6.1. 2D synaptic interconnects

The hardware of ANNs can be used for different tasks or modes, which place several requirements on the optical synaptic interconnections between the neuron nodes. If the optical neural network hardware is used to train a new network, the synaptic interconnect weights have to be adjusted after each (small) batch of training samples. In this case, training time and power consumption can be reduced significantly by eliminating time and energy-consuming data transfers between the synaptic interconnect and external weight memory and processing units. This can be achieved by storing and updating the synaptic connection weights locally in the synaptic interconnect [96]. If the optical neural network is used in inference mode and thus performs a pre-trained classification task on a stream of input data, the synaptic weights are updated seldomly and external control and storage of the connection weights does not incur a large power penalty. In this case the holding power that is required to keep the synaptic connection weights at a fixed level dominates the energy consumption [8].

In this Section, we discuss two paths toward a tunable 2D synaptic interconnect. The first approach consists of an electrically tunable synaptic interconnect based on the well-established silicon photonics

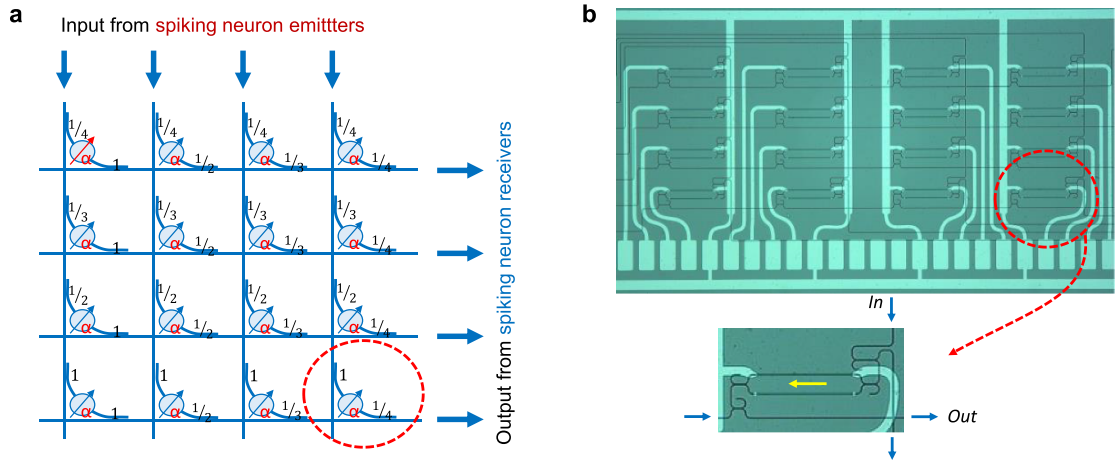

Figure 19. Silicon photonics-based interconnect. (a) Schematic layout. (b) Die photograph of the fabricated interconnect.

technology. In a second approach, we present on an exploratory concept, where the optical synaptic interconnections are obtained by diffraction on refractive index gratings written in a photorefractive crystal [97, 98]. This approach has been demonstrated in optical systems built from discrete elements [99], but we intend to integrate the photorefractive material and all controlling optics onto a single silicon photonics chip. With the latter concept, the synaptic weights are stored locally in the interconnect and the full weight matrix can be adjusted in one single step, which makes this interconnect well suited for application in a neural network training system.

6.1.1. Silicon photonics interconnects

The silicon photonics interconnect architecture is based on a mesh of crossing input and output waveguides, with cross-coupling elements incorporating a variable attenuator at each crossing. The schematic layout of this interconnect mesh, together with a die photograph of the fabricated interconnects are shown in figures 19(a) and (b), respectively. Assuming that the light from the source neurons is completely incoherent we can split and combine the light in the optical power domain only. Thus, the crossing points of the mesh are designed such that at each crossing the same amount of the input light power is coupled out from the input waveguide, attenuated according to the required synaptic interconnect strength, and coupled to the output waveguide. The corresponding power coupling factors are denoted at each cross-point element in the schematic layout, as shown in figure 19(a).

The physical layout of the synaptic cross-point elements consists of a directional coupler that taps light out of the vertical input waveguide, followed by a thermally tuneable Mach–Zehnder interferometer that serves as a variable optical attenuator. The attenuated light is then coupled into the horizontal output waveguide by a second directional coupler. A photograph of a representative fabricated silicon photonics interconnects is shown in figure 19(b). The current devices could be implemented in an envisioned SNN using off-chip coupling (e.g. via optical fibres) from sources-emitters to the synaptic interconnects. The designs presented here can be adapted to include optical interfaces matching the spiking optical elements for a complete on-chip hybrid co-integration.

6.1.2. Photorefractive optical interconnects

In the photorefractive concept, the synaptic connections are stored as refractive index gratings in a photorefractive layer on an integrated photonic circuit. The refractive gratings are written and read by sets of collimated optical beams. The collimated beams are obtained by placing lenses between the photorefractive medium and arrays of transmitters and detectors. These basic operations are illustrated in figure 20.

The physical implementation of the photorefractive interconnect in integrated optics is illustrated in figure 21. The photorefractive layer in the current version of the device is produced by bonding a semi-insulating GaAs wafer to an oxide cladding layer on a silicon wafer and thinning the GaAs layer down to $\sim$1 μm by grinding and chemical-mechanical polishing. Through two-wave mixing measurements performed perpendicularly through the photorefractive interconnect die, as shown in figure 21(c), we could confirm the photorefractive effect in the interaction region of the interconnect devices. Our current focus is on the experimental validation of the full interconnect functionality of the devices towards realizing photorefractive crossbar arrays for neural network inference and training.

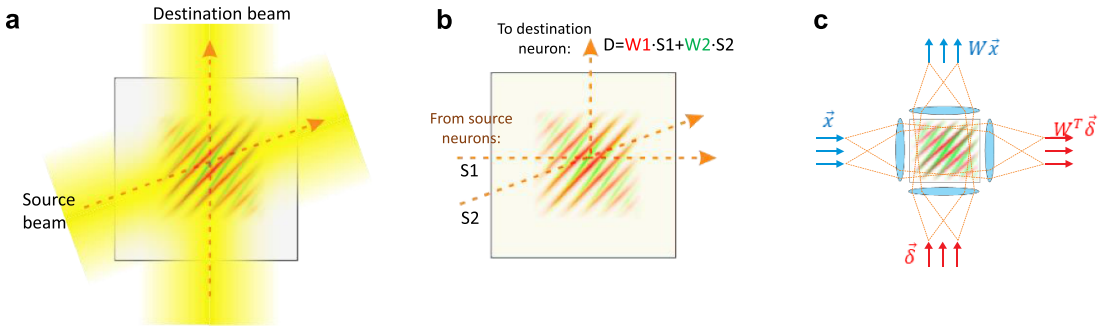

Figure 20. Basic operations for writing and read-out of synaptic connections in a photorefractive crystal. (a) Writing of a refractive index grating by the interference pattern of two coherent collimated beams. (b) Optical Multiply-Add operation on two superimposed refractive index gratings. (c) Conversion between diverging/converging light beams from arrays of transmitters and detectors and sets of collimated beams by collimating lenses. Forward and backward readout of the synaptic matrix as required for backpropagation training of an artificial neural network.

