All-photonic artificial-neural-network processor via nonlinear optics

Jasvith Raj Basani,[1,2] Mikkel Heuck,[3,4] Dirk R. Englund,[3] and Stefan Krastanov[3,5,*]

[1] *Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science, Hyderabad Campus, Pilani, Telangana 500078, India*

[2] *Department of Electrical and Computer Engineering, Institute for Research in Electronics and Applied Physics, and Joint Quantum Institute, University of Maryland, College Park, Maryland 20742, USA*

[3] *Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA*

[4] *Department of Electrical and Photonics Engineering, Technical University of Denmark, Building 343, Kgs. Lyngby 2800, Denmark*

[5] *College of Information and Computer Sciences, University of Massachusetts Amherst, 140 Governors Drive, Amherst, Massachusetts 01003, USA*

(Received 5 July 2022; revised 30 April 2024; accepted 24 May 2024; published 3 July 2024)

Optics and photonics have recently captured interest as a platform to accelerate linear matrix processing, otherwise a bottleneck in traditional digital electronics. In this paper we propose an all-photonic computational accelerator wherein information is encoded in the amplitudes of frequency modes stored in a single ring resonator. Interaction among these modes is enabled by nonlinear optical processes. Both the matrix multiplication and elementwise activation functions on these modes (the artificial neurons) are performed through coherent processes, enabling the direct representation of negative and complex numbers without having to pass through digital electronics, a common limitation in today's photonic architectures. This design also has a drastically lower hardware footprint compared with today's electronic and optical accelerators, as the entirety of the matrix multiplication happens in a single multimode resonator on chip. Our architecture is unique in providing a completely unitary, reversible mode of computation, enabling on-chip analog Hamiltonian-echo backpropagation for gradient descent and other self-learning tasks. Moreover, the computational speed increases with the power of the pumps to arbitrarily high rates, as long as the circuitry can sustain the higher optical power. Lastly, the design presented here is a less demanding version of a future room-temperature quantum computational device. Therefore, while this architecture is already viable today, direct reinvestments in it would be enabling its evolution into quantum computational hardware.

DOI: 10.1103/PhysRevApplied.22.014009

I. INTRODUCTION

The last decade has witnessed phenomenal advances in the domain of machine learning, with applications ranging from natural language processing [1], structural biology [2], and even game playing [3]. With the growing accessibility of large datasets and larger computational power, machine learning models have been increasing in complexity to tackle a multitude of problems. The requirement for better performance in these networks has necessitated the development of hardware accelerators, specifically for the training of deep neural networks. Recently, with advances in silicon photonics [4,5], optical computing has been introduced as an attractive platform to carry out large-scale computational schemes. Properties of light, such as coherence and superposition, blended with the vast array of CMOS-compatible optical devices has made photonics a fruitful direction of exploration for efficiently and effectively implementing computational schemes.

Photonic implementations of neural networks have been proposed and successfully realized in free-space environments using spatial light modulators [6–9], vertical-cavity surface-emitting laser arrays [10], diffractive media [11,12], and homodyne detection [13]. A number of techniques have been used to construct optical neural networks via photonic integrated circuitry, particularly with interferometric meshes [14–20], electro-optics [21], and time-wavelength multiplexing [22,23]. These architectures have been exploited to build scalable devices for spiking neural networks [24–26] and reservoir computing [27–30]. The photonic platform has garnered interest from scientists and engineers alike, to leverage the massive parallelism being offered by the multiple degrees of freedom of light (wavelength, polarization, phase, etc.) Photonic solutions also greatly reduce energy consumption due to data transfer

[*] Contact author: aps.acc@krastanov.org

and computational operations by performing operations via passive optical interactions [12–14,31–33].

Our proposal for a fully photonic implementation of an artificial neural network is based on nonlinear optical intermodulation. In contrast to previous approaches [11,12,14,31,34,35], we encode information in the complex amplitudes of frequency states that act as neurons, in a multimode cavity. Information regarding the linear operations that the neuron modes undergo is encoded in the amplitudes of controlled pump modes and is enabled via four-wave mixing (FWM) [36]. Furthermore, unlike other optical [33,37–39] and opto-electronic [40–44] approaches, we also propose a scheme to perform the elementwise activation function coherently via nonlinear optical processes [45–49]. This approach lets us represent negative (or even complex) activation values, a problem plaguing other optical approaches.

The proposed processor is rapidly reprogrammable, and can be realized using only microring resonators (which can be fabricated easily via well-established lithography techniques). Moreover, the entirety of the computation performed by the proposed hardware is, in principle, reversible and unitary, opening up many possibilities for low-power (even reversible) computation, and on-chip efficient analog Hamiltonian-echo backpropagation [50] for *in situ* learning tasks. We also find that the speed of computation performed by our device scales with the pump power, hence providing for extremely fast operations, to within limitations imposed by the hardware. Finally, the accelerator presented here can serve as a near-term commercially viable stepping stone for more demanding quantum hardware, particularly for room-temperature quantum computation [51].

This paper is organized as follows. In the following Sec. II, we introduce the scheme for matrix multiplication via the method *active* coupling of “neuron” pulses in a multimode optical cavity. We discuss the Hamiltonian and matrix transformation implemented by the optical cavity and establish the time dynamics of the neuron modes. The limitations of the available operations and methods for overcoming these limitations are discussed. Section III discusses our implementation of the nonlinear activation function. In Sec. IV we perform simulations to train our neural network accelerator on the Modified National Institute of Standards and Technology (MNIST) dataset [52] to illustrate the performance of our hardware design in different parameter regimes. Our paper concludes with a discussion of the results, the potential for *in situ* training, and the prospects for experimental realization of this work.

II. PROGRAMMABLE TRANSFORMATIONS VIA FOUR-WAVE MIXING

Deep neural networks (DNNs) are a class of artificial neural networks that, fundamentally, consist of multiple stacked layers of neurons, each connected via a matrix multiplication ($\vec{x} \mapsto W\vec{x}$) and an elementwise nonlinear activation function ($x_i \mapsto \sigma(x_i)$). For a DNN of arbitrary depth, the input to the $(k+1)$th layer is related to the input of the kth layer as

$$x_i^{(k+1)} = \sigma\left(\sum_j W_{i,j}^{(k)} x_j^{(k)}\right). \tag{1}$$

We propose realizing the matrix multiplication by $W^{(k)}$ in a multimode optical cavity. For instance, consider an optical cavity implemented as a microring resonator that supports a frequency comb in the telecommunication range (around 1550 nm). The frequency states supported by the microring resonator are chosen to be either “pump” or “neuron” modes, that interact with each other via the process of FWM. Our design encodes information to be processed in the complex amplitudes of the neuron modes, while the matrix-multiplication operations are enabled by interaction with controlled pump modes. With FWM being an inherently third-order nonlinear optical process, the microring resonator will have to be fabricated from a material that facilitates the third-order nonlinear optical response described with a large $\chi^{(3)}$ susceptibility coefficient. The neural network weights that act as interconnects between the network’s layers are encoded in the strength of the pumps.

