PAPER • OPEN ACCESS

Massively scalable wavelength diverse integrated photonic linear neuron

To cite this article: Matthew van Niekerk *et al* 2022 *Neuromorph. Comput. Eng.* **2** 034012

View the article online for updates and enhancements.

You may also like

- Deuteron electromagnetic form factors and tensor polarization observables in the framework of the hard-wall AdS/QCD model
  Narmin Huseynova, Shahin Mamedov and Jannat Samadov

- Extended reaction kinetics model for non-thermal argon plasmas and its test against experimental data
  Marjan Stankov, Markus M Becker, Tomas Hoder et al.

- Wafer-level testing of inverse-designed and adjoint-inspired dual layer Si-SiN vertical grating couplers
  Thomas Van Vaerenbergh, Sean Hooten, Mudit Jain et al.

NEUROMORPHIC
Computing and Engineering

CrossMark

Massively scalable wavelength diverse integrated photonic linear neuron

Matthew van Niekerk[1],[*] , Anthony Rizzo[2] , Hector Rubio[1] , Gerald Leake[3],
Daniel Coleman[3], Christopher Tison[4], Michael Fanto[4],
Keren Bergman[2] and Stefan Preble[1]

1   Microsystems Engineering, Rochester Institute of Technology, Rochester, NY, United States of America
2   Department of Electrical Engineering, Columbia University, New York, NY, United States of America
3   State University of New York Polytechnic Institute, Albany, NY, United States of America
4   Information Directorate, Air Force Research Laboratory, Rome, NY, United States of America
*   Author to whom any correspondence should be addressed.

E-mail: mv7146@rit.edu

Abstract

As computing resource demands continue to escalate in the face of big data, cloud-connectivity and the internet of things, it has become imperative to develop new low-power, scalable architectures. Neuromorphic photonics, or photonic neural networks, have become a feasible solution for the physical implementation of efficient algorithms directly on-chip. This application is primarily due to the linear nature of light and the scalability of silicon photonics, specifically leveraging the wide-scale complementary metal-oxide-semiconductor manufacturing infrastructure used to fabricate microelectronics chips. Current neuromorphic photonic implementations stem from two paradigms: wavelength coherent and incoherent. Here, we introduce a novel architecture that supports coherent *and* incoherent operation to increase the capability and capacity of photonic neural networks with a dramatic reduction in footprint compared to previous demonstrations. As a proof-of-principle, we experimentally demonstrate simple addition and subtraction operations on a foundry-fabricated silicon photonic chip. Additionally, we experimentally validate an on-chip network to predict the logical 2 bit gates AND, OR, and XOR to accuracies of 96.8%, 99%, and 98.5%, respectively. This architecture is compatible with highly wavelength parallel sources, enabling massively scalable photonic neural networks.

1. Introduction

In recent years, integrated photonics has emerged as a viable and highly appealing candidate for neural network (NN) hardware implementation. Photonic neural networks (PNN) hold particular promise compared to traditional electronic implementations due to the highly favorable physical properties of photons compared to electrons for performing linear operations; namely, (i) the bosonic nature of light permits high degrees of parallelism in spatial mode and wavelength, (ii) such linear transformations are naturally implemented using standard photonic elements (i.e., beamsplitters and phase shifters) at the speed of light, and (iii) these operations can consume orders of magnitude less energy than their electronic counterparts [1–5]. In particular, silicon photonics provides an ideal platform for PNNs due to its compatibility with the standard complementary metal-oxide-semiconductor infrastructure used to fabricate microelectronic chips and the high index contrast between silicon and silicon dioxide which permits unparalleled device density [3]. Numerous approaches to optical NNs have been demonstrated in various platforms, including as spiking [6, 7], phase-change materials [8–10], metal-sulfide microfibers [11, 12], photoelectric multiplication [10, 13], and semiconductor optical amplifiers [14, 15]. However, two primary categories have emerged for PNNs in the silicon photonics platform: wavelength coherent and incoherent [3].

Wavelength coherent designs rely on careful placement and tuning of Mach–Zehnder interferometers (MZI), such that optical interference provides the multiplication and summation operations. Carefully

designed MZI mesh circuits can implement any real-valued matrix on-chip by reconfiguring phase shifters according to singular value decomposition [1, 16]. The coherent optical linear neuron (COLN) architecture implements vector–vector multiplication and summation by leveraging MZIs as inputs and weights for multiplication and coherent interference as the summation operator [17, 18]. Wavelength incoherent designs rely on multi-wavelength operation, often employing a broadband photodetector to perform summation, but through this method the phase of the signal is lost to optical-electrical (OE) conversion. For example, in the 'Broadcast & Weight' protocol, multiple wavelength channels are launched as inputs along the same waveguide, which allows for parallel weighting with micro-ring resonator weight banks [19]. A balanced photodetector pair sums the vector–vector multiplication; however, multiple micro-ring weight banks can operate in parallel for vector-matrix operation, as demonstrated in [20].

In this work, we present a novel architecture that leverages both coherent *and* incoherent aspects to create a massively scalable PNN specific linear operator. This architecture requires fewer components, provides orders-of-magnitude footprint reduction, and consumes substantially less power than comparable designs. As an initial proof-of-principle, we experimentally demonstrate simple addition and subtraction on-chip. Additionally, we implement and demonstrate a NN task on-chip, designed to perform logic gate operations (AND, OR, and XOR). Due to the compact footprint, low energy consumption, and scalability of the demonstrated linear neuron, these results pave the way toward large-scale PNNs with hundreds to thousands of neurons on a single silicon chip.