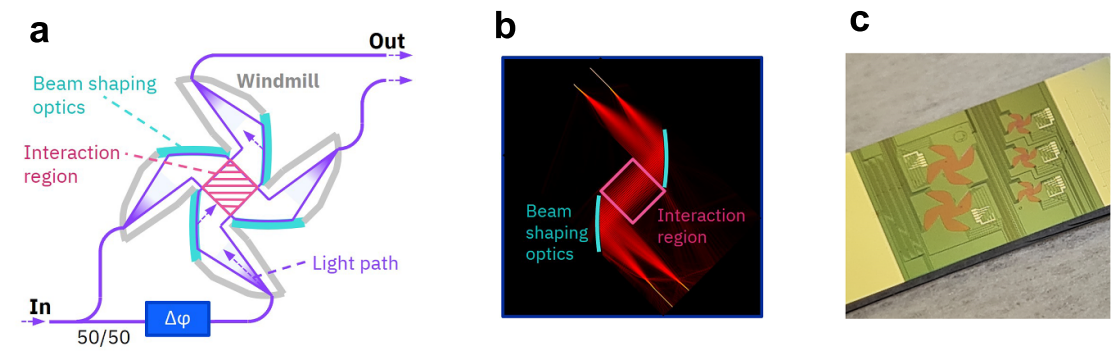

Figure 21. Physical implementation of the integrated photorefractive interconnect. (a) Mask layout showing the input and output waveguide arrays, collimating mirrors and photorefractive interaction region. (b) Simulation of light propagation through the interconnect. (c) Photograph of a completed die, containing five interconnects of different sizes (Reprinted with permission from [100] © The Optical Society.).

6.2. 3D suspended waveguide interconnections

Chip-to-chip connections are usually performed using optical fibres and grating couplers, which can incur in up to 10 dB losses due to large core size and mode mismatches between the fibre and the on-chip components. This is a severe limiting factor for neural network scaling, as the cumulative loss in complex interconnected systems quickly becomes prohibitively large. The microprinting of 3D polymeric structures using TPP offers potential to tackle this, with recent works reporting fibre-to-chip-couplers [101], optical fibre taper couplers [102, 103], and free-form inter-chip couplers [104, 105].

Noteworthy, 3D polymer structures have been proposed recently for neural networks in reservoir computation [34]. In fact, there is a vast potential in recently-developed 3D interconnection design, simulation, and manufacturing methods for novel inter-chip optical wire-bonding approaches. Such approaches resort in TPP microprinting methods to achieve freeform 3D structures with optically clear resins [106]. The flexibility of such polymeric waveguides holds excellent prospects for connecting various nanolight sources and detectors in comparatively small area footprints, as the optical connections can cross each other in 3D. Examples of recent 3D-printed polymeric photonic elements include beam-shaping structures [107], sub-micron waveguides [108], and waveguide splitter-based convolution kernels for DNNs [34], among other waveguide devices [109, 110]. On the other hand, free-form on-chip 3D waveguides and couplers for individual device interconnections still remain relatively unexplored. Recent works have proposed designs based on continuously [111] and discretely [109] supported 3D waveguides for planar photonic circuits and 3D splitter waveguides [112].

In this Section, we discuss the design and fabrication of 3D structures (waveguides and couplers, figures 22(a)–(c)) to exploit the interconnectivity of multiple photonic neuromorphic elements in small footprint areas (see concept in figure 22(d)). Differently from the synaptic interconnects discussed in the previous Section, here there is no functionality for active weighting in 3D polymers. Future developments could include active functional materials for either thermal, magnetic, or other ways of controlling light

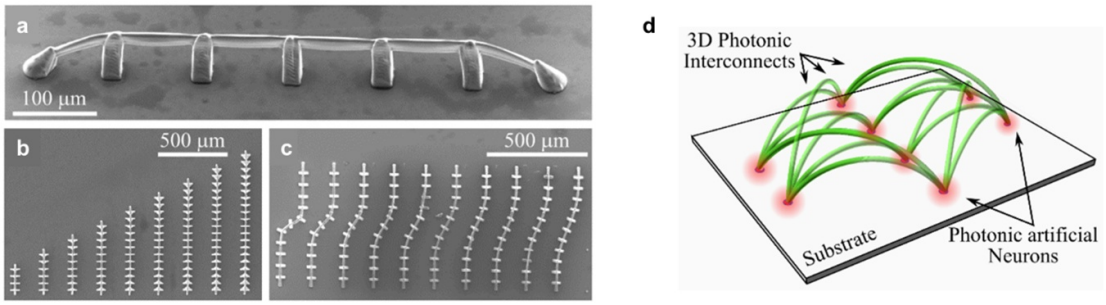

Figure 22. 3D photonic interconnects. (a)–(c) Scanning electron microscope (SEM) images of 3D waveguides fabricated using two photon polymerization (TPP). (a) Straight waveguide with tapering. (b), (c) Waveguides for systematic optical transmission characterization of straight (b) and s-bend (c) designs with variable length (number of support structures) and bending radius, respectively (panels (a)–(c) Reprinted with permission from [115] © The Optical Society.). (d) Concept of photonic artificial neural nodes interconnected by polymeric 3D photonic waveguides.

transmission properties. Alternatively, the passive 3D polymeric waveguides could be integrated with recently-developed active transmission weighting technologies (section 6.1).

For the experimental realization of 3D waveguides, we focus on the TPP technique [113, 114], which uses a femtosecond laser to polymerize a photosensitive material in a two-photon absorption process. A suitable photosensitive polymer should meet several requirements for optical interconnects: (i) optical transparency in the desired wavelength range, (ii) suitable refractive index contrast for waveguiding, and (iii) displaying suitable post-exposure stiffness (Young's modulus) for mechanically stable 3D architectures (in some scenarios bridging over several hundred micrometers between functional elements is required). Some polymers contain photoinitiators that can induce avalanche effects that increase the TPP cross-section (reducing the fluence threshold required for TPP). However, this approach also results in larger minimum effective writing voxels and the size of the recoverable features. Lastly, materials with low shrinkage and suitable viscosity are also relevant. For example, low viscosity for simple spin coating or higher viscosity to achieve better bonding to the substrate in inverted TPP fabrication systems, can facilitate the fabrication process. Such properties could improve the quality of the 3D microstructures, leading to higher-fidelity 3D CAD design reproductions.

Recently, we have developed a dedicated algorithm to predict the 3D morphology of TPP writing structures, taking into account the experimental setup and the material parameters [115]. This tool enables to predict crucial experimental and design parameters (such as laser fluence and line density) for optimal structural reproducibility. It should be noted that the limited stiffness of typical TPP optical-grade materials for the neuromorphic chip dimensions (spanning several hundred microns) can require the use of support structures in suspended waveguides (figure 22(a)). Such structures introduce losses in extended 3D waveguides, characterized by the systematic fabrication of straight waveguides with variable height and number of support structures (figures 22(b) and (c)). The support structures introduce scattering hotspots, leading to propagation losses far above the typical insertion losses in planar waveguides employed in semiconductor materials. Thus, mechanical stability (through stiffer polymers and shortened designs) is central for reducing the number of support structures and transmission losses. The fabrication results shown here revealed stable structures with at least 100 μm gaps between support structures.