A. Method of active coupling for programmable linear transformations

Our protocol for implementing a fully connected neural net layer employs actively capturing and storing neuron modes into a microring resonator as depicted in Fig. 1. In order to realize such active capturing, we consider the resonator to be coupled to the waveguide via a tunable coupler that controls the coupling coefficient $\gamma(t)$. The mixing of the neuron modes to perform linear operations on them is enabled via FWM with time-dependent control pumps. We term this scheme as the method of active coupling. Similar dynamics can be realized without actively capturing the pump and neuron modes—by simply allowing them to propagate and interact while propagating through a series of resonators [53]. The latter scheme, which we term as the method of passive coupling, is an experimentally less demanding version, but a hardware inefficient version of active coupling. The method of passive coupling is described in Appendix A. To understand the mechanism through which the neurons are intermodulated, we first consider a resonator with only four modes, i.e., two neural modes and two pump modes. The lower two modes are the pumps that drive the system, denoted by operators $(\hat{p}_1, \hat{p}_2)$. The two higher-frequency modes act as neurons, denoted by $(\hat{a}_1, \hat{a}_2)$. The Hamiltonian associated with the

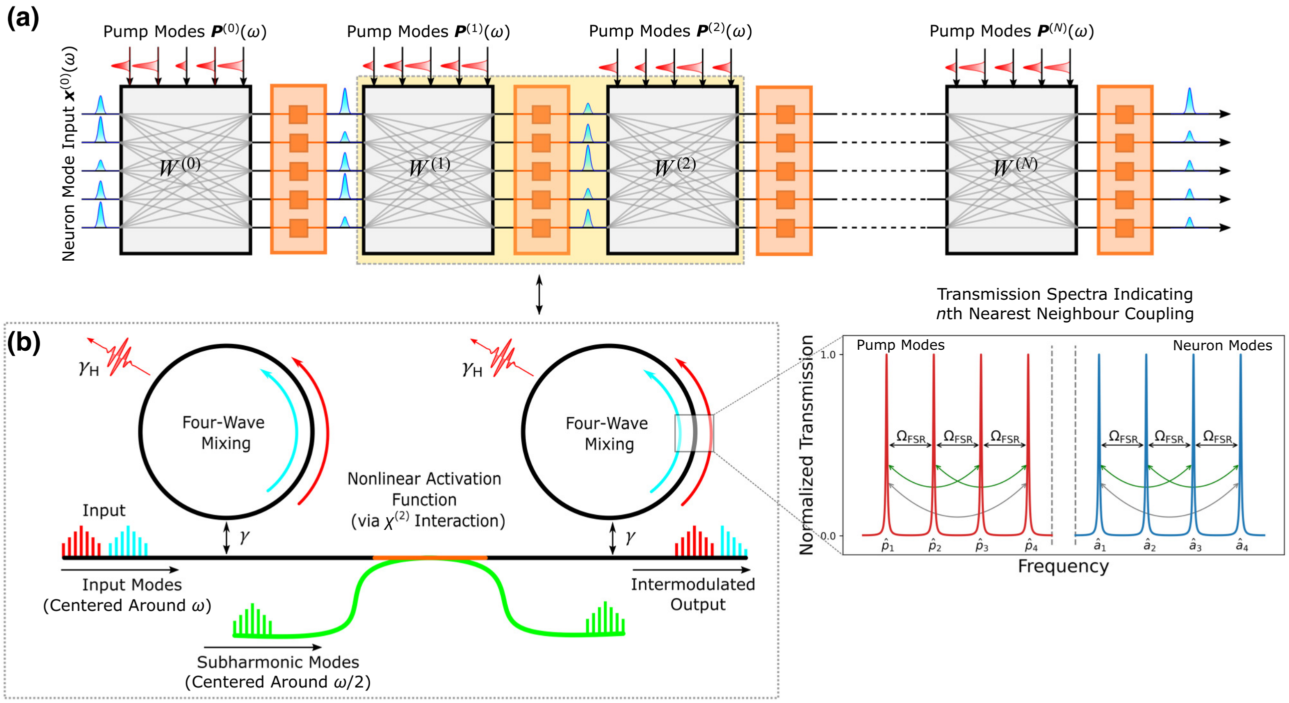
(a)
Pump Modes $\boldsymbol{P}^{(0)}(\omega)$
Pump Modes $\boldsymbol{P}^{(1)}(\omega)$
Pump Modes $\boldsymbol{P}^{(2)}(\omega)$
Pump Modes $\boldsymbol{P}^{(N)}(\omega)$
Neuron Mode Input $\boldsymbol{x}^{(0)}(\omega)$
$W^{(0)}$
$W^{(1)}$
$W^{(2)}$
$W^{(N)}$
(b)
γ_H
γ_H
Four-Wave Mixing
Four-Wave Mixing
Nonlinear Activation Function (via $\chi^{(2)}$ Interaction)
Input
γ
γ
Input Modes (Centered Around ω)
Intermodulated Output
Subharmonic Modes (Centered Around $\omega/2$)
Transmission Spectra Indicating nth Nearest Neighbour Coupling
Normalized Transmission
1.0
0.0
Pump Modes
Neuron Modes
Ω_{FSR}
Ω_{FSR}
Ω_{FSR}
Ω_{FSR}
Ω_{FSR}
Ω_{FSR}
$\hat{p}_1$
$\hat{p}_2$
$\hat{p}_3$
$\hat{p}_4$
$\hat{a}_1$
$\hat{a}_2$
$\hat{a}_3$
$\hat{a}_4$
Frequency

FIG. 1. **(a)** Schematic of the neural network represented as a sequence of N layers. The information being processed is encoded in the amplitudes of neuron modes, i.e., frequency modes (blue), while the linear transformations $W^{(i)}$ are implemented via interaction with strong classical pump modes (red). The nonlinear elementwise activation function (given by orange blocks) occurs during propagation through waveguides via nonlinear optical interactions of the neuron modes with additional pump modes (green). **(b)** Hardware for consecutive layers of the optical neural network are shown: microring resonators connected via a waveguide. Each ring resonator is coupled to the waveguide with a coupling constant $\gamma(t)$ and experiences internal losses $\gamma_{\rm H}$. The transmission spectra of the microring resonator is shown alongside, where the nth nearest neighbour pump and neuron modes are coupled (given by green and grey arrows).

interaction of the four waves is

$$\hat{H} = \hbar\chi\left(\hat{p}_1\hat{p}_2^\dagger\hat{a}_1\hat{a}_2^\dagger\right) + \text{H.c.} \tag{2}$$

The coupling coefficient χ determines the strength of interaction, incorporating effects from several parameters including the nonlinear susceptibility of the material of our cavity, phase matching, and mode volume realized in the cavity. The pumps in each timestep are assumed to be strong classical modes of light and their operators can be replaced by a classical complex amplitude $\hat{p}_i \mapsto p_i = \sqrt{\langle\hat{n}_i\rangle}e^{i\theta}$, involving the expectation value of the number of photons n_i in the given pump mode and its phase θ. Furthermore, these pumps are much stronger than the other modes and, hence, are nondepletive. We assume that the resonances of the modes obey the FWM energy matching condition, such that $\omega_{p_2} - \omega_{p_1} = \omega_{a_2} - \omega_{a_1}$. During capture or release γ is increased in order to transfer the neuron modes from the waveguide into the resonator or vice versa. During the FWM process γ is kept at its minimal value to avoid information loss via leakage into the environment, thus, the total loss rate Γ can be written as $\Gamma = \gamma + \gamma_{\rm H} = \gamma_{\rm H}$.