2. Proposed architecture

We begin our design derivation with inspiration from the COLN [17]. The operating principle of the COLN employs four key stages: fan-out, input, weighting, and fan-in, as shown in figure 1(a). At the fan-out, a continuous-wave (CW) optical carrier signal is split onto N copies. Here, the input and weight stage is represented by amplitude modulators, which are implemented in hardware as Mach–Zehnder modulators (MZM) and impose the inputs and weights onto the CW carrier signal. This signal is recombined from N paths down to 1 path, performing summation of the N input/weight products at the fan-in through constructive interference. Figure 1(b) demonstrates the equivalent NN diagram for this circuit, which performs the linear summation of the inputs multiplied by the synaptic weights, preparing the carrier signal for the activation stage. This optical circuit and the NN both reduce to the general form:
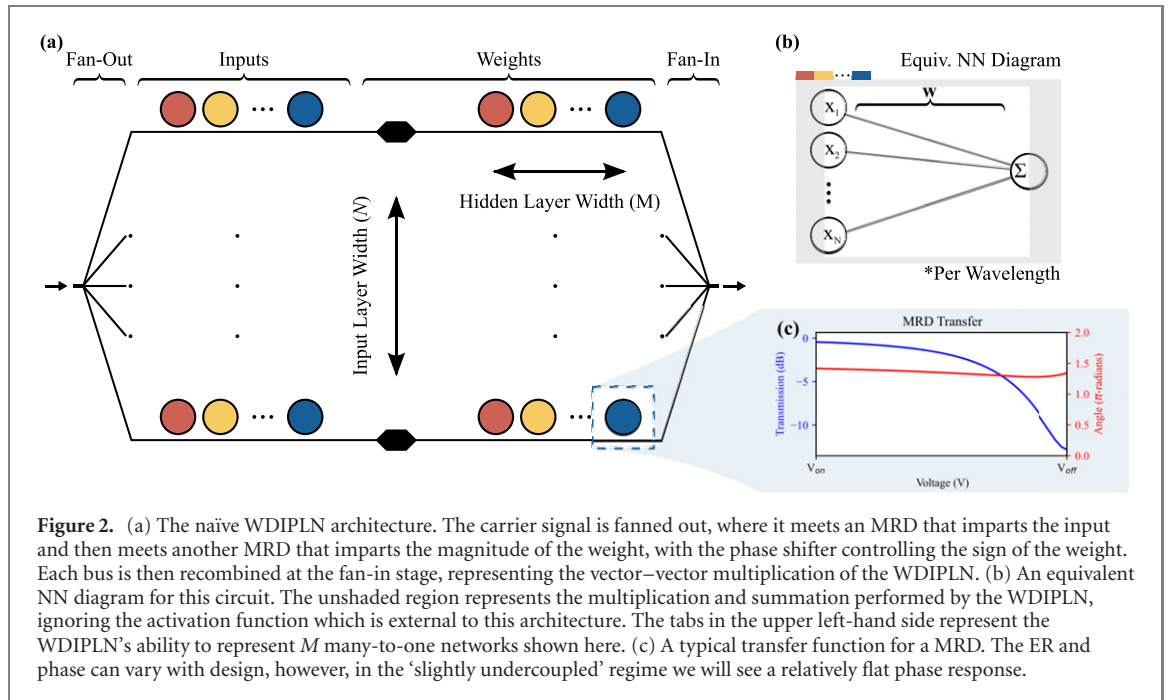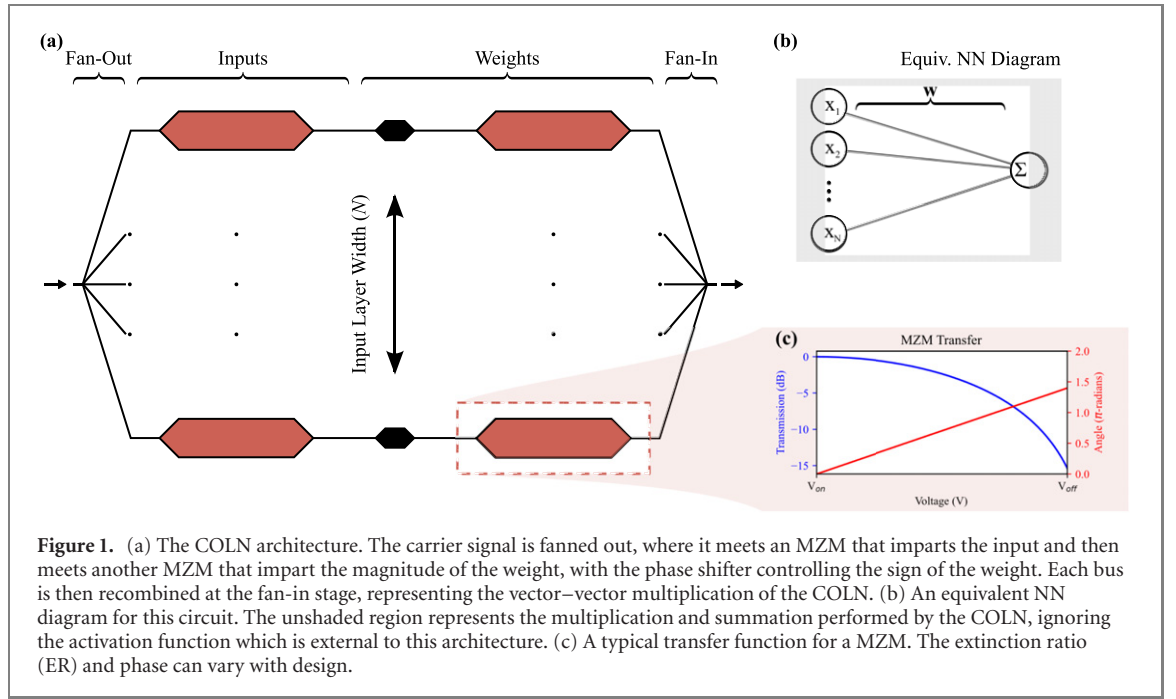
$$\text{out} = \sum_{n}^{N} \mathbf{w}_n \mathbf{x}_n, \tag{1}$$

where, for the optical circuit, out is the electric field output, N as the number of layers in the width (or fan-out), $\mathbf{w}$ is the weight vector, and $\mathbf{x}$ is the input vector. In this formulation, $\mathbf{w}$ and $\mathbf{x}$ are represented by the electro-optic transfer function of the MZM in each layer similar to that seen in figure 1(c). A key feature of the COLN circuit is that the output is in a state which, with minimal additional passive circuitry, is fully compatible both with an activation based on OE, electrical processing, and electrical-optical (EO) conversion [21], direct optical-electrical-optical activation [22] or all-optical, non-linear activation [23, 24].

In a fiber-optic based system, phase-shifters and splitters are readily available off-the-shelf [17]. However, when attempting to scale these down for an integrated photonic system, the size of MZM translates into a practical barrier. Current PN diode, free-carrier MZM devices each occupy a physical space of >1 mm^2 in on-chip area, severely limiting the scalability of this architecture. The first on-chip demonstration of the COLN in a silicon photonic chip utilizes a 25 mm^2 chip to implement a 4 input/weight linear neuron, implementing the weights with thermal phase shifters, which are dramatically smaller at the cost of update speed [25, 26].