A systematic optical characterization of the bending losses in fabricated straight (figure 22(b)) and s-bend (figure 22(c)) structures was performed (figures 23(a) and (b)). The results revealed a minimum bending radius of about 100 μm is required to minimize the losses. The 3D CAD designs used for fabrication were employed for modelling of the waveguide optical properties (using either mode-solving, finite-difference time-domain (FDTD), or custom oscillator-based O-FDTD simulations [116, 117]). As illustrated in figure 23(b), the mode-solving simulations are in good agreement with experimental results. Such simulator engines provide the description of the mode profiles in the waveguides, enabling for example the optimization of the 3D geometries of optical tapers to achieve a high coupling efficiency between waveguides and the active neuron emitters/receivers [115], which will be crucial in future interconnected nodes. Additional relevant design considerations shall include the design of tapers optimized for the emission profile of nanolight sources, and the implementation of 3D interconnects linking on-chip nanophotonic components (such as a neural network of nanoRTD-LD/LEDs and nanoRTD-PDs). In an advanced stage of development, the design of 3D manifolds that act as network building blocks and form branched neural networks, and the conceptual design of chip-to-chip interconnects shall be considered for complex neural networks.

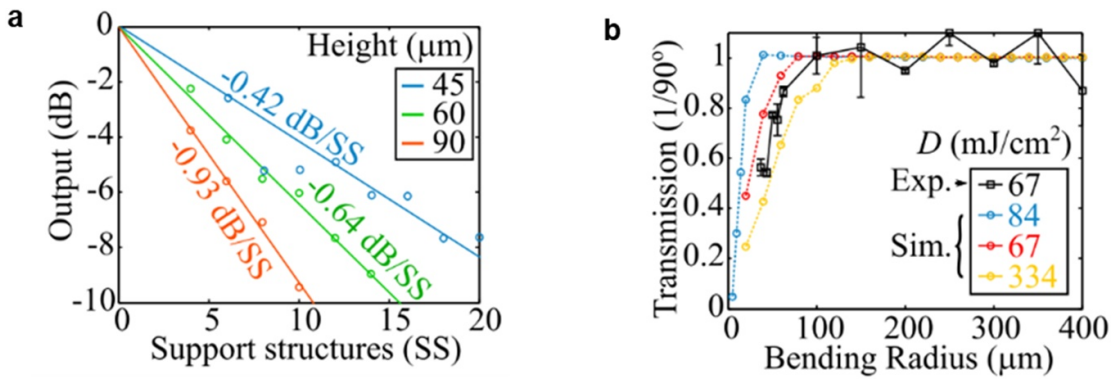

Figure 23. Characterization of the optical transmission losses for (a) straight and (b) s-bend 3D waveguides elevated by support structures (Reprinted with permission from [115] © The Optical Society.).

7. Spatio-temporal spike-based information processing

By drawing inspiration from the brain, the propagation and processing of information via spikes is one of the key differentiating aspects of neuromorphic platforms when compared to other methods for hardware-based AI acceleration. SNNs, as a third generation of ANNs, hold significant promise for their further development, particularly due to their prospects for lower energy requirements arising from the intrinsic data representation sparsity in SNNs as well as their use of the temporal domain for computation and data representation. Simultaneously, SNNs are not well suited for implementation on traditional CPU or GPU architectures as the mismatch between the principles of synchronous digital computation and time-based spiking processing leads to high power overhead and reduced speeds. Meanwhile, as a new technology competing with already well established deep learning approaches, SNNs face some ongoing challenges. These include the development of new efficient learning algorithms (as SNNs typically cannot be trained via conventional backpropagation algorithms), and the handling of data represented solely via temporal spikes. Novel approaches to learning in SNNs are under investigation in both scientific research and industrial settings. These include algorithms such as equilibrium propagation [118], modified backpropagation [119], supervised learning methods such as SuperSpike [120] or ReSuMe [121], and even approaches that directly account for slight hardware variations of the active components on neuromorphic chips [122].

Since the RTD-based optoelectronic devices explored in this article naturally exhibit excitable spiking responses as a result of their highly nonlinear characteristics (see section 3), they are prime candidates for light-enabled hardware realizations of spiking neurons and, with the implementation of additional scaling-up strategies, full photonic SNNs. In a recent study [95], we have numerically investigated the spike-based information processing capabilities using both single and multiple interconnected (in feed-forward architecture) nanoRTD-based optoelectronic spiking nodes. In the single node case, we considered a single nanoRTD-PD capable of performing spike-based information processing tasks, as it will be discussed in section 7.1. In the feed-forward networked architecture, discussed in section 7.2, we consider the following interconnected system: (i) master nodes, realizing the electrical-to-optical (E/O) link by means of a nanoRTD element coupled to a nanolaser (see emitter neurons presented in section 5), and (ii) a receiver node, realizing the optical-to-electrical (O/E) functionality via a nanoRTD with an incorporated photodetecting input (see receiver neurons presented in section 5).

7.1. Spike-based information processing methods using single neuron nodes

For a system to successfully operate as an artificial spiking neuron in scope of a larger SNN, it should exhibit certain key dynamical functionalities. These include the ability to sum and threshold input signals both spatially (realizing fan-in) and temporally; thus providing the ability to respond to incoming pulses (stimuli) with different timing while exhibiting a (short term) memory. Lastly, the artificial neuron should be capable of providing a fan-out, sending its output to multiple downstream nodes in a network. Recently, we have numerically demonstrated all of these key functionalities in nanoRTD-powered spiking optoelectronic nodes [95]. Firstly, we have shown that RTDs, when biased either in the valley or peak regions (in their highly nonlinear I–V characteristic), are capable of deterministic firing spikes in response to either electrical or optical (in the case of nanoRTD-PD) incoming signals. The fired spikes exhibit all the hallmarks of typical all-or-nothing excitable responses, that is, incoming subthreshold perturbations produce negligible system outputs, whilst superthreshold perturbations elicit firing of spikes that are independent of the amplitude of

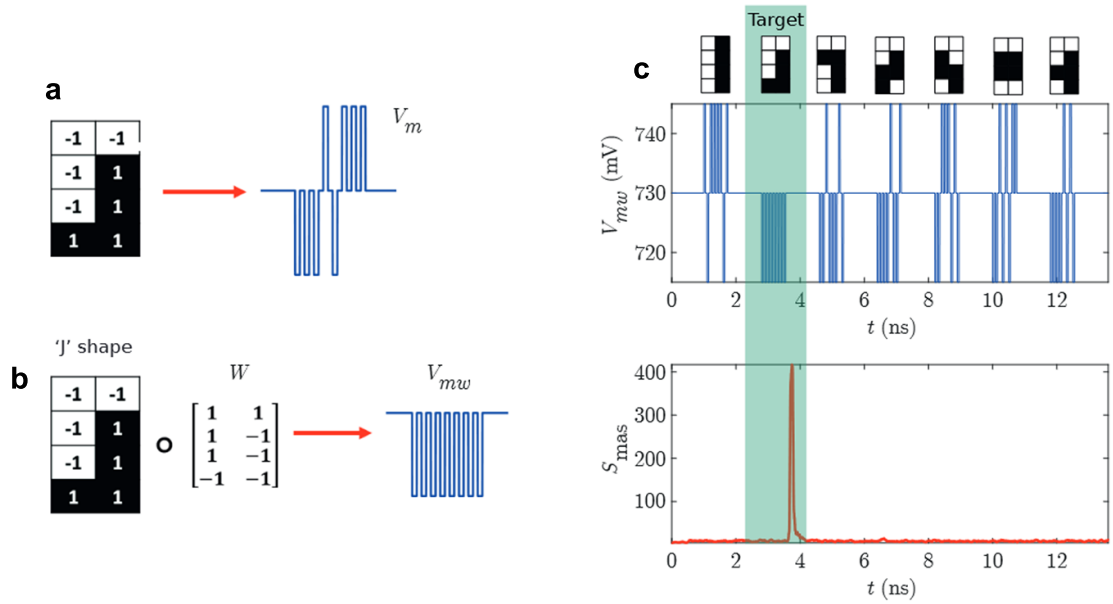

Figure 24. Information processing task performed with a single nanoRTD node. (a) A 4 × 2 binary pattern is serialized into a waveform of return-to-zero bits at 100 ps bit^{-1} with 50% duty cycle. (b) By pre-weighting the pattern with a weight matrix, the values represented in the waveform are remapped. Using the temporal integration of input perturbations in the nanoRTD node, a pattern recognition task (c) is performed where only a target results in the firing of a spike at the RTD node's output (Reprinted (figure) with permission from [95], Copyright (2022) by the American Physical Society.).