The time dynamics of the modes can be solved using coupled mode theory [54,55]. The exact form of the coupled amplitude equations can be found in Appendix A. In the general case for a deep neural network with N neurons, we can extend this formalism to see that pumps that are nth nearest neighbours [i.e., have a frequency difference of $n \times \Omega_{\rm FSR}$ for a ring with the free spectral range (FSR) $\Omega_{\rm FSR}$] couple all the neuron modes at that frequency difference. Without loss of generality, we make the simplifying assumption that the first pump P_1 is much stronger than the other pumps, permitting us to neglect the cross-coupling terms, leading to the following coupled amplitude equations:

$$\dot{P}_i(t) = 0 = -\frac{\Gamma}{2}P_i(t) - \sqrt{\gamma}S_{\text{in},P}(t), \tag{3}$$

$$\begin{aligned}\frac{\mathrm{d}A_i}{\mathrm{d}t} &= \left(-\frac{\Gamma}{2} + i\chi\left|P_1\right|^2\right)A_i \\ &\quad - \chi\left[\sum_{j>i}^{N}\left(P_1P_j^*\right)A_j - \sum_{j<i}^{i-1}\left(P_1^*P_j\right)A_j\right].\end{aligned} \tag{4}$$

Here A_i and P_i represent, respectively, the amplitude of the ith neuron mode and the amplitude of the ith pump mode *inside of the resonator*. The pump amplitudes are set to a scale much higher than the scale of the neuron activations, to permit neglecting the direct neuron-neuron interactions. The encoded data is introduced into the system via the input waveguide mode, denoted by $S_{\text{in},P}$ (representing the activation values of the neurons). In terms of matrix-vector operations, Eq. (4) can be written as $\dot{\vec{A}} = \mathbf{P}\vec{A}$, where the matrix **P** is constant by diagonal (also known as a Toeplitz matrix). The *n*th off-diagonal elements of matrix **P** takes the value $P_1 P_n$. The P_i values might need a correction to account for nonlinear interactions purely between the pumps, however, this is a straightforward matrix inversion problem that does not affect the neural dynamics.

The solution to this system of equations (at the end of a period Δt during which **P** is constant) is $\vec{A}(t = \Delta t) = e^{\Delta t \mathbf{P}} \vec{A}(t = 0)$. While we assumed piecewise constant **P** for simplicity in this example, a freely evolving **P** is just as easy to work with. It is important to note here that the Toeplitz nature of the $N \times N$ matrix **P** gives us only N degrees of freedom, as opposed to N^2 degrees of freedom encoded in the weights of a fully connected deep neural network. This implies that the transformation imposed on the input optical modes during a single timestep (one instance of FWM over the period of Δt), would span only a fraction of the space that would otherwise be spanned by the full group of unitary transformations. To quantify the group of operations that can be spanned by matrices of the form $e^{\Delta t \mathbf{P}}$, we introduce the concept of expressivity.

The expressivity is the average fidelity with which a transformation $\mathbf{T}(\mathbf{P})$ parametrized as $e^{\Delta t \mathbf{P}}$ can represent an arbitrary unitary operation **U**. Numerically, we estimate the expressivity by sampling M Haar-random unitaries $\{\mathbf{U}_i\}_{1 \leq i \leq M}$ and, for each one, we use gradient descent to find the $\mathbf{T}(\mathbf{P}_i)$ that approximates it most closely. We estimate the expressivity using the trace distance as

$$\mathbb{F} = 1 - \frac{1}{M} \sum_{i=1}^{M} \sqrt{\operatorname{tr}\left[(\mathbf{T}_i - \mathbf{U}_i)(\mathbf{T}_i - \mathbf{U}_i)^\dagger\right]}, \qquad (5)$$

which both accounts for imperfections due to losses (deviations from unitarity) and insufficient degrees of freedom. Since this function is convex [56] in both $\mathbf{T}_i$ and $\mathbf{U}_i$, the gradient descent always converges to the global optimum.

The transformation performed by a single layer, i.e., a single matrix of the form $e^{\Delta t \mathbf{P}}$, does not reach expressivity large enough to perform arbitrary unitary transformations. To solve this problem, we leverage the time dependence of the pumps to perform multiple instances of FWM by varying the amplitudes of the pumps in each timestep Δt. Physically, this corresponds to each pump mode consisting of a series of piecewise constant segments, each with a duration of Δt. Each layer of the neural network is now implemented by several noncommuting cascaded matrices of the form $e^{\Delta t \mathbf{P}}$. Thus, after a time of $N\Delta t$, N instances of FWM would be performed resulting in the net transformation having N^2 degrees of freedom, spanning a larger group of operations and, hence, increasing the expressivity. By estimating the expressivity of these compound operations as a function of matrix dimension and number of sublayers, we see that for larger matrices, at higher sublayers, the expressivity reaches unity as illustrated in Fig. 2. This implies that by cascading multiple matrices in a single layer, we can span the group of unitary operations.

A factor that negatively influences the expressivity is the presence of loss, Γ. In the ideal case ($\Gamma \Delta t = \gamma_{\text{H}} \Delta t = 0$) as shown in Fig. 2, we see that the expressivity grows upon cascading sublayers just as in the previous case, approaching unity. For a much more pessimistic case where $\frac{1}{2}\Gamma \Delta t = 1$ (e.g., corresponding to a large cavity loss rate $\Gamma = 2\ \text{ns}^{-1}$ and a control pulse resolution of $\Delta t = 1$ ns), however, there is a high sensitivity to the loss. We observe an increase in the average fidelity upon cascading a few sublayers, beyond which the expressivity begins to decrease due to the pulses entering the decay regime. This arises as a result of the trade-off between losses and the coverage of the $N \times N$ unitary group. As the number of pump steps increases, the number of free parameters increases, therefore allowing us to access a larger fraction of the unitary group. However, increasing the number of pump steps also results in increased losses in the system, thereby restricting the group of accessible unitaries. The result of this trade-off is the rise and then fall in expressivity values seen in the right panel of Fig. 2. As the number of pump steps continues increasing, losses begin to dominate, further restricting the group of accessible unitaries, resulting in a drop in the expressivity. In this case, the final expressivity, even after cascading enough layers to obtain N^2 degrees of freedom, does not reach unity.

Transformations of the form $e^{\Delta t \mathbf{P}}$ can be realized via three-wave mixing [57] as well, with a single pump mode instead of two as proposed above. Solving for the transformation matrix using the coupled mode equations give us a similar result to the one presented above—the difference being that three-wave mixing does not give rise to cross-coupling between different neuron modes. The Hamiltonian associated with the interaction of the three interacting waves would be $\hat{H} = \chi(\hat{p}\hat{a}\hat{b}^\dagger) + \text{H.c.}$, where $\hat{p}$ is the single pump mode. These modes obey the energy matching condition that $\omega_p = \omega_b - \omega_a$. Experimentally implementing this system via three-wave mixing, however, presents engineering challenges in the design of the microring resonator. The energy matching condition requires the frequency of the pump mode to be equal to the difference in frequencies of the neuron modes. This would result in pump modes operating at frequencies much smaller than the neuron modes, i.e., integer multiples of the FSR of the microring resonator. This ring would therefore have to support modes over multiple octaves in order to perform these transformations via three-wave mixing. Spanning across