Our circuit architecture, which we subsequently refer to as wavelength diverse integrated photonic linear neuron (WDIPLN), begins as a COLN with naïve replacement of MZMs with micro-resonant devices (MRD), either micro-ring resonators or a micro-disks, as shown in figure 2(a). With MRDs, we immediately benefit from the decrease in footprint. Single-channel MRDs are commonly <100 μm^2 in size [27]. In addition, careful design of the MRD allows for a desirable adaptation of the previous electro-optic transfer function, displayed in figure 2(c). The MRD acts as an amplitude filter both in the input and weight stage of the network, for each bus in the layer width. MRDs are sensitive, but if the coupling condition is designed to be slightly under-coupled, it is possible to mitigate some of the phase sensitivity (for example, a depletion biased MRD would remain under-coupled during operation). This enables the coherent summation at the fan-in stage to remain relatively stable for each channel.
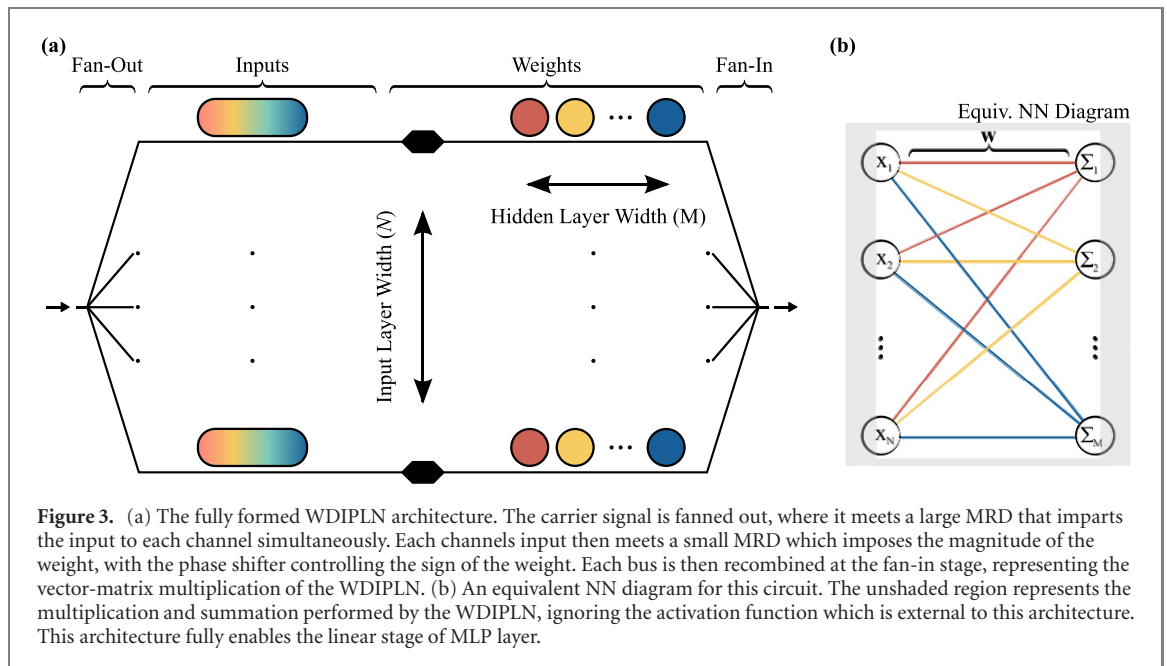
Naïve WDIPLN introduces a feature beyond the COLN, namely wavelength diversity. For sufficiently small radii MRD (or, conversely, sufficiently large free-spectral-range (FSR)), we can place additional MRDs at new

Figure 1. (a) The COLN architecture. The carrier signal is fanned out, where it meets an MZM that imparts the input and then meets another MZM that impart the magnitude of the weight, with the phase shifter controlling the sign of the weight. Each bus is then recombined at the fan-in stage, representing the vector–vector multiplication of the COLN. (b) An equivalent NN diagram for this circuit. The unshaded region represents the multiplication and summation performed by the COLN, ignoring the activation function which is external to this architecture. (c) A typical transfer function for a MZM. The extinction ratio (ER) and phase can vary with design.

Figure 2. (a) The naïve WDIPLN architecture. The carrier signal is fanned out, where it meets an MRD that imparts the input and then meets another MRD that imparts the magnitude of the weight, with the phase shifter controlling the sign of the weight. Each bus is then recombined at the fan-in stage, representing the vector–vector multiplication of the WDIPLN. (b) An equivalent NN diagram for this circuit. The unshaded region represents the multiplication and summation performed by the WDIPLN, ignoring the activation function which is external to this architecture. The tabs in the upper left-hand side represent the WDIPLN's ability to represent *M* many-to-one networks shown here. (c) A typical transfer function for a MRD. The ER and phase can vary with design, however, in the 'slightly undercoupled' regime we will see a relatively flat phase response.

resonant wavelengths (or along specific channels, similar to wavelength-division-multiplexing (WDM) in optical communications). For example, in silicon photonics a sufficiently large channel spacing has been demonstrated with a 2.5 μm radius microdisk modulator, with >55 μm, full width at half maximum of ~0.2 nm, and high quality factor (>8000), which indicates high channel isolation [27].

Recently, a similar WDM scheme was proposed using MZMs, but the WDIPLN implementation with ring resonators have the advantage that the wavelength multiplexing/de-multiplexing is done intrinsically using the single device in a compact footprint [18]. We show a schematic for the naïve WDIPLN in figure 2(a). This enables wavelength diverse operation of the coherent circuit *at each wavelength channel*, such that we have an incoherent circuit which acts like a coherent operator at each channel (λ), with a slight adaptation of equation (1) to

$$\text{out}_\lambda = \sum_n^N \mathbf{w}_{\lambda,n} \mathbf{x}_{\lambda,n}. \tag{2}$$

Figure 3. (a) The fully formed WDIPLN architecture. The carrier signal is fanned out, where it meets a large MRD that imparts the input to each channel simultaneously. Each channels input then meets a small MRD which imposes the magnitude of the weight, with the phase shifter controlling the sign of the weight. Each bus is then recombined at the fan-in stage, representing the vector-matrix multiplication of the WDIPLN. (b) An equivalent NN diagram for this circuit. The unshaded region represents the multiplication and summation performed by the WDIPLN, ignoring the activation function which is external to this architecture. This architecture fully enables the linear stage of MLP layer.

We can claim equation (2) for WDIPLN design with sufficiently isolated channels, where the interaction of each resonance occurs only in and around that wavelength. As discussed above, this is achieved by using sufficiently small MRDs with large FSRs. Using isolated wavelength channels enables the resonant based circuit to act as any m individual COLN circuit for $m \in M$, where M here represents the number of wavelength channels in operation *and* the number of COLN circuits represented by a single WDIPLN circuit. Therefore, this design not only leverages smaller footprint devices, but enables further footprint reduction of a given system of linear neurons from the limit of one optical circuit per perceptron to one optical circuit per fully connected multi-layer perceptron (MLP) by using the input/weight pairs at each wavelength as shown in figure 2(b).