the incoming perturbations. Furthermore, the photodetection property in the nanoRTD nodes enables directly spatial integration of upstream signals, while the RTD itself exhibits the functionality of temporal signal integration, drawing a general parallel to the well-known leaky-integrate-and-fire neuronal model.

We have utilized this capability of the nanoRTD node to demonstrate spike-based information processing tasks at ultrafast rates and with low-input energy signals. We have shown the ability to perform a pattern recognition task, where an 8-bit sequence corresponding to a serialized matrix of 4 × 2 binary values (representing Tetris-like blocks) is used as an input to the nanoRTD node, figure 24(a). A weighting process (figure 24(b)) is applied as follows: first the 8-bit pulse sequences are multiplied elementwise by a vector of weights W bounded between (−1,+1), and this weighted sequence is injected into the RTD as a series of return-to-zero pulses of 50 ps separated by 50 ps interleaves (50% duty cycle). In such a configuration, only the pre-weighted sequences with a sufficient number (in this case, eight) of downward pulses carry enough combined perturbation energy to go over the excitability threshold, eliciting a spiking response in the system (figure 24(c)). Therefore, the RTD is capable of performing a classification task on the pre-weighted data, only firing a spike for a specifically targeted data sequence.

7.2. Photonic SNNs

In the previous Section, we have shown that an nanoRTD node can not only operate as an artificial spiking neuron, but also that a single node can already perform information processing tasks, such as the classification of binary sequences. However, for more complex data processing, and for the realization of advanced functional tasks, it is essential to go beyond single neuron systems and implement network architectures built from multiple interconnected nodes.

To investigate this network mode of operation, we have analysed numerically a 5-to-1 feedforward SNN built with nanoRTDs (figure 25(a)). This is formed by five master (E/O) nodes in the input layer and a single receiver (O/E) node on the output. All the upstream (PRE) nodes are connected to the output (POST) node via unidirectional links (e.g. waveguides) and considering a simple weighting function, in this case intensity/power attenuation. The PRE nodes are encoded with five binary data streams, eliciting input spatio-temporal spiking patterns that are processed (integrated) in the POST downstream node. The latter fires a spike only when a particular (target) spatio-temporal input pattern is detected, remaining quiescent otherwise. In this network, the temporal separation between each 5-bit input pattern is set to 420 ps, corresponding to full network processing capacity of 11.9 Gbps, with prospects for faster processing speeds, via increased parallelization (higher number of parallel nodes) or faster input sequences.

Similarly to common ANNs, the system needs to go first through a learning phase to be able to perform inference (recognition) on input data. As SNNs are not readily suitable for conventional error

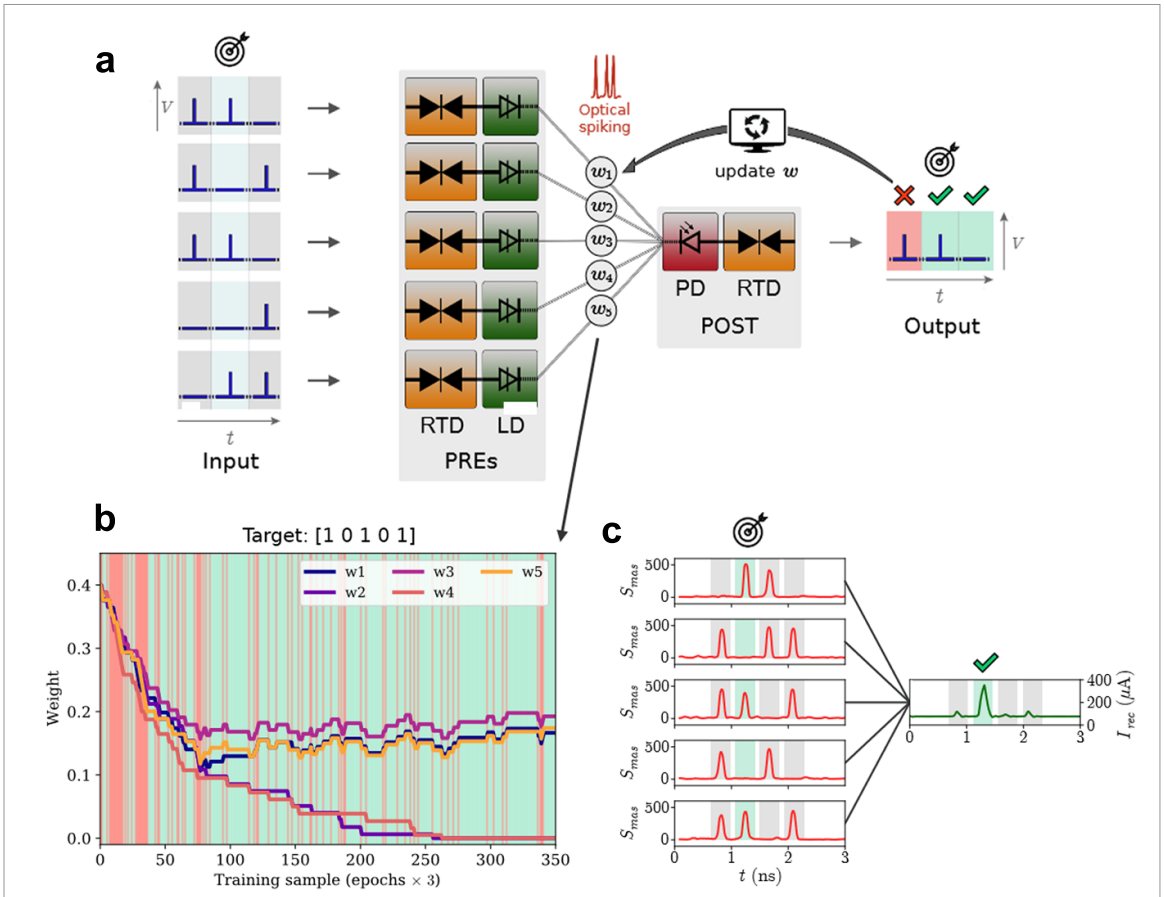

Figure 25. Information processing tasks performed with an RTD-powered network. (a) A five-to-one feedforward SNN is used for the classification of spike-based spatio-temporal binary patterns. The system can be trained with an off-chip, supervised, spike-timing based learning rule (b) and after training, it is capable of performing spike-based spatio-temporal pattern recognition (c). (Reprinted (figure) with permission from [95], Copyright (2022) by the American Physical Society.)

backpropagation training algorithms, an alternative solution for network weight adjustments is required. In [95] a supervised, off-chip, spike-timing dependent network learning rule is introduced, and demonstrated on the nanoRTD spiking network model (figure 25(b)). The rule utilizes labels on input data and temporal separation between pre- and postsynaptic spiking activity of the downstream node to train the network. The aforementioned 5-to-1 network was trained using this approach, reaching 94%+ accuracy during the subsequent inference phase on the spatio-temporal patterns (figure 25(c)). The simulation studies indicate the demonstration of a functional feedforward ANN utilizing spikes to compute and that can be realized in hardware with neural nodes based on optoelectronic nanoRTDs.