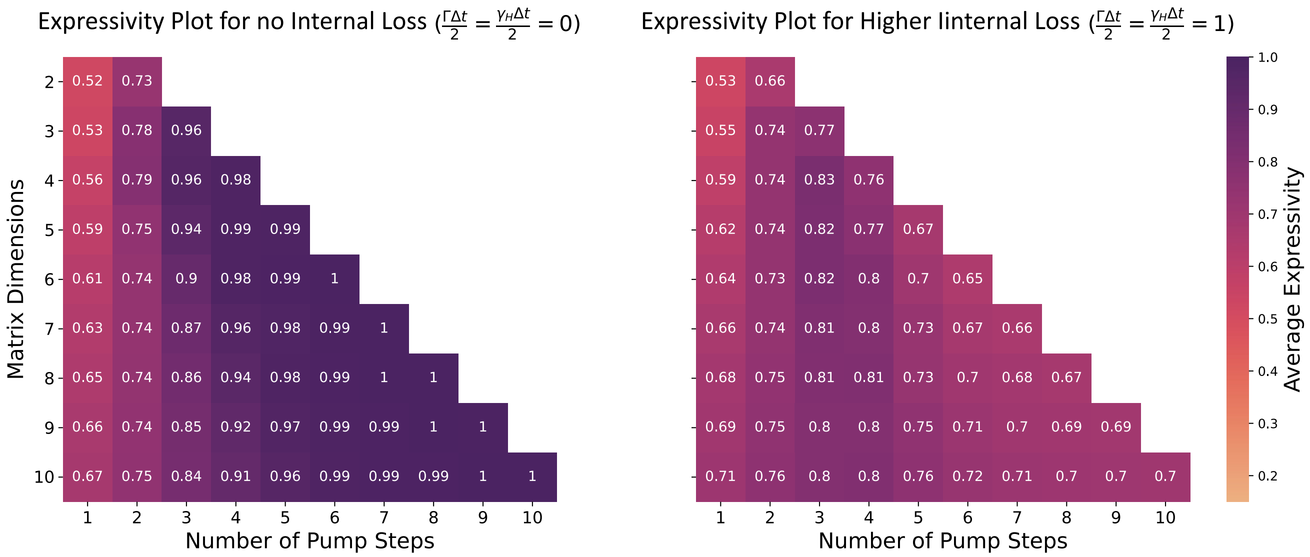

FIG. 2. The expressivity of the *active coupling* transformation of the form $\prod e^{\Delta t\mathbf{P}}$ in different parameter regimes. Each plot displays the average fidelity as we vary the number of timesteps (the horizontal axis) for a given matrix dimension (the vertical axis). On the left, the ideal case of no internal loss ($\Gamma = \gamma_{\mathrm{H}} = 0$) where the expressivity reaches unity at sufficiently many timesteps. On the right, the expressivity at higher loss ($(\Gamma\Delta t/2) = (\gamma_{\mathrm{H}}\Delta t/2) = 1$) never approaches unity. Of note is that the expressivity initially climbs up with the number of layers, until the loss becomes too significant, exponentially growing with the number of layers. Importantly, the expressivity as defined here includes both infidelity due to missing degrees of freedom, and amplitude decay due to leakage from the cavity.

multiple octaves gives rise to differences in refractive indices and Q factors for modes at different frequencies. This leads to difficulties in maintaining the resonance condition and phase matching required for high-efficiency three-wave mixing. Alternatively, pump and neuron modes across multiple octaves could be implemented as an electro-optic frequency comb [58]; this approach would, however, be limited by the speed of the electronics used to couple modes across large frequency bands.

B. Hardware resources and computational speed

As we have seen in the previous section, the rate at which the wave-mixing interactions happen scales as $\chi P'P''$, where P' and P'' denote the pump amplitudes of the main pump and an arbitrary secondary pump. Therefore, the higher the pump power is, the faster the computation can be executed, up to loading and heating constraints. The value for χ for a given piece of hardware is derived below, giving us realistic engineering constraints on the computational speed. From Ref. [60], we see that the nonlinear component of the Hamiltonian is given by

$$\hat{H} = \int \frac{\chi^{(3)}\hat{\mathbf{D}}^4}{4\varepsilon_0^3\eta^8}\mathrm{d}\mathbf{r}, \tag{6}$$

where $\chi^{(3)}$ is the FWM nonlinear susceptibility of the material, ε_0 and η are the vacuum permitivity and refractive index of the material, and $\hat{\mathbf{D}}$ is the electrical displacement field operator. The field operator $\hat{\mathbf{D}}$ is the sum of pump or neuron modes $\hat{m}$ that can be written in terms of the eigenmode $\mathbf{d}(\mathbf{r})$ as [60]

$$\hat{\mathbf{D}}_m(\mathbf{r}) = \sqrt{\frac{\hbar\omega_m}{2}}\hat{m}\mathbf{d}_m(\mathbf{r}) + \text{H.c.}, \tag{7}$$

where $\hat{m}$ is the creation operator for the given mode and the normalization condition $\int |\mathbf{d}(\mathbf{r})|^2\mathrm{d}\mathbf{r} = \varepsilon_0\eta^2$ is fulfilled. Taking into account the energy matching conditions for two neuron modes $\hat{a}_1$ and $\hat{a}_2$ and two pump modes $\hat{p}_1$ and $\hat{p}_2$, and identifying with Eq. (2) gives us

$$\hbar\chi = \frac{3}{2}\frac{\chi^{(3)}}{\varepsilon_0\eta^4 V_{\mathrm{FWM}}}\sqrt{\hbar^4\omega_{a_1}\omega_{a_2}\omega_{p_1}\omega_{p_2}}, \tag{8}$$

where we define the FWM mode volume V_{FWM} as

$$\frac{1}{V_{\mathrm{FWM}}} = \frac{\int_{\mathrm{nl}} d^i_{a_1} d^{j*}_{a_2} d^k_{p_1} d^{l*}_{p_2}\mathrm{d}\mathbf{r}}{\sqrt{\int |\mathbf{d}_{a_1}|^2\mathrm{d}\mathbf{r}\int |\mathbf{d}_{a_2}|^2\mathrm{d}\mathbf{r}\int |\mathbf{d}_{p_1}|^2\mathrm{d}\mathbf{r}\int |\mathbf{d}_{p_2}|^2\mathrm{d}\mathbf{r}}}. \tag{9}$$

The $\int_{\mathrm{nl}}$ denotes integration over the volume of the nonlinear material and i, j, k, l denote the spacial components of the fields between which nonlinear interaction is enabled.

If we are to use a silicon nitride resonator ($\eta = 2.02$ and $\chi^{(3)} = \frac{4}{3}\eta^2 n_2\varepsilon_0 c \approx 3.5 \times 10^{-21}(\mathrm{m}^2/\mathrm{V}^2)$ [61]) with good phase matching such that the FMW mode volume V_{FWM} is comparable to the geometric volume (approximately equal to 1300 μm^3 for a 115 μm radius, 2.5 μm width and 0.73 μm height) [62], we find that $\chi \approx 4.2$ s^{-1}.

The period of complete exchange of energy between two neuron modes can be calculated via the coupled mode equations derived from Eq. (2), leading to $\Delta t = 2\pi/(\chi\langle P_1\rangle\langle P_2\rangle)$, where the maximum amplitudes $\langle P_*\rangle$ are measured in square root of average number of photons. We use these amplitudes as a worst-case estimate of the energy requirements for our design. As we have seen from Fig. 2, increasing $\Delta t\Gamma$ beyond unity significantly decreases the performance of our hardware due to losses, which leads to the requirement $\langle P_1\rangle\langle P_2\rangle > (2\pi\Gamma/\chi)$. For a modern silicon nitride resonator, we can expect a $Q \approx 10^6$ and $\Gamma = \gamma_{\mathrm{H}} = (\omega/Q) \approx 1$ ns^{-1}, therefore, $(2\pi\Gamma/\chi) \approx 10^9$. This implies that we need of the order of 1×10^9 photons in the main pump mode, leading to thermal heating losses from the main pump of the order of $\Gamma\hbar\omega\langle P\rangle^2 \approx 100$ mW.