In the naïve case, we have simply swapped each MZM with an MRD, such that the circuit requires $2NM$ MRDs, for layer width N and wavelength channels (or layer depth) M. This case is the same as using M COLN circuits, which is equivalent only in abstracted device count complexity. Thus, an actual network with COLNs would require M copies of the circuit resulting in a larger implementation. However, we recognize that the input stage MRDs are redundant. Therefore, to realize full WDIPLN we replace all M input MRDs for a given layer with a carefully designed, larger radius MRD with an FSR that matches the channel spacing, as outlined in figure 3(a). This adjustment lowers the total device count to $N(M + 1)$, which represents a dramatic improvement in device count scalability in addition to the physical size. As seen in figure 3(b), the WDIPLN circuit in this form represents the linear weighting and summation stage of a fully connected MLP. A caveat of the fully-realized WDIPLN architecture is the inability to realize phase shifting per weight. In this way, we must limit the weight attribution to $[-1, 0]$ or $[0, 1]$. The phase shifters in this architecture are here as a global change to all weights simultaneously, serving both as path balancing and switching between subtractive or additive behavior. Table 1 summarizes the key differences between the COLN and WDIPLN, including scaling rules, physical size, electrical I/O, and power consumption estimates from available foundry process design kits (PDKs) [28–30]. We note that the physical size is calculated purely as fundamental device area, although both the COLN and WDIPLN require additional optical and electrical routing to connect to the outside world. This routing footprint was omitted from the total since it is highly implementation-specific and depends heavily on packaging constraints (i.e., edge coupled versus grating coupled, flip-chip bonded or wire-bonded, etc). However, is is clear that the extremely small device footprint of the WDIPLN architectures leave ample room for additional circuitry, such that optical couplers and electrical bond pads will dominate the final device area. For example, 288 electrical pads (60 μm × 60 μm) on a 16 × 18 grid at 150 μm spacing take up 6.48 mm^2 in on-chip area. If we add optical couplers (400 μm × 20 μm) on the edge of the chip at a large pitch for ease of packaging, the total on-chip area is <7 mm^2 for the WDIPLN of $N = 8, M = 8$. In addition, we note the power consumption values are also simply estimated from available information in [29], which utilize thermal isolation methods to achieve a $P_\pi \sim 2.8$ mW. Importantly, these values serve to illustrate the improved power budget of the WDIPLN architecture—regardless of technology implementation.
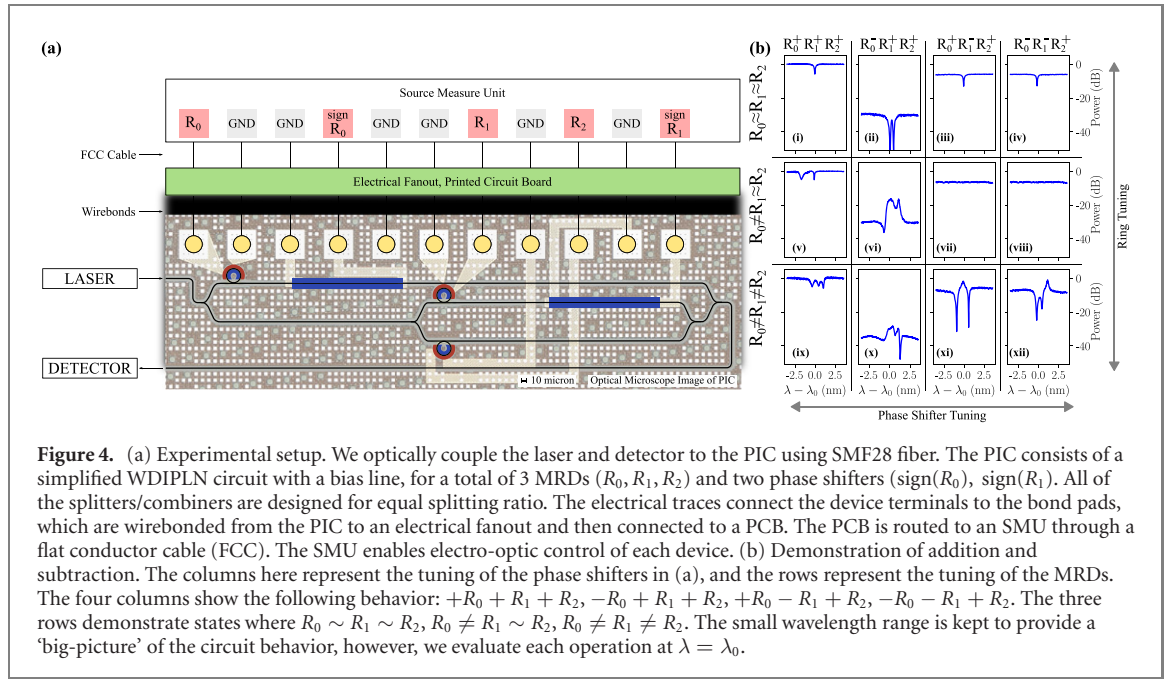
Table 1. Table comparing the primary differences between the COLN and WDIPLN architectures. We consider the nominal COLN, a COLN where weights are done with thermal MZIs, the naïve WDIPLN and nominal WDIPLN. For physical size calculations, we only consider the footprint of individual elements and recognize additional routing, both optical and electrical, will be necessary for each design. The physical size and power consumption values were estimated from available foundry PDKs in [28–30].

Architecture (variation)	COLN (nominal)	COLN (W/thermal MZI)	WDIPLN (naïve)	WDIPLN (nominal)
Input element	MZM	MZM	Small MRD	Large MRD
Approx. element size (mm^2)	0.8	0.8	10^{-4}	10^{-2}
Weight element	MZM	Thermal MZI	Small MRD	Small MRD
Approx. element size (mm^2)	0.8	10^{-1}	10^{-4}	10^{-4}
Scaling rule	$2NM$	$2NM$	$2NM$	$N(M+1)$
Physical size (mm^2)	Component size only (no routing considered)			
$N = 8, M = 1$	12.8	7.2	1.6×10^{-3}	8.08×10^{-2}
$N = 8, M = 8$	102.4	57.6	1.28×10^{-2}	8.64×10^{-2}
Electrical I/O	Consider 4 electrical I/O per element (2 EO, 2 thermal)			
$N = 8, M = 1$	64	64	64	64
$N = 8, M = 8$	512	512	512	288
Power consumption (mW)	Same configuration (2 EO, 2 thermal)			
	EO power—$(P_{\text{on/off}})$			
$N = 8, M = 1$	2.72	1.36	1.6	0.48
$N = 8, M = 8$	21.76	10.88	12.8	3.84
	Thermal power—(P_{π})			
$N = 8, M = 1$	89.6	89.6	89.6	25.2
$N = 8, M = 8$	716.8	716.8	716.8	201.6