8. Conclusions and future perspectives

Today's most successful AI algorithms are based on ANNs. However, unlike our highly efficient brains, running these algorithms on conventional computers based on CMOS electronics consumes very large amounts of energy. In brain-inspired nanophotonic spike computing, light is used to perform computations in a manner analogous to our understanding of the brain. The key aspect is the use of light pulses, or spikes, to encode information. These spikes are then processed by an optoelectronic device, which performs the necessary computations. This approach is inspired by the way the brain processes information, as spikes are used to encode and transmit information between neurons. This could enable to build novel photonics-based SNNs. Here, we covered various initiatives reported recently to realize all-optical and optoelectronic SNNs. We note however active optical spiking components integrated on-chip in a viable scalable SNN architecture remain at an early stage of development.

In this perspective paper we highlight the potential of subwavelength photonics nanotechnology to deliver the compact, high-bandwidth and energy-efficient central processing spiking nodes (i.e. nanophotonic spiking neurons) needed to implement a future brain-inspired nanophotonic computing architecture. Specifically, we discussed the use of semiconductor RTD nanostructures embedded in subwavelength nanoLEDs, nanolasers, and nanophotodetectors, 100 times smaller than conventional

optoelectronic light sources and photodetectors, to efficiently confine, emit, detect and process neuron-like electrical and optical spikes. The advantage of the proposed nanophotonic spiking architecture is that it may perform a variety of complex spatio-temporal neuron-like functionalities in both the optical and electrical domains such as spike firing (excitable model), stochastic spiking and bursting, self-oscillations, mixed-mode-oscillations, resonate-and-fire, bistability, autaptic, which are challenging to achieve in other platforms, namely optics-based approaches. These neuron-like functions are similar to those that are many times only possible using complex circuit design in electronics-based neuron models. Additionally, nanophotonic spiking neurons can deliver performance-equivalent energy consumption and spike event timescales to some of the best reported photonic and optoelectronic neuron-like systems. In perspective, the development of such an energy-efficient neuro-architecture integrated with 2D and 3D synaptic optical interconnections, and configured with appropriate spike-based algorithms will pave the way towards the emerging field of extremely efficient neuromorphic optical computing with the potential to revolutionize many different fields, from medicine to finance, robotics and AI.

Data availability statement

The data cannot be made publicly available upon publication because they are not available in a format that is sufficiently accessible or reusable by other researchers. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

European Union, H2020-FET-OPEN project 'ChipAI' (Grant 828841). European Union, Horizon Europe project 'InsectNeuroNano' (Grant 101046790). UK Research and Innovation (UKRI) Turing AI Acceleration Fellowships Programme (EP/V025198/1).

ORCID iDs

Bruno Romeira https://orcid.org/0000-0002-1485-6665
Ricardo Adão https://orcid.org/0000-0002-9864-3922
Jana B Nieder https://orcid.org/0000-0002-4973-1889
Qusay Al-Taai https://orcid.org/0000-0001-5741-0527
Weikang Zhang https://orcid.org/0000-0001-7797-8734
Robert H Hadfield https://orcid.org/0000-0002-8084-4187
Edward Wasige https://orcid.org/0000-0001-5014-342X
Matěj Hejda https://orcid.org/0000-0003-4493-9426
Antonio Hurtado https://orcid.org/0000-0002-4448-9034
Victor Dolores Calzadilla https://orcid.org/0000-0001-9774-1574
D Castro Alves https://orcid.org/0000-0001-7026-2514
José M L Figueiredo https://orcid.org/0000-0001-5668-7073
Ignacio Ortega-Piwonka https://orcid.org/0000-0002-7175-0285
Julien Javaloyes https://orcid.org/0000-0001-9131-4483
J Iwan Davies https://orcid.org/0000-0003-1068-6344
Bert J Offrein https://orcid.org/0000-0001-6082-0068

References

[1] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436
[2] Xu X, Ding Y, Hu S X, Niemier M, Cong J, Hu Y and Shi Y 2018 Scaling for edge inference of deep neural networks *Nat. Electron.* **1** 216–22
[3] Wu Y, Deng L, Li G, Zhu J and Shi L 2018 Spatio-temporal backpropagation for training high-performance spiking neural networks *Front. Neurosci.* **12** 331
[4] Merolla P A *et al* 2014 A million spiking-neuron integrated circuit with a scalable communication network and interface *Science* **345** 668–73
[5] Davies M *et al* 2018 Loihi: a neuromorphic manycore processor with on-chip learning *IEEE Micro* **38** 82–99
[6] Sung C, Hwang H and Yoo I K 2018 Perspective: a review on memristive hardware for neuromorphic computation *J. Appl. Phys.* **124** 151903
[7] Shastri B J, Tait A N, Ferreira de Lima T, Pernice W H P, Bhaskaran H, Wright C D and Prucnal P R 2021 Photonics for artificial intelligence and neuromorphic computing *Nat. Photon.* **15** 102–14
[8] Feldmann J *et al* 2021 Parallel convolutional processing using an integrated photonic tensor core *Nature* **589** 52–58
[9] Miller D A B 2017 Attojoule optoelectronics for low-energy information processing and communications *J. Lightwave Technol.* **35** 346–96