To summarize, increasing the power of the pumps ($\propto \langle P\rangle^2$) would linearly increase the rate at which computations are performed ($\chi\langle P\rangle^2$) and linearly increase the power dissipated during the computation ($\Gamma\hbar\omega\langle P\rangle^2$). For a typical ring resonator today [59], this implies a computational speed of 1 GHz (1×10^9 sublayer matrix multiplications per second) at dissipation from the main pump of 100 mW. As seen in Fig. 3, both of these figures of merit can be drastically improved in the very near term by employing already demonstrated techniques (higher $\chi^{(3)}$ in slightly more exotic materials like silicon-rich silicon nitride or aluminum gallium arsenide and better Q factors). Curiously, there is a lower bound for the computational speed of our device: we need to provide enough pump power such that the computation happens faster than the rate of decay of the neuron modes.

In practice, the computational speed of the device cannot be scaled arbitrarily by increasing the power of the pumps. The upper limit of the computational speed (and, hence, the pump power) is determined by thermal properties of the material used to fabricate the device. Knowing the safe operating temperature is essential for ensuring that there are no thermally induced nonlinear effects or material damage. The rate of increase of temperature can be calculated from the power dissipated as $\Delta T = \left(\Gamma\hbar\omega\langle P\rangle^2\right)/mc_{\mathrm{p}}$, where m is the mass and c_{p} is the specific heat capacity of the material. Based on the safe operating range of the temperature, the upper limit of the pump power can be calculated. Additional hardware for cooling would allow for increased pump power and, therefore, even faster computational speeds. As mentioned above, there is a restriction on the lower bound for the speed of the device as well. To calculate this lower bound, at least one instance of FWM (of period Δt) would have to be executed before the neurons leak from the cavity. If the value of $\Gamma\Delta t < 1$, one instance of FWM occurs before the neurons decay. This places a lower bound on the computational speed of the device, given by $1/\Gamma$.

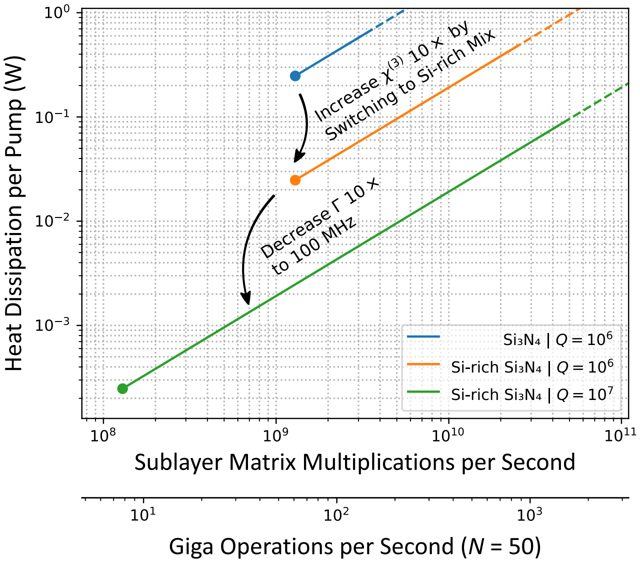

FIG. 3. Estimated computational performance of our design. The horizontal axis is the rate at which sublayer multiplications are performed (i.e., the rate of matrix-vector multiplications, where the matrix is a restricted unitary matrix). The vertical axis is an estimate of the heat dissipation expected in a single ring due to leakage from the pump that encodes the matrix parameters. The top blue line represents a typical silicon nitride ring [59] with $\Gamma = 1$ ns^{-1} and $V_{\mathrm{FWM}} = 1300$ μm^3. Two near-term evolutions are presented as well, first (in orange) using silicon-rich material that significantly increases the $\chi^{(3)}$ susceptibility, and second (in green) developing higher-Q resonators. Lowering the mode volume of the ring would provide similar performance improvements. The curves are cut off to the left due to the constraint seen in Fig. 2 that the computational rate is faster than the decay rate. The second axis shows the equivalent number of giga operations per second (GOPS), taking into account the limited degrees of freedom in a single instance of FWM. This is a function of the number of frequency modes N, which is set to 50 that can be easily realized with today's frequency combs.

To compare the throughput of DNN accelerator architectures, it is helpful to introduce the tera operations per second (TOPS) figure of merit [63], the number of scalar multiplication (and addition) operations implicitly performed by the accelerator. Recent experimental demonstrations for optical neural networks have achieved processing speeds approximately equal to 10–100 TOPS [22,23,64,65], while heuristically designed state-of-the-art digital electronic DNN accelerators operate at approximately similar speeds [66]. As we have established, a single sublayer matrix multiplication (single instance of FWM) modulates all the neuron modes simultaneously. Therefore, during one FWM period of duration Δt we perform the equivalent of $\mathcal{O}(N)$ multiply-accumulate (MAC) operations. A general matrix-vector multiplication would involve $\mathcal{O}(N^2)$ MACs, but as discussed, we need multiple sublayer multiplications (multiple instances of FWM) to achieve that. For a numerical performance estimate, we choose $N = 50$ since matrix multiplication can be readily implemented on 50 frequency modes today [67]. With present day hardware parameters, this scheme reaches processing speeds of 10–100 giga operations per second at comparatively low thermal overhead (Fig. 3). With improved hardware parameters (such as larger quality factors, lower mode volumes, improved effective nonlinear susceptibility) and more neurons (N in the hundreds of modes thanks to frequency combs), such an architecture would efficiently scale into the TOPS regime. The quality factor determines the rate at which neurons decay from the cavity. A larger quality factor implies that the neurons decay slowly from the cavity, and therefore, more instances of FWM can be executed. While this does not directly affect the computational speed of the device, the lower bound of Δt is determined by it. The mode volume, on the other hand, is directly related to the computational speed of the device, given by Eqs. (8) and (9). In our calculations, we estimate a mode volume approximately equal to 1300 μm^3 based on the geometric volume of the resonator. However, photonic crystal cavities can reach mode volumes of the order of 0.1 μm^3, which is 4 orders of magnitude smaller [68]. This suggests that the computational speed can be increased by about 4 orders of magnitude, to operate at speeds of approximately 100 TOPS. Finally, increasing the number of neurons linearly increases the number of multiply-accumulate operations per second as described above. Since frequency combs can have hundreds of modes, increasing the number of neurons could potentially scale up the computational speed of the device by another order of magnitude.

III. NONLINEAR ACTIVATION

The nonlinear activation function is indispensable to the operation of the neural network. Previous implementations of the nonlinear activation function have relied on the use of thermo-optic effects [14,69], hybrid optical-electronic schemes [48,69–71], semiconductor lasers [72, 73], and saturable absorption [74,75]. The nonlinearity we propose relies on nonlinear interactions facilitated by a $\chi^{(2)}$ medium, followed by controllable capture into a ring resonator.