3. Experimental demonstration of simple addition

As an initial demonstration, we show the ability of WDIPLN to add and subtract in a simplified form. This represents a fundamental building block of the architecture. While tuning the MRDs ultimately imposes the input/weight onto the circuit, each of the N buses in the circuit are additionally reconfigurable via a phase shifter. The phase shifter rotates the phase of a given bus so that we are able to express the sign coefficient of the weights as either -1 or 1. Bias plays a role in many NN architectures, effectively shifting the set point of the linear fit. Similarly, in this architecture, we can introduce a bias to change our set point. From an optical perspective, this allows us to validate the operation of the circuit via direct measurement through a photodetector at the output [17]. The summation operator occurs as optical interference, which will 'add' (constructively interfere) when the phase is the same, and 'subtract' (destructively interfere) when the phase is opposite. For an initial demonstration, we consider the optical experiment in figure 4(a). We designed an WDIPLN circuit with an added bias. There are five active circuit elements: the bias ring, R_0, bias phase shifter, PS_0, top arm ring R_1, top arm phase shifter PS_1 and bottom arm ring R_2. Each element is electrically connected via a wirebond to an electrical fanout, which is in turn wirebonded to a printed circuit board (PCB) and connected to a source measure unit (SMU). The circuit is optically coupled from two single mode fibers (SMF28) embedded in a tilted fiber array to two grating couplers on either side of the circuit. A laser and detector pair are connected to the other ends of the respective SMF28 fibers. The MRD we employ is a carrier injection-based PIN micro-ring resonator. We employ a PIN for proof of principle due to the large wavelength change, however, future systems will utilize application specific, optimized MRDs operating in carrier depletion mode. The photonic integrated circuit (PIC) was fabricated in American Institute for Manufacturing (AIM) Photonics' 300 mm silicon photonics process [28].

We demonstrate a variety of circuit configurations as summarized in figure 4(b). We achieve these configurations by tuning the phase shifters to 0 or π (columns) and setting the rings at various resonant wavelengths (rows), for which we indicate the state of each ring by the label of R_0, R_1, or R_2. The bias path is slightly shorter than the internal WDILPN, which means we can utilize the phase shifter and the natural wavelength-dependent interference to extract the set-point behavior. Sub-figures (i), (v) and (ix) show three states of addition with phase shifters tuned to 0 where the rings are all aligned to a single resonance ($R_0 = R_1 = R_2$, (i)), R_1 is tuned off ($R_1 \neq R_0 = R_2$, (v)), and all the rings are at different resonances ($R_0 \neq R_1 \neq R_2$, (ix)). These states manifest as one, two, or three resonance peaks, respectively. Sub-figures (ii), (vi) and (x) demonstrate the corresponding states where the bias phase is at π rather than 0. Due to manufacturing variations, we observe behavior that deviates from the designed point. In (ii), we expect no signal to pass through; however, due to path and y-branch imbalance, the circuit floor is ~ -30 dB—additionally, small differences in the

Figure 4. (a) Experimental setup. We optically couple the laser and detector to the PIC using SMF28 fiber. The PIC consists of a simplified WDIPLN circuit with a bias line, for a total of 3 MRDs (R_0, R_1, R_2) and two phase shifters (sign(R_0), sign(R_1)). All of the splitters/combiners are designed for equal splitting ratio. The electrical traces connect the device terminals to the bond pads, which are wirebonded from the PIC to an electrical fanout and then connected to a PCB. The PCB is routed to an SMU through a flat conductor cable (FCC). The SMU enables electro-optic control of each device. (b) Demonstration of addition and subtraction. The columns here represent the tuning of the phase shifters in (a), and the rows represent the tuning of the MRDs. The four columns show the following behavior: $+R_0 + R_1 + R_2$, $-R_0 + R_1 + R_2$, $+R_0 - R_1 + R_2$, $-R_0 - R_1 + R_2$. The three rows demonstrate states where $R_0 \sim R_1 \sim R_2$, $R_0 \neq R_1 \sim R_2$, $R_0 \neq R_1 \neq R_2$. The small wavelength range is kept to provide a 'big-picture' of the circuit behavior, however, we evaluate each operation at $\lambda = \lambda_0$.

ring are seen in the transmission. Therefore, the maximum value in (ii) is ~ -33 dB at λ_0, where the resonant wavelength sits, which indicates a slight imbalance but overall high suppression of any difference in the rings. Sub-figure (vi) shows two peaks above the minimum, as $R_0 \neq R_1 = R_2$, and (x) shows three distinguishable peaks as $R_0 \neq R_1 \neq R_2$. The peaks are overlapping such that distinguishing individual peaks is difficult. Sub-figures (iii) and (iv) demonstrate a single resonance, similar to (i); however, note the maximum power is roughly -6 dB since no light passes through in the lower path and each y-branch contributes additional loss of 3 dB. Sub-figures (vii) and (viii) show detuned R_0, and $R_1 = R_2$, where these rings fully cancel out each others' magnitudes (subtraction). Sub-figures (xi) and (xii) demonstrate $R_0 \neq R_1 \neq R_2$ for the subtraction operator, which is seen by the distinct peak difference above and below the bias line. Finally, R_1, R_2 each flip signs between (xi) and (xii) due to the phase control of the bias line.

4. Experimental demonstration of logic gates

We designed a second circuit in order to demonstrate a learning task. This design reflects the WDIPLN architecture for $N = 2, M = 1$, as seen in figure 5(a). This circuit places both input and weight rings along the inner buses, along with a bias branch that also contains a ring. We wirebond out to an electrical carrier, which we then connect to programmable SMU channels in order to electrically control each element. We optically couple the laser and detector through fiber to grating couplers on the input and output waveguides, respectively.

We selected a simple network architecture that is able to reconfigure the weights in order to represent AND, OR and XOR gates, shown in figure 5(b). We trained these networks on a CPU using PyTorch [31]. For the activation function we employ the rectified linear-unit (ReLU) for all stages. We can create a simple map from the trained model to our physical chip by reconfiguring the chip in between the three linear stages, leading to the three non-input nodes h_1, h_2 and out. For the ring resonators, we define a value of '0' as $V_{\text{off}} = 1.4$ V and '1' as $V_{\text{on}} = 1.2$ V. This ensures we are operating in the slightly undercoupled regime for both '0' and '1'. We note that for this demonstration, the V_{off} and V_{on} values are globally set for all rings. In addition, the bias ring is set to V_{off} and the bias phase shifter is set to $\pi/2$, specifically for visibility of the full range of $[-1, 1]$ through direct detection as opposed to inference [17].