[10] Shen Y et al 2017 Deep learning with coherent nanophotonic circuits Nat. Photon. 11 441–6
[11] Appeltant L, Soriano M C, Van der Sande G, Danckaert J, Massar S, Dambre J, Schrauwen B, Mirasso C R and Fischer I 2011 Information processing using a single dynamical node as complex system Nat. Commun. 2 468
[12] Izhikevich E M 2010 Dynamical Systems in Neuroscience the Geometry of Excitability and Bursting (Cambridge: MIT Press)
[13] Wu Y et al 2022 Brain-inspired global-local learning incorporated with neuromorphic computing Nat. Commun. 13 65
[14] El Srouji L, Krishnan A, Ravichandran R, Lee Y, On M, Xiao X and Ben Yoo S J 2022 Photonic and optoelectronic neuromorphic computing APL Photonics 7 51101
[15] Peng H-T, Nahmias M A, de Lima T F, Tait A N and Shastri B J 2018 Neuromorphic photonic integrated circuits IEEE J. Sel. Top. Quantum Electron. 24 1–15
[16] Barbay S, Kuszelewicz R and Yacomotti A M 2011 Excitability in a semiconductor laser with saturable absorber Opt. Lett. 36 4476–8
[17] Selmi F, Braive R, Beaudoin G, Sagnes I, Kuszelewicz R and Barbay S 2014 Relative refractory period in an excitable semiconductor laser Phys. Rev. Lett. 112 183902
[18] Rasmussen T S, Yu Y and Mork J 2020 All-optical non-linear activation function for neuromorphic photonic computing using semiconductor Fano lasers Opt. Lett. 45 3844–7
[19] Cheng Z, Ríos C, Pernice W H P, Wright C D and Bhaskaran H 2017 On-chip photonic synapse Sci. Adv. 3 e1700160
[20] Tait A N, de Lima T, Nahmias M A, Miller H B, Peng H-T, Shastri B J and Prucnal P R 2019 Silicon photonic modulator neuron Phys. Rev. Appl. 11 64043
[21] Romeira B, Javaloyes J, Ironside C N, Figueiredo J M L, Balle S and Piro O 2013 Excitability and optical pulse generation in semiconductor lasers driven by resonant tunneling diode photo-detectors Opt. Express 21 20931
[22] Seo S et al 2018 Artificial optic-neural synapse for colored and color-mixed pattern recognition Nat. Commun. 9 5106
[23] McCaughan A N, Verma V B, Buckley S M, Allmaras J P, Kozorezov A G, Tait A N, Nam S W and Shainline J M 2019 A superconducting thermal switch with ultrahigh impedance for interfacing superconductors to semiconductors Nat. Electron. 2 451–6
[24] Hill M T and Gather M C 2014 Advances in small lasers Nat. Photon. 8 908
[25] Leuthold J et al 2013 Plasmonic communications: light on a wire Opt. Photonics News 24 28–35
[26] Shambat G, Ellis B, Majumdar A, Petykiewicz J, Mayer M A, Sarmiento T, Harris J, Haller E E and Vučković J 2011 Ultrafast direct modulation of a single-mode photonic crystal nanocavity light-emitting diode Nat. Commun. 2 539
[27] Dolores-Calzadilla V, Romeira B, Pagliano F, Birindelli S, Higuera-Rodriguez A, van Veldhoven P J, Smit M K, Fiore A and Heiss D 2017 Waveguide-coupled nanopillar metal-cavity light-emitting diodes on silicon Nat. Commun. 8 14323
[28] Ellis B, Mayer M A, Shambat G, Sarmiento T, Harris J, Haller E E and Vučković J 2011 Ultralow-threshold electrically pumped quantum-dot photonic-crystal nanocavity laser Nat. Photon. 5 297
[29] Huang K C Y, Seo M-K, Sarmiento T, Huo Y, Harris J S and Brongersma M L 2014 Electrically driven subwavelength optical nanocircuits Nat. Photon. 8 244
[30] Crosnier G, Sanchez D, Bouchoule S, Monnier P, Beaudoin G, Sagnes I, Raj R and Raineri F 2017 Hybrid indium phosphide-on-silicon nanolaser diode Nat. Photon. 11 297
[31] Eltes F et al 2020 An integrated optical modulator operating at cryogenic temperatures Nat. Mater. 19 1164–8
[32] Huang C, Bilodeau S, Ferreira de Lima T, Tait A N, Ma P Y, Blow E C, Jha A, Peng H-T, Shastri B J and Prucnal P R 2020 Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits APL Photonics 5 40803
[33] Dietrich P-I et al 2018 In situ 3D nanoprinting of free-form coupling elements for hybrid photonic integration Nat. Photon. 12 241–7
[34] Moughames J, Porte X, Thiel M, Ulliac G, Larger L, Jacquot M, Kadic M and Brunner D 2020 Three-dimensional waveguide interconnects for scalable integration of photonic neural networks Optica 7 640–6
[35] Ironside C, Romeira B and Figueiredo J M L 2019 Resonant Tunneling Diode Photonics (San Rafael, CA: Morgan & Claypool Publishers)
[36] Maekawa T, Kanaya H, Suzuki S and Asada M 2016 Oscillation up to 1.92 THz in resonant tunneling diode by reduced conduction loss Appl. Phys. Express 9 24101
[37] Blakesley J C, See P, Shields A J, Kardynal B E, Atkinson P, Farrer I and Ritchie D A 2005 Efficient single photon detection by quantum dot resonant tunneling diodes Phys. Rev. Lett. 94 67401
[38] Weng Q, An Z, Zhang B, Chen P, Chen X, Zhu Z and Lu W 2015 Quantum dot single-photon switches of resonant tunneling current for discriminating-photon-number detection Sci. Rep. 5 9389
[39] Romeira B, Avó R, Figueiredo J M L, Barland S and Javaloyes J 2016 Regenerative memory in time-delayed neuromorphic photonic resonators Sci. Rep. 6 19510
[40] ChipAI Energy-efficient and high-bandwidth neuromorphic nanophotonic Chips for Artificial Intelligence systems (available at: www.chipai.eu/)
[41] Nandakumar S R, Kulkarni S R, Babu A V and Rajendran B 2018 Building brain-inspired computing systems: examining the role of nanoscale devices IEEE Nanotechnol. Mag. 12 19–35
[42] Rumelhart D E, Hinton G E and Williams R J 1986 Learning representations by back-propagating errors Nature 323 533–6
[43] Konopik M, Korten T, Lutz E and Linke H 2023 Fundamental energy cost of finite-time parallelizable computing Nat. Commun. 14 447
[44] Dellaferrera G, Woźniak S, Indiveri G, Pantazi A and Eleftheriou E 2022 Introducing principles of synaptic integration in the optimization of deep neural networks Nat. Commun. 13 1885
[45] Hendy H and Merkel C 2022 Review of spike-based neuromorphic computing for brain-inspired vision: biology, algorithms, and hardware J. Electron. Imaging 31 10901
[46] Laughlin S B, de Ruyter van Steveninck R R and Anderson J C 1998 The metabolic cost of neural information Nat. Neurosci. 1 36
[47] Ielmini D and Wong H-S P 2018 In-memory computing with resistive switching devices Nat. Electron. 1 333–43
[48] Owen-Newns D, Robertson J, Hejda M and Hurtado A 2023 GHz rate neuromorphic photonic spiking neural network with a single vertical-cavity surface-emitting laser (VCSEL) IEEE J. Sel. Top. Quantum Electron. 29 1–10
[49] Robertson J, Hejda M, Bueno J and Hurtado A 2020 Ultrafast optical integration and pattern classification for neuromorphic photonics based on spiking VCSEL neurons Sci. Rep. 10 6098
[50] Prucnal P R, Shastri B J, de Lima T F, Nahmias M A and Tait A N 2016 Recent progress in semiconductor excitable lasers for photonic spike processing Adv. Opt. Photonics 8 228–99