The nonlinearity we propose is based upon a second-order nonlinear interaction (e.g., in a lithium niobate waveguide, characterized by its $\chi^{(2)}$ susceptibility coefficient) [76–78]. We release the neuron mode from the resonator in which the matrix multiplication was performed into the waveguide. We aim to distort the temporal envelope of the neuron mode via the nonlinear interaction with an externally pumped pulse (that we term as the subharmonic mode). This subharmonic mode has a frequency of half of the neuron mode. Following the distortion, we selectively capture [79,80] the neuron mode into the microring resonator that forms the subsequent layer of the neural network. Thus, the distorted pulses are selectively absorbed into the ring, with absorption efficiency dependent on the amount of distortion. The nonlinear distortion is stronger for higher-amplitude pulses, giving rise to a total effective nonlinearity. Figure 4 provides a sketch of the setup and Fig. 5 shows the realized nonlinear activation function. To avoid interactions between different neural modes, i.e., keep the activation function elementwise, a waveguide segment with dispersion can be used to offset the modes in time. The required distance between these dispersed modes for realizing the elementwise activation function will be determined by the optical modulators used to generate the subharmonic modes. State-of-the-art electro-optic modulators have been shown to reach speeds of 40 GHz, meaning the modes will have to be separated by at least 25 ps in time. With an appropriately dispersion-engineered waveguide, silicon has shown to have dispersion coefficients of approximately 4400 ps/nm/km at 1550 nm [81]. If the neurons are assumed to be initially spaced by 1 nm in wavelength, a waveguide approximately 5 m long will be required to separate them out by at least 25 ps in time. Alternatively, racetrack resonators with a large dispersion coefficient (shown to reach approximately 600 ps/km/nm at 1550 nm in silicon nitride [82]) can also be used to disperse the neuron modes.

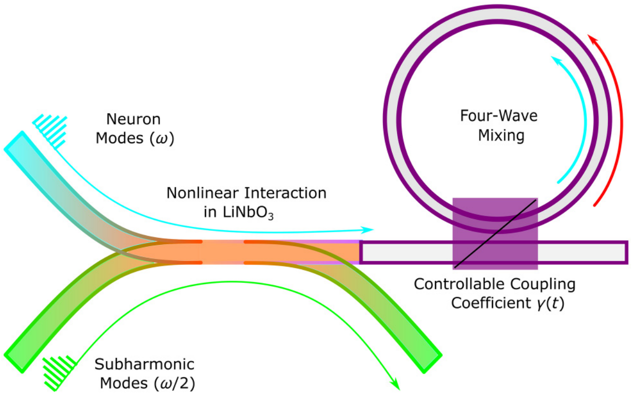
Neuron Modes (ω)
Nonlinear Interaction in LiNbO$_3$
Four-Wave Mixing
Controllable Coupling Coefficient $\gamma(t)$
Subharmonic Modes ($\omega/2$)

FIG. 4. Schematic of the propagating “neural” pulses undergoing the activation function. Input pulses (cyan) are distorted via second-order nonlinear interaction in the $\chi^{(2)}$ waveguide before being captured in the ring resonator. The controllable coupling coefficient $\gamma(t)$ allows us to selectively absorb pulses, with efficiency dependent on how distorted a pulse is.

First, we explore the envelope distortion dynamics for a neural pulse interacting with a subharmonic pump pulse in a waveguide. We parametrize both envelopes as $E_\text{n}(z,t)$ and $E_\text{sub}(z,t)$, where z is the spacial coordinate along the length of the waveguide. As elaborated in the archived simulation code, these envelopes obey [83]

$$\frac{\partial E_\text{n}}{\partial z} + \frac{\eta}{c}\frac{\partial E_\text{n}}{\partial t} = -\kappa E_\text{sub}^2 - \alpha E_\text{n}, \tag{10}$$

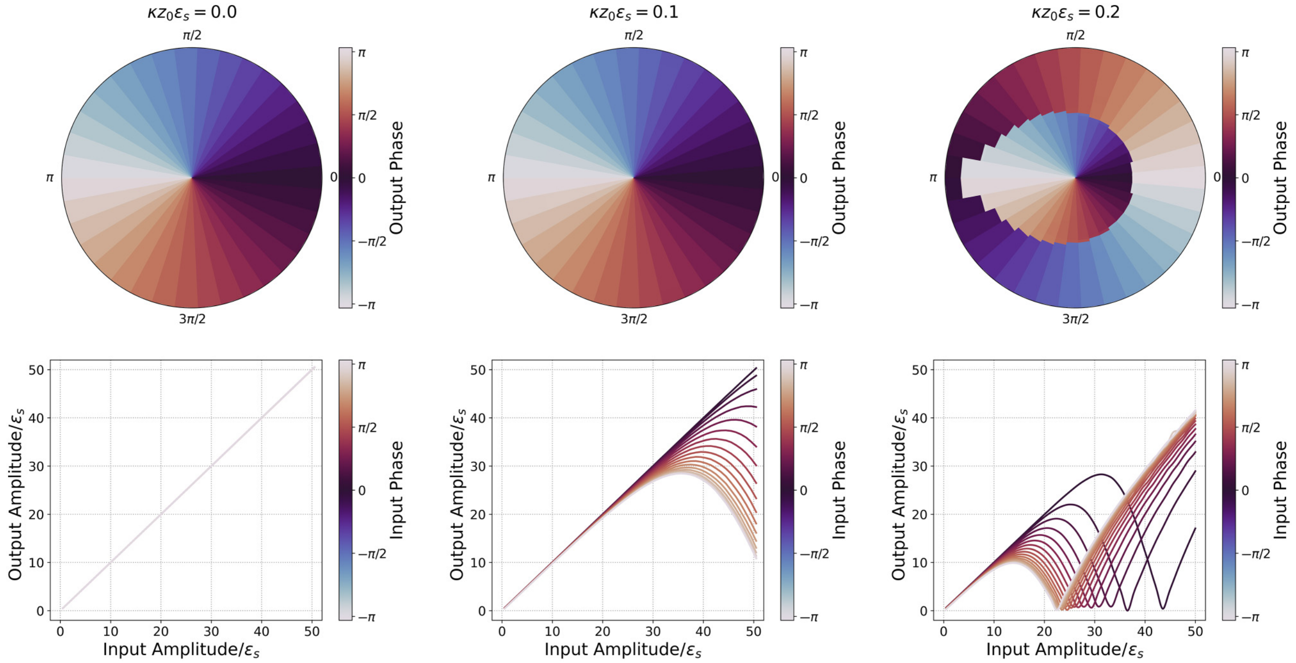

FIG. 5. The “neural” activation function realized in our design. The top row of polar plots gives the phase of a neural mode post-activation function (in color) versus the phases of an input mode (polar coordinate) and its amplitude (radial coordinate). We plot the nonlinearity for three different values of the dimensionless parameter $\kappa z_0 \epsilon_s \in \{0.0, 0.1, 0.2\}$. The bottom row presents the output amplitudes (vertical axis) versus the input amplitudes (horizontal axis), scaled to the fixed amplitude of the pump pulses ϵ_s. In the absence of a nonlinear interaction, i.e., $\kappa z_0 \epsilon_s = 0.0$, we see a linear activation function. The nonlinearity of the activation function becomes more pronounced as the rate of optical nonlinear interactions increases.

$$\frac{\partial E_{\mathrm{sub}}}{\partial z} + \frac{\eta}{c}\frac{\partial E_{\mathrm{sub}}}{\partial t} = \kappa E_{\mathrm{n}} E^{*}_{\mathrm{sub}} - \alpha E_{\mathrm{sub}}, \tag{11}$$

$$\kappa = \frac{\omega}{c}\chi^{(2)} s, \tag{12}$$

where s is a unitless measure of the mode overlap between the neural and subharmonic modes, ω is the frequency of the neural mode, and α is the waveguide loss. For specificity, we consider Gaussian wave packets for the input neural modes (released from the ring that has been performing the matrix multiplication of the previous layer) of the form $E_{\mathrm{n}} = \epsilon_{\mathrm{n}} e^{-(((-(z/c)+t-t_0)^2)/2w^2)} e^{-i\varphi_0}$, where w is the temporal length of the packet, φ_0 is the relative phase of the neuron activation to the subharmonic field, and ϵ_{n} gives the field amplitude scale. Each neuron can have a different phase φ_0, depending on the four-wave-mixing process. Similarly, for the subharmonic pump, we set $E_{\mathrm{sub}} = \epsilon_{\mathrm{s}} e^{-(((-(z/c)+t-t_0)^2)/2w^2)}$; however, an equally valid option would be a continuous wave $E_{\mathrm{sub}} = \epsilon_{\mathrm{s}}$. The subharmonics are assumed to be phase matched, and maintain a constant phase relative to each other. In an experimental setting, the subharmonic will have to be phase locked to a field centered at ω only once as a calibration step. We solve for the evolution of $E_{\mathrm{n}}(z, t)$ numerically. The dimensionless parameters that emerge as chiefly governing these dynamics are the effective strength of the nonlinear interaction $\kappa \epsilon_s z_0$ and the strength of the neural mode relative to the fixed subharmonic mode, $\epsilon_{\mathrm{n}}/\epsilon_{\mathrm{s}}$. The length of the $\chi^{(2)}$ waveguide is denoted z_0. The simulations and the necessary algebraic manipulations are detailed in the interactive archived simulation code [84].