The process of reconfiguration and implementation of the two hidden nodes and the output is shown in figures 5(c)–(e). For the two hidden nodes, h_1 and h_2, we configure the weights of the optical circuit, namely $|W_1|$ and $|W_2|$, according to the pairs of $[w_{11}, w_{21}]$ and $[w_{12}, w_{22}]$, respectively. In addition, if the sign of the weight is negative, we adjust the corresponding phase shifter, namely sign(W_1) and sign(W_2), from 0 to π. Once we configure the chip for h_1, we feed the four input pairs $[0, 0], [0, 1], [1, 0], [1, 1]$ into X_1, X_2. For each input, we measure the output from the detector, apply the ReLU activation function used in training with a '3 dB cutoff', and record the result. After the two hidden nodes are measured, we reconfigure the circuit for the final stage, according to the final weight stage from training t_1, t_2. Finally, we feed the four recorded pairs of h_1, h_2 into the circuit as inputs, measure the output, apply the activation function and record the result.
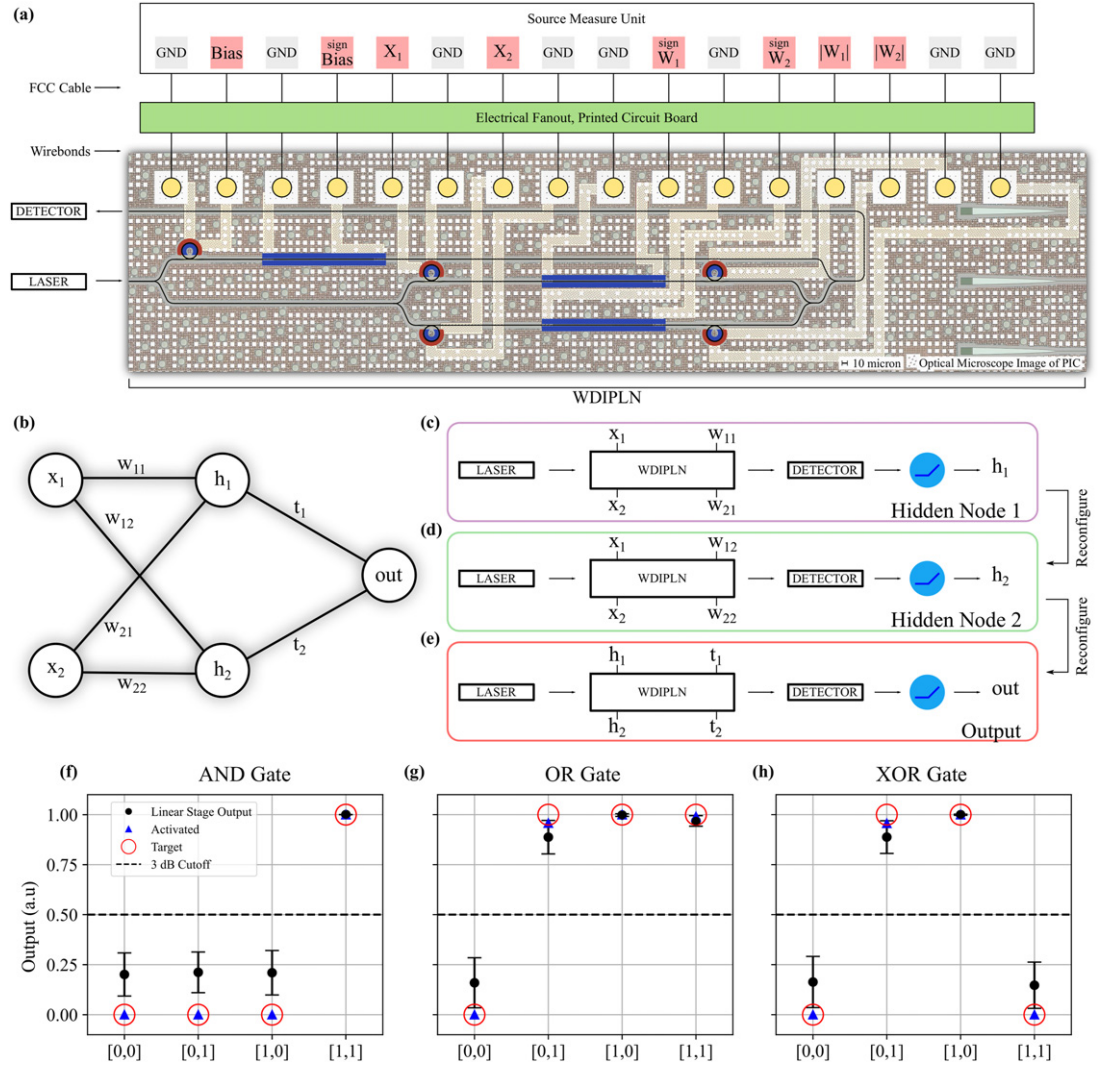
Figure 5. Photonic neural network experimental demonstration. (a) Experimental set-up, similar to figure 4. We optically couple the laser and detector to the PIC. The PIC consists of a WDIPLN circuit with $N = 2, M = 1$ and a bias line, for a total of 5 MRDs and 3 phase shifters. All of the splitters/combiners are designed for equal splitting ratio. The electrical wiring goes out the bond pads, which are wirebonded from the PIC to an electrical fanout and then connected to a PCB. The PCB is routed to a SMU through a FCC. The SMU enables electro-optic control of each device. (b) The selected architecture for the simple NN we implement for this task. Inputs (x_1, x_2) are connected by a weight matrix, **w**, to a hidden layer with nodes h_1, h_2. From here, a simple weight vector, **t**, connects the hidden layer to the output. (c)–(e) The configure-recycle process for the circuit in (a). We send in the input pairs $[0, 0], [0, 1], [1, 0], [1, 1]$ as x_1, x_2 for the first two stages, and subsequently send in h_1, h_2 in the final stage. We configure the weights of the WDIPLN to match that in (b) step-by-step from pairs $[w_{11}, w_{21}], [w_{12}, w_{22}]$, and $[t_1, t_2]$ for each stage of the network, respectively. Each configure-to-measurement cycle takes approximately 1 s, which is dominated by the read/write speeds of the SMUs. (f)–(h) Results of the experimental demonstration. The AND, OR and XOR gate correctly predict the outputs with accuracy 96.8%, 99%, and 98.5%, respectively.

According to the training-to-implementation scheme described above and in figures 5(c)–(e), we demonstrate the circuit's performance for three different 2 bit logic gates, namely AND, OR and XOR. In training, the accuracy of this simple network reaches ∼100%. The resulting outputs of the three gates are shown in figures 5(f)–(h). The black dots are the square root (i.e. magnitude) of the output values measured by the detector at the laser probe wavelength, which in this experiment was globally set to $\lambda_0 = 1,526$ nm. The error bars represent the standard output deviation for a 100 pm window around, $\lambda_0 \pm 50$ pm. The blue triangles show the outputs of the circuit after activation, and the red circles show the target for the gates. The AND, OR and XOR gates achieve predictive accuracy of 96.8%, 99%, and 98.5%, respectively.