[51] Feldmann J, Youngblood N, Wright C D, Bhaskaran H and Pernice W H P 2019 All-optical spiking neurosynaptic networks with self-learning capabilities *Nature* **569** 208–14
[52] Inagaki T, Inaba K, Leleu T, Honjo T, Ikuta T, Enbutsu K, Umeki T, Kasahara R, Aihara K and Takesue H 2021 Collective and synchronous dynamics of photonic spiking neurons *Nat. Commun.* **12** 2325
[53] Lee Y-J, On M B, Xiao X, Proietti R and Yoo S J B 2022 Photonic spiking neural networks with event-driven femtojoule optoelectronic neurons based on Izhikevich-inspired model *Opt. Express* **30** 19360–89
[54] Khan S *et al* 2022 Superconducting optoelectronic single-photon synapses *Nat. Electron.* **5** 650–9
[55] Casaburi A and Hadfield R H 2022 Superconducting circuits that mimic the brain *Nat. Electron.* **5** 627–8
[56] Morozov D V, Casaburi A and Hadfield R H 2021 Superconducting photon detectors *Contemp. Phys.* **62** 69–91
[57] Melikyan A *et al* 2012 Chip-to-chip plasmonic interconnects and the activities of EU project NAVOLCHI *2012 14th Int. Conf. on Transparent Optical Networks (ICTON)* pp 1–3
[58] Izhikevich E M 2000 Neural excitability, spiking and bursting *Int. J. Bifurcation Chaos* **10** 1171–266
[59] Ortega-Piwonka I, Piro O, Figueiredo J, Romeira B and Javaloyes J 2021 Bursting and excitability in neuromorphic resonant tunneling diodes *Phys. Rev. Appl.* **15** 34017
[60] Romeira B, Figueiredo J M L and Javaloyes J 2017 Delay dynamics of neuromorphic optoelectronic nanoscale resonators: perspectives and applications *Chaos* **27** 114323
[61] Miyamoto T, Yamaguchi A and Mukai T 2016 Terahertz imaging system with resonant tunneling diodes *Jpn. J. Appl. Phys.* **55** 32201
[62] Romeira B, Pessoa L M, Salgado H M, Ironside C N and Figueiredo J M L 2013 Photo-detectors integrated with resonant tunneling diodes *Sensors* **13** 9464–82
[63] Rothmayr F, Pfenning A, Kistner C, Koeth J, Knebl G, Schade A, Krueger S, Worschech L, Hartmann F and Höfling S 2018 Mid-infrared GaSb-based resonant tunneling diode photodetectors for gas sensing applications *Appl. Phys. Lett.* **112** 161107
[64] Encomendero J, Protasenko V, Sensale-Rodriguez B, Fay P, Rana F, Jena D and Xing H G 2019 Broken symmetry effects due to polarization on resonant tunneling transport in double-barrier nitride heterostructures *Phys. Rev. Appl.* **11** 34032
[65] Hartmann F, Pfenning A, Rebello Sousa Dias M, Langer F, Höfling S, Kamp M, Worschech L, Castelano L K, Marques G E and Lopez-Richard V 2017 Temperature tuning from direct to inverted bistable electroluminescence in resonant tunneling diodes *J. Appl. Phys.* **122** 154502
[66] Pettinari G, Balakrishnan N, Makarovsky O, Campion R P, Polimeni A, Capizzi M and Patanè A 2013 A micrometer-size movable light emitting area in a resonant tunneling light emitting diode *Appl. Phys. Lett.* **103** 241105
[67] Makarovsky O, Kumar S, Rastelli A, Patanè A, Eaves L, Balanov A G, Schmidt O G, Campion R and Foxon C T 2010 Direct laser writing of nanoscale light-emitting diodes *Adv. Mater.* **22** 3176–80
[68] Van Hoof C, Genoe J, Mertens R, Borghs G and Goovaerts E 1992 Electroluminescence from bipolar resonant tunneling diodes *Appl. Phys. Lett.* **60** 77–79
[69] Slight T J and Ironside C N 2007 Investigation into the integration of a resonant tunnelling diode and an optical communications laser: model and experiment *IEEE J. Quantum Electron.* **43** 580–7
[70] Gravé I, Kan S C, Griffel G, Wu S W, Sa'ar A and Yariv A 1991 Monolithic integration of a resonant tunneling diode and a quantum well semiconductor laser *Appl. Phys. Lett.* **58** 110–2
[71] Hartmann F, Gammaitoni L, Höfling S, Forchel A and Worschech L 2011 Light-induced stochastic resonance in a nanoscale resonant-tunneling diode *Appl. Phys. Lett.* **98** 242109
[72] Zhang W, Al-Khalidi A, Figueiredo J, Al-Taai Q R A, Wasige E and Hadfield R H 2021 Analysis of excitability in resonant tunneling diode-photodetectors *Nanomaterials* **11** 1590
[73] Hejda M *et al* 2023 Artificial optoelectronic spiking neuron based on a resonant tunnelling diode coupled to a vertical cavity surface emitting laser *Nanophotonics* **12** 857–67
[74] Hänggi M and Chua L O 2001 Cellular neural networks based on resonant tunnelling diodes *Int. J. Circuit Theory Appl.* **29** 487–504
[75] Schulman J N, Santos H J D and Chow D H 1996 Physics-based RTD current-voltage equation *IEEE Electron Device Lett.* **17** 220–2
[76] Nishida Y, Nishigami N, Diebold S, Kim J, Fujita M and Nagatsuma T 2019 Terahertz coherent receiver using a single resonant tunnelling diode *Sci. Rep.* **9** 18125
[77] Ortega-Piwonka I, Teruel A E, Prohens R, Vich C and Javaloyes J 2021 Simplified description of dynamics in neuromorphic resonant tunneling diodes *Chaos* **31** 113128
[78] Romeira B, Figueiredo J M L and Javaloyes J 2020 NanoLEDs for energy-efficient and gigahertz-speed spike-based sub-λ neuromorphic nanophotonic computing *Nanophotonics* **9** 4149–62
[79] Romeira B, Borme J, Fonseca H, Gaspar J and Nieder J B 2020 Efficient light extraction in subwavelength GaAs/AlGaAs nanopillars for nanoscale light-emitting devices *Opt. Express* **28** 32302
[80] Jacob B, Camarneiro F, Borme J, Bondarchuk O, Nieder J B and Romeira B 2022 Surface passivation of III–V GaAs nanopillars by low-frequency plasma deposition of silicon nitride for active nanophotonic devices *ACS Appl. Electron. Mater.* **4** 3399–410
[81] Purcell E M 1946 Spontaneous emission probabilities at radio frequencies *Phys. Rev.* **69** 681
[82] Romeira B and Fiore A 2020 Physical limits of nanoLEDs and nanolasers for optical communications *Proc. IEEE* **108** 735–48
[83] Romeira B and Fiore A 2018 Purcell effect in the stimulated and spontaneous emission rates of nanoscale semiconductor lasers *IEEE J. Quantum Electron.* **54** 1–12
[84] Pfenning A, Jurkat J, Naranjo A, Köck D, Hartmann F and Höfling S 2019 Resonant tunneling diode photon number resolving single-photon detectors *Infrared Remote Sensing and Instrumentation XXVII* vol 11128, ed M Strojnik and G E Arnold (SPIE) pp 47–56
[85] Tsakmakidis K L, Boyd R W, Yablonovitch E and Zhang X 2016 Large spontaneous-emission enhancements in metallic nanostructures: towards LEDs faster than lasers [Invited] *Opt. Express* **24** 17916–27
[86] Suhr T, Gregersen N, Yvind K and Mørk J 2010 Modulation response of nanoLEDs and nanolasers exploiting Purcell enhanced spontaneous emission *Opt. Express* **18** 11230–41
[87] Richard O, Blais S, Arès R, Aimez V and Jaouad A 2020 Mechanisms of GaAs surface passivation by a one-step dry process using low-frequency plasma enhanced chemical deposition of silicon nitride *Microelectron. Eng.* **233** 111398
[88] Bedford R and Fallahi M 2004 Analysis of high-reflectivity metal–dielectric mirrors for edge-emitting lasers *Opt. Lett.* **29** 1010–2
[89] Fang C-Y *et al* 2019 Lasing action in low-resistance nanolasers based on tunnel junctions *Opt. Lett.* **44** 3669–72
[90] Ding K and Ning C Z 2012 Metallic subwavelength-cavity semiconductor nanolasers *Light Sci. Appl.* **1** e20