The distorted neuron envelopes are then actively captured into the next ring via a controllable ring-waveguide coupling $\gamma(t)$. The dynamics of the capture without interactions from the pump modes is governed by [55,79,85,86]

$$\frac{\mathrm{d}A}{\mathrm{d}t} = -\frac{(\gamma(t) + \gamma_{\mathrm{H}})}{2} A + \sqrt{\gamma(t)} S_{\mathrm{in}}, \tag{13}$$

$$S_{\mathrm{out}} = S_{\mathrm{in}} + \sqrt{\gamma(t)} A, \tag{14}$$

where $S_{\mathrm{in}}(t) = E_{\mathrm{n}}(0, t)$ is the incoming neural mode’s envelope, S_{out} is the outgoing (not captured) signal, and A is the neuron mode amplitude captured in the resonator. By fixing $S_{\mathrm{out}} = 0$ we can solve for the $\gamma(t)$ that would completely capture a given envelope S_{in}. The analytical solution for a Gaussian wave packet is given in the archived simulation code. However, high neural activations would lead to strong envelope distortions, which in turn cause the mode to not be fully captured, thus providing for the equivalent of a nonlinear elementwise activation function in our neural network architecture. Importantly, this implementation naturally supports negative and complex activations, unlike the vast majority of optical approaches.

From the numerical experiments we see that $\kappa\epsilon_s z_0 \approx 0.2$ provides for a saturating activation function. For a waveguide of length $z_0 = 1$ cm, with good mode overlap $s \approx 1$, in lithium niobate with $\chi^{(2)} = 31$ (pm/V) we obtain $\epsilon_s = 160$ (kV/m). Such a field strength amplitude corresponds to a peak power of approximately $\varepsilon_0\sqrt{\eta}c\epsilon_s^2 a = 20$ μW for a waveguide with a cross section of $a = 0.2$ μm^2. Such pump powers are easy to achieve and should pose no problem for the realization of our device. This device could be constructed using a material with only a large $\chi^{(2)}$ coefficient. As mentioned above, this would require the linear operations to be performed with three-wave mixing which poses difficulties in maintaining the resonance and phase-matching conditions. Alternatively, the whole device could be realized using a single material with a large $\chi^{(3)}$ coefficient. The nonlinear activation in this case will have to be performed using four-wave mixing. Since the strength of the $\chi^{(3)}$ coefficient is generally lower than the $\chi^{(2)}$ coefficient, significantly higher pump powers will be required to realize a similar distortion in the temporal waveform of the neurons. Depending on the platform, especially if one wants to avoid heterogenous integration, other materials with a high $\chi^{(2)}$ such as gallium arsenide [87], aluminium gallium arsenide [88], and silicon carbide [89] can be used. Such materials with both high $\chi^{(2)}$ and $\chi^{(3)}$ coefficients would allow the entire device to be integrated into a single material platform with appropriately rescaled pump powers. [90,91]. Methods to ensure good coupling between heterogeneously integrated materials require adiabatically transitioning the optical mode from one platform to another. Experimentally, silicon nitride and lithium niobate waveguides have been coupled using a terracelike structure, where coupling losses of the order of 0.8 dB have been reported [92]. Alternatively, high-efficiency coupling among different material platforms has also been achieved using tapered waveguides. Experimentally, tapered waveguides coupled to optical fibers have also achieved high coupling efficiencies of approximately 97% [93]. With the appropriate inverse design, simulations suggest that materials with a large index contrast can be coupled with efficiencies >99% [94]. Coupling efficiencies on that order would suggest that networks that are approximately 35 layers deep would lose approximately 50% of the input power solely because of transmission from one material to another.

The nonlinear activation function can also be used to circumvent losses experienced in the ring resonators during the four-wave-mixing process. Particularly, by using the second-harmonic pumps instead of the subharmonic, the nonlinear interactions can be engineered to provide an activation function with a slope greater than 1. In this case, the energy lost during the FWM process can be compensated for, by transferring energy from the second-harmonic pumps into the neurons. An example of such an activation function is discussed in Appendix D.

IV. CASE STUDY: IMAGE CLASSIFICATION

We benchmark the performance of the proposed hardware designs in a simulated neural network for image classification. Each layer of the network is implemented through the method of active coupling, followed by a nonlinear activation function discussed in the previous section. The results of the simulations performed are provided for the MNIST [52] classification task. The preprocessing of these images involved low-pass filtering with a window size $N = 8$ and is identical to the procedures followed in Ref. [95]. The network we train consists of two layers—the first of size 64 (as a result of the chosen window size, recast into a vector), with a variable number of timesteps and the second layer of size 10 (corresponding to the 10 output classes). These vectors are encoded into the initial complex amplitudes of the modes of the simulated microring resonator, i.e., the input layer of the neural network. The digital differentiable model was trained with the standard Adam [96] optimizer and mean-squared error loss function.

In our simulations, we test the performance of the network in different loss regimes while varying the number of timesteps, i.e., the piecewise constant steps of the pumps. In previous sections we observed that expressivity initially grows with the number of sublayers, until losses due to the prolonged operations become detrimental. This observation is confirmed in Fig. 6 where we show the classification accuracy of our model versus the various hardware parameters: some minimum number of sublayers is required to reach sufficiently good accuracy, after which accuracy degrades due to losses. We find that under ideal conditions, where $\Gamma = 0$, the network reaches approximately 90% classification accuracy after just 3 pump steps, and further increases to a peak accuracy of approximately 95%. This performance is similar to that achieved by other optical neural networks on the MNIST dataset. Since a fully expressive network has $\mathcal{O}(N^2)$ degrees of freedom that can be trained, this architecture is also expected to perform similar to other optical neural network architectures on other datasets. Increasing the size of the network is expected to further increase the performance of the network [97,98]. On the other hand, in the more practical case using a state-of-the-art cavity with $\Gamma = 0.2$ ns^{-1}, the peak accuracy of approximately 90% is reached after about five instances of FWM, after which the performance of the network drops. As the cavities get increasingly lossy, we see a similar trend—an increase in the performance of the network for a small number of pump steps, followed by a precipitous drop in the accuracy. As discussed in

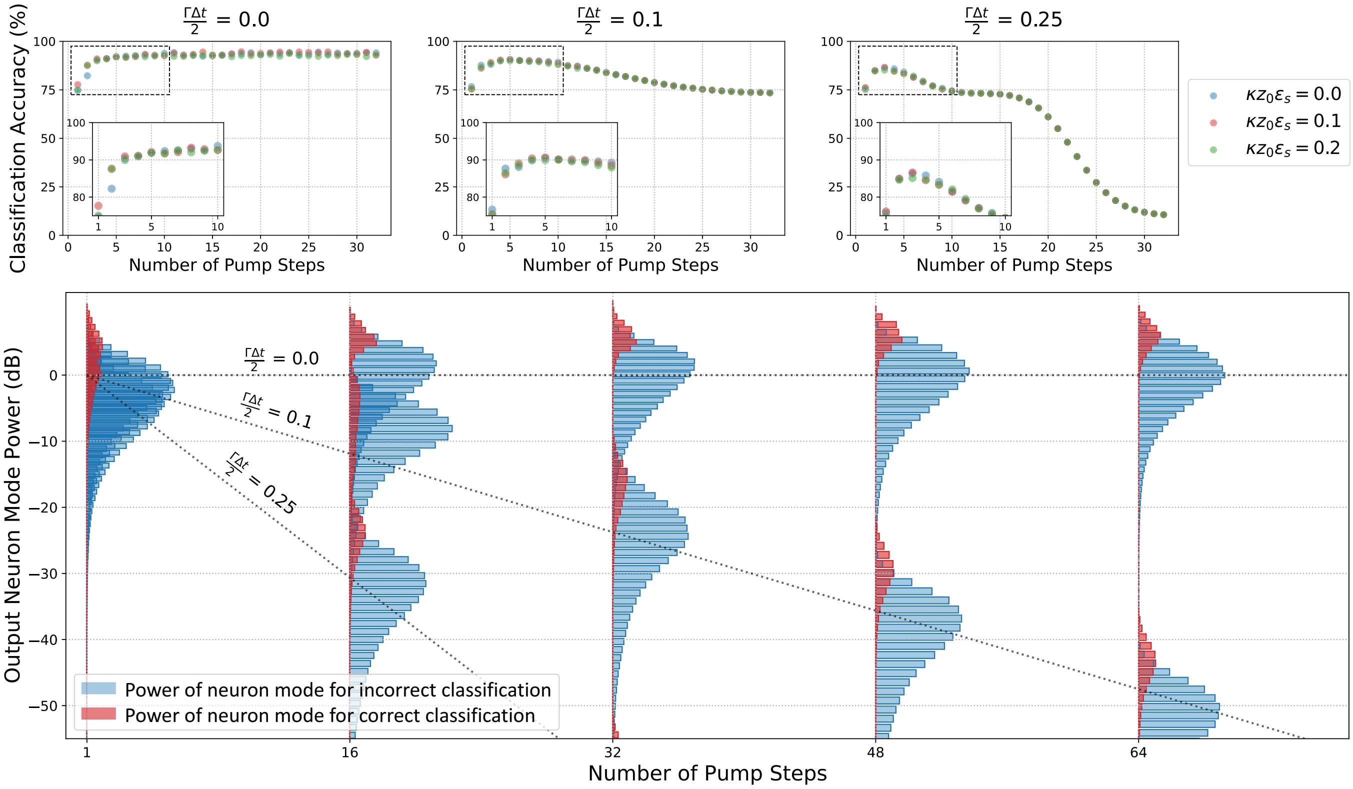

FIG. 6. The classification performance of an all-optical neural net against the MNIST dataset depending on optical losses, effective waveguide nonlinearity, and network size. We use a varying number of 64-neuron sublayers (as depicted on the horizontal axis) followed by ten 10-neuron layers. In the top row we have classification accuracy (vertical axis) versus the number of sublayers, i.e., distinct piecewise constant steps in the control pumps (horizontal axis). The three top facets depict different decay rates Γ, e.g., $(\Gamma\Delta t/2) = 0.25$ corresponds to $\Gamma = 0.5\ \text{ns}^{-1}$ for a step duration of $\Delta t = 1$ ns. The insets in each of these facets enlarge the trend of the classification accuracy for the first 10 pump steps. The strength of the nonlinear interaction in the waveguides between ring-resonators is depicted in the color of the marker. While, initially, increasing the number of sublayers improves the performance thanks to the higher expressivity of the encoded operation, a further increase is detrimental as it causes nonunitary behavior and a decrease in expressivity. In the third facet one can additionally observe the precipitous drop in performance when, due to the increasing losses, the shot noise starts dominating the measurement result. The bottom plot shows histograms of the power carried by the output neural modes. For various number of sublayers (horizontal axis) and various loss rates (annotated with dashed lines), we plot the distribution of energy per neuron mode (histograms with respect to the vertical axis). The blue histograms correspond to the “incorrect class” neurons, while the red are the “correct class” neurons (which are fewer). The energy carried by the “correct class” neurons is consistently higher, indicating effective classification. Moreover, at high expressivity the red histogram has noticeably smaller spread. The zero decibel reference corresponds to a maximum of 10^6 photons per mode.

previous sections, the presence of loss reduces the expressivity of the network. On other datasets, the accuracy achieved by the network depends on the expressivity, and whether the optimal solution can be realized by the group of operations that can be expressed by the network. For larger networks in the presence of loss, we expect to see a similar trend as that shown in Fig. 6—increasing performance for a small number of pump steps followed by a faster precipitous drop. To explicitly illustrate this loss regime, we also plot histograms of the energy carried by each neuron mode at the output of the neural network. Unsurprisingly, the energy of the neurons decays exponentially as we increase the number of sublayers. Even before measurements become shot-noise limited, the performance of the network drops. For a state-of-the-art cavity ($\Gamma = 0.2\ \text{ns}^{-1}$) and control pulse resolution of $\Delta t = 1$ ns, we obtain excellent classification performance and less than 5 dB of loss. However, for larger networks, the pumping schemes discussed in previous sections would be crucial for the reliable performance of the system.

V. DISCUSSION AND CONCLUSION

This work presents a novel architecture for an all-optical coherent artificial-neural-network processor that relies solely on coherent nonlinear optical processes. The proposed scheme encodes information in the complex amplitudes of frequency states and are modulated via four-wave mixing in a $\chi^{(3)}$ medium, that enables the process of matrix multiplication. This scheme can be realized experimentally on a chip requiring only microring resonators.

The proposed neural network processor has multiple advantages over previous implementations of optical and electronics neural networks. As opposed to digital matrix-vector multiplication that typically take $\mathcal{O}(N^2)$ timesteps (where N is the size of the vector), the proposed model has a time complexity of only $\mathcal{O}(N)$. An added benefit is that the speed of a single linear operation is a constant that depends on the parameter Δt and not on the number of neurons. This arises as a result of the parallel nature of the FWM process. The number of on-chip components is also very low, as all neural modes occupy the same resonator. Another feature of the architecture we propose is that the speed of the operations is directly proportional to the power of the pumps, letting us freely increase the computational speed, to limits imposed by heating and leakage from the hardware. Since a practical device would be trained to accelerate inference on a specific dataset requiring only one set of weights, the same sequence of pumps can be recirculated, significantly lowering power requirements, up to the need for amplification to guard against optical losses.

A particularly exciting benefit of the processor design is the fully reversible (unitary) dynamics realised by our accelerator. This opens up many future avenues for ultrafast, low-energy computing and the possibility of *in situ* on-chip training [99]. One application is the possibility of executing Hamiltonian-echo backpropagation [50], which is an extremely efficient form of analog gradient descent. Successfully realizing this scheme would result in self-learning devices that converge to the ground state of a given Hamiltonian. Other types of computational accelerators like reservoir computers [27–30,53] and Ising [100] machines could also be studied on this hardware. Even more crucially, given the extreme