5. Conclusion

Silicon photonics holds promise for implementing large-scale PNNs directly on chip with substantially improved performance compared to electronic implementations. In this work, we propose and demonstrate a novel PNN architecture, WDIPLN, which exhibits highly favorable characteristics compared to previous

literature. The WDIPLN derives from the COLN architecture detailed in [17], enabling similar vector–vector multiplication and summation using an optical carrier signal with the additional benefits of wavelength parallelism, drastically reduced device footprint, and significantly lower energy consumption. The majority of these beneficial properties come from exchanging large, power-hungry MZMs with compact MRDs. While we can trivially exchange the fundamental elements (MRDs for MZMs) in a one-to-one manner to enable multi-channel operation, we show that by replacing the input MRDs with a single, multi-channel MRD with an FSR equal to the weighting channel spacing we can reduce the total device count significantly. As a first proof-of-principle, we experimentally demonstrated simple addition and subtraction between rings in a WDIPLN circuit. We then show a WDIPLN configured with $N = 2$ and $M = 1$, which we configure and recycle to perform 2 bit logic gates (AND, OR and XOR). We observe the implementation accuracy for each gate to be 96.8%, 99%, and 98.5%, respectively. Furthermore, the natural wavelength parallelism of the proposed WDIPLN architecture can be exploited using chip-based Kerr frequency comb sources [32] for massive scaling in the frequency domain, similar to recent demonstrations in the silicon photonics platform for high bandwidth data communications [33]. This demonstration opens new opportunities in massively parallel silicon photonic PNNs and paves the way to large-scale systems in the thousand-neuron regime on a single chip with minimal energy consumption.

Acknowledgments

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Appendix A. Operating conditions and variation of MRDs

The mechanism of tuning for MRDs here is assumed to be electro-optic (where thermo-optic is an accepted simplification) and well defined by free-carrier plasma dispersion [34]. The transfer function from the electrical to optical domain is defined as

$$E_{\text{all-pass-ring}}(\lambda; V) = e^{i(\pi + \theta(\lambda, V))} \frac{a(\lambda, V) - r\, e^{-i\theta(\lambda, V)}}{1 - ra(\lambda, V)e^{i\theta(\lambda, V)}}, \tag{A.1}$$

where a the voltage (V) dependent absorption (or loss) in the cavity lumped with intrinsic cavity loss, θ is the wavelength- and voltage-dependent phase shift, r is the self-coupling coefficient (with a corresponding cross coupling coefficient, κ, such that $\kappa^2 + r^2 = 1$). Equation (A.1) depicts the coupled effect of tuning, which both changes the loss and the relative cavity size (i.e. phase shift).

In this work, we employ carrier injection-based ring resonators through the well-known free carrier plasma dispersion effect in silicon. These rings demonstrate a large spectral tuning range, as seen in figure A1(a). However, this shift comes at the cost of change of coupling conditions due to the strongly voltage-dependent round-trip loss. Changing the coupling conditions from over coupled to critically coupled to under coupled has a large impact on the phase shift associated the ring. It is precisely this phase that we aim to leverage in this architecture. Therefore, we operate the rings in the slightly-to-moderately under coupled regime in this experiment, which corresponds to voltages $V > 1.1$ V, as seen in figure A1(b).

Additionally, this demonstration did not employ thermal tuners to shift the rings coarsely. This leaves the demonstration susceptible to fabrication variations. For large-scale demonstrations, these thermal phase shifters can be implemented with isolated heaters with high efficiency [35, 36] in concert with reverse-biased MRDs for low-power electro-optic configuration [27, 37].
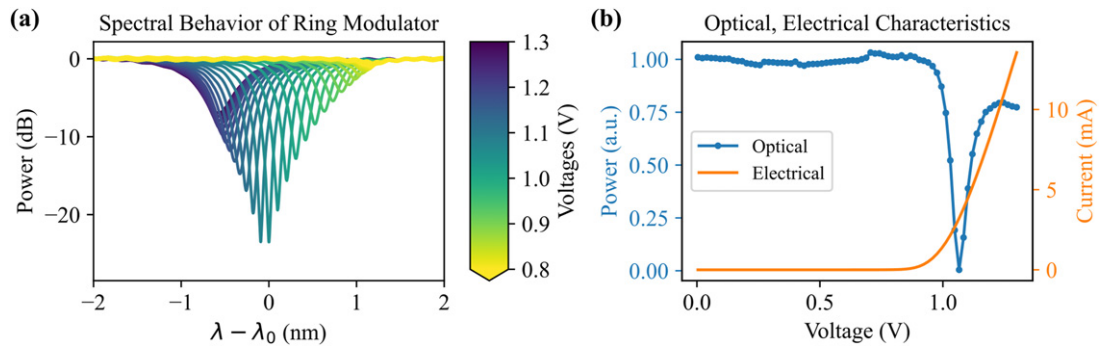
Figure A1. (a) The spectral behavior of the ring modulator used in our demonstration. As the voltage increases, the resonant wavelength blue-shifts to a lower wavelength. In addition, we see that the ring traverses the initial condition (over-coupled), to critical coupled, finally to under-coupled. This ring was selected due to the large tuning range for initial demonstration. (b) Optical and electrical characterization of the ring. A single wavelength, $\lambda - \lambda_0 = 0$ from (a), is plotted to show the change in optical transmission as a function of applied bias. From an electrical perspective, we see that the operating regime for this diode is indeed in forward-bias, with a maximum power for full tunability of approximately 12 mW.
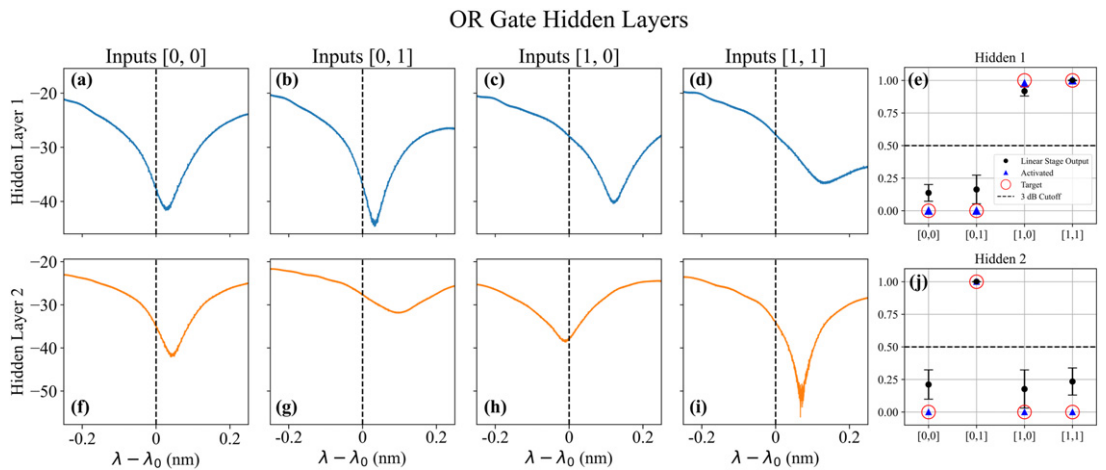

Figure A2. Full implementation sequence for the hidden layers of the OR gate. (a)−(d) The circuit response when configured to the trained weights of hidden node 1 for the incoming inputs [0, 0], [0, 1], [1, 0], and [1, 1]. (e) Circuit output for hidden node 1 at the probe wavelength, with a range of ± 100 pm. (a)−(d) Show the circuit response when configured to the trained weights of hidden node 2 for the incoming inputs [0, 0], [0, 1], [1, 0], and [1, 1]. (e) Circuit output for hidden node 2 at the probe wavelength, with a range of ± 100 pm.

Appendix B. Hidden layer demonstration results

The hidden layers are not explicitly necessary in order to run the network since the final stage weights are known. However, they are important in demonstrating the full generality of the WDIPLN architecture, including the ability to reconfigure. In figure A2, we show an in-depth look at the procedure of validating the hidden nodes, h_1 and h_2. First, we configure the circuit to measure h_1, as shown in figure 5(c). Figures A2(a)−(d) show the optical circuit response in the hidden 1 configuration for the four logic inputs, with (e) showing the outputs at the probe wavelength. Figures A2(f)−(i) shows the optical circuit response in the hidden 2 configuration for the four logic inputs, with (j) showing the outputs at the probe wavelength.

ORCID iDs

Matthew van Niekerk https://orcid.org/0000-0002-5367-0062
Hector Rubio https://orcid.org/0000-0001-7315-6039

References

[1] Shen Y *et al* 2017 *Nat. Photon.* **11** 441−6
[2] Totovic A R, Dabos G, Passalis N, Tefas A and Pleros N 2020 *IEEE J. Sel. Top. Quantum Electron.* **26** 1−15

[3]   Shastri B J, Tait A N, Ferreira de Lima T, Pernice W H, Bhaskaran H, Wright C D and Prucnal P R 2021 *Nat. Photon.* **15** 102–14
[4]   Nahmias M A, De Lima T F, Tait A N, Peng H T, Shastri B J and Prucnal P R 2019 *IEEE J. Sel. Top. Quantum Electron.* **26** 1–18
[5]   Cheng Q, Kwon J, Glick M, Bahadori M, Carloni L P and Bergman K 2020 *Proc. IEEE* **108** 1261–82
[6]   Nahmias M A, Shastri B J, Tait A N and Prucnal P R 2013 *IEEE J. Sel. Top. Quantum Electron.* **19** 1–12
[7]   Shastri B J, Tait A N, Nahmias M A and Prucnal P R 2014 (arXiv:1407.2917)
[8]   Cheng Z, Ríos C, Pernice W H, Wright C D and Bhaskaran H 2017 *Sci. Adv.* **3** e1700160
[9]   Wu C, Yu H, Lee S, Peng R, Takeuchi I and Li M 2021 *Nat. Commun.* **12** 96
[10] Feldmann J *et al* 2021 *Nature* **589** 52–8
[11] Gholipour B, Bastock P, Craig C, Khan K, Hewak D and Soci C 2015 *Adv. Opt. Mater.* **3** 635–41
[12] Sun J *et al* 2018 *Adv. Funct. Mater.* **28** 1804397
[13] Hamerly R, Bernstein L, Sludds A, Soljačić M and Englund D 2019 *Phys. Rev. X* **9** 021032
[14] Shi B, Calabretta N and Stabile R 2019 *IEEE J. Sel. Top. Quantum Electron.* **26** 1–11
[15] Fouskidis D E, Zoiros K E and Hatziefremidis A 2020 *IEEE J. Sel. Top. Quantum Electron.* **27** 1–15
[16] Zhang H *et al* 2021 *Nat. Commun.* **12** 1–11
[17] Mourgias-Alexandris G, Totović A, Tsakyridis A, Passalis N, Vyrsokinos K, Tefas A and Pleros N 2019 *J. Lightwave Technol.* **38** 811–9
[18] Totovic A, Giamougiannis G, Tsakyridis A, Lazovsky D and Pleros N 2022 *Sci. Rep.* **12** 1–13
[19] Tait A N, Nahmias M A, Shastri B J and Prucnal P R 2014 *J. Lightwave Technol.* **32** 3427–39
[20] Tait A N, De Lima T F, Zhou E, Wu A X, Nahmias M A, Shastri B J and Prucnal P R 2017 *Sci. Rep.* **7** 1–10
[21] Williamson I A, Hughes T W, Minkov M, Bartlett B, Pai S and Fan S 2019 *IEEE J. Sel. Top. Quantum Electron.* **26** 1–12
[22] Tait A N, De Lima T F, Nahmias M A, Miller H B, Peng H T, Shastri B J and Prucnal P R 2019 *Phys. Rev. Appl.* **11** 064043
[23] Jha A, Huang C and Prucnal P R 2020 *Opt. Lett.* **45** 4819–22
[24] Mourgias-Alexandris G, Tsakyridis A, Passalis N, Tefas A, Vyrsokinos K and Pleros N 2019 *Opt. Express* **27** 9620–30
[25] Giamougiannis G *et al* 2021 Silicon-integrated coherent neurons with 32GMAC/sec/axon compute line-rates using EAM-based input and weighting cells *2021 European Conf. Optical Communication (ECOC)* (Piscataway, NJ: IEEE) pp 1–4
[26] Mourgias-Alexandris G, Totovic A, Passalis N, Dabos G, Tefas A and Pleros N 2020 Neuromorphic computing through photonic integrated circuits *Smart Photonic and Optoelectronic Integrated Circuits XXII* vol 11284 (International Society for Optics and Photonics) p 1128403
[27] Timurdogan E, Sorace-Agaskar C M, Sun J, Shah Hosseini E, Biberman A and Watts M R 2014 *Nat. Commun.* **5** 4008
[28] Fahrenkopf N M, McDonough C, Leake G L, Su Z, Timurdogan E and Coolbaugh D D 2019 *IEEE J. Sel. Top. Quantum Electron.* **25** 1–6
[29] Giewont K *et al* 2019 IEEE J. Sel. Top. Quantum Electron. 25 1–11
[30] Littlejohns C G *et al* 2020 *Appl. Sci.* **10** 8201
[31] Paszke A, Gross S, Massa F, Lerer A *et al* 2019 Pytorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems* vol 32 ed H Wallach, H Larochelle, A Beygelzimer, F d' Alché-Buc, E Fox and R Garnett (Curran Associates) pp 8024–35
[32] Gaeta A L, Lipson M and Kippenberg T J 2019 *Nat. Photon.* **13** 158–69
[33] Rizzo A *et al* 2021 arXiv:2109.10297
[34] Soref R and Bennett B 1987 *IEEE J. Quantum Electron.* **23** 123–9
[35] Coenen D, Oprins H, Ban Y, Ferraro F, Pantouvaki M, Van Campenhout J and De Wolf I 2022 *J. Lightwave Technol.* **40** 4357–63
[36] van Niekerk M *et al* 2022 Wafer-scale-compatible substrate undercut for ultra-efficient SOI thermal phase shifters *Conference on Lasers and Electro-Optics* (*Technical Digest Series*) (Optical Society of America) p JTh3B. 24
[37] van Niekerk M *et al* 2022 High extinction micro-ring modulator *Conference on Lasers and Electro-Optics* (*Technical Digest Series*) (Optical Society of America) p STh4K. 5