[91] Kuramochi E, Nozaki K, Shinya A, Takeda K, Sato T, Matsuo S, Taniyama H, Sumikura H and Notomi M 2014 Large-scale integration of wavelength-addressable all-optical memories on a photonic crystal chip *Nat. Photon.* **8** 474–81
[92] Jiao Y *et al* 2020 InP membrane integrated photonics research *Semicond. Sci. Technol.* **36** 13001
[93] Malysheva E, Pellegrino D, Fiore A, Williams K and Calzadilla V D 2021 Impact of high temperature post-treatment on photoluminescence performance of passivated InP/In0.53Ga0.47As/InP nanopillars *2021 Conf. on Lasers and Electro-Optics Europe European Quantum Electronics Conf. (Cleo/europe-eqec)* p 1
[94] Al-Taai Q, Morariu R, Wang J, Al-Khalidi A, Al-Moathin A, Romeira B, Figueiredo J and Wasige E 2022 Towards an excitable microwave spike generator for future neuromorphic computing *2021 16th European Microwave Integrated Circuits Conf. (EuMIC)* pp 386–9
[95] Hejda M, Alanis J A, Ortega-Piwonka I, Lourenço J, Figueiredo J, Javaloyes J, Romeira B and Hurtado A 2022 Resonant tunneling diode nano-optoelectronic excitable nodes for neuromorphic spike-based information processing *Phys. Rev. Appl.* **17** 24072
[96] Gokmen T and Vlasov Y 2016 Acceleration of deep neural network training with resistive cross-point devices: design considerations *Front. Neurosci.* **10** 333
[97] Yeh P 1989 Two-wave mixing in nonlinear media *IEEE J. Quantum Electron.* **25** 484–519
[98] Kukhtarev N V, Markov V B, Odulov S G, Soskin M S and Vinetskii V L 1978 holographic storage in electrooptic crystals. II. beam coupling—light amplification *Ferroelectrics* **22** 961–4
[99] Owechko Y and Soffer B H 1995 Holographic neurocomputer utilizing laser diode light source *Optical Implementation of Information Processing* vol 2565, ed B Javidi and J L Horner (SPIE) pp 12–19
[100] Vlieg E A, Dangel R, Horst F and Offrein B J 2021 Photonic photorefractive MAC accelerator for machine learning *Photonics in Switching and Computing 2021* (Optica Publishing Group) p M2B.3
[101] Gehring H, Blaicher M, Hartmann W, Varytis P, Busch K, Wegener M and Pernice W H P 2019 Low-loss fiber-to-chip couplers with ultrawide optical bandwidth *APL Photonics* **4** 010801
[102] Vanmol K, Tuccio S, Panapakkam V, Thienpont H, Watté J and Van Erps J 2019 Two-photon direct laser writing of beam expansion tapers on single-mode optical fibers *Opt. Laser Technol.* **112** 292–8
[103] Vanmol K, Baghdasaryan T, Vermeulen N, Saurav K, Watté J, Thienpont H and Van Erps J 2020 3D direct laser writing of microstructured optical fiber tapers on single-mode fibers for mode-field conversion *Opt. Express* **28** 36147
[104] Perez E F, Lu X, Westly D and Srinivasan K 2020 Robust and automated direct on-axis laser writing of coupling elements for photonic chips *Opt. InfoBase Conf. Pap.* vol 28 pp 39340–53
[105] Yu S, Zui H, Sun X, Liu J, Gu T and Hu J 2020 Optical free-form couplers for high-density integrated photonics (OFFCHIP): universal optical interface *J. Lightwave Technol.* **38** 3358–65
[106] Lindenmann N 2018 *Photonic Wire Bonding as a Novel Technology for Photonic Chip Interfaces* vol 21 (Karlsruhe: KIT Scientific Publishing)
[107] Yu J, Bai Z, Zhu G, Fu C, Li Y, Liu S, Liao C and Wang Y 2020 3D nanoprinted kinoform spiral zone plates on fiber facets for high-efficiency focused vortex beam generation *Opt. Express* **28** 38127–39
[108] Panusa G, Pu Y, Wang J, Moser C and Psaltis D 2020 Fabrication of sub-micron polymer waveguides through two-photon polymerization in polydimethylsiloxane *Polymers* **12** 04852485
[109] Gao H, Chen G F R, Xing P, Choi J W, Low H Y and Tan D T H 2020 High-resolution 3D printed photonic waveguide devices *Adv. Opt. Mater.* **8** 2000613
[110] Oh M C, Chu W S, Shin J S, Kim J W, Kim K J, Seo J K, Lee H K, Noh Y O and Lee H J 2016 Polymeric optical waveguide devices exploiting special properties of polymer materials *Opt. Commun.* **362** 3–12
[111] Baghdasaryan T, Vanmol K, Thienpont H, Berghmans F, Geernaert T and Erps J V 2020 Design and two-photon direct laser writing of low-loss waveguides, tapers and S-bends *J. Phys. Photon.* **4** 045001
[112] Moughames J, Porte X, Larger L, Jacquot M, Kadic M and Brunner D 2020 3D printed multimode-splitters for photonic interconnects *Opt. Mater. Express* **10** 2952
[113] Adão R M R, Romeira B and Nieder J B 2021 Design and fabrication of 3D interconnects for photonic neural networks using two-photon polimerization *Conf. on Lasers and Electro-Optics, OSA Technical Digest* (*Optica Publishing Group*) p ATh1R.7
[114] Romeira B, Nieder J B, Jacob B, Adão R M R, Camarneiro F, Arturo J, Hejda M, Hurtado A, Lourenço J and Alves D C 2021 Subwavelength neuromorphic nanophotonic integrated circuits for spike-based computing: challenges and prospects *Proc. SPIE* **11804** 118040D
[115] Adão R M R, Alves T, Maibohm C, Romeira B and Nieder J B 2022 Two-photon polymerization-based simulation and fabrication of 3D microprinted suspended waveguides for on-chip optical interconnects *Opt. Express* **30** 9623–42
[116] Adão R M R, Caño-García M, Maibohm C and Nieder J B 2021 Photonic polymeric structures and electrodynamics simulation method based on a coupled oscillator finite-difference time-domain (O-FDTD) approach *Opt. Express* **29** 11903–16
[117] Adão R M R, Caño-Garcia M, Maibohm C, Romeira B and Nieder J B 2021 Oscillator finite-difference time-domain (O-FDTD) electric field propagation model: integrated photonics and networks *EPJ Web of Conf.* vol 255 p 01005
[118] Martin E, Ernoult M, Laydevant J, Li S, Querlioz D, Petrisor T and Grollier J 2021 EqSpike: spike-driven equilibrium propagation for neuromorphic implementations *iScience* **24** 102222
[119] Bohte S M, Kok J N and Poutré H L 2002 Error-backpropagation in temporally encoded networks of spiking neurons *Neurocomputing* **48** 17–37
[120] Zenke F and Ganguli S 2018 SuperSpike: supervised learning in multilayer spiking neural networks *Neural Comput.* **30** 1514–41
[121] Ponulak F and Kasiński A 2010 Supervised learning in spiking neural networks with ReSuMe: sequence learning, classification, and spike shifting *Neural Comput.* **22** 467–510
[122] Büchel J, Zendrikov D, Solinas S, Indiveri G and Muir D R 2021 Supervised training of spiking neural networks for robust deployment on mixed-signal neuromorphic processors *Sci. Rep.* **11** 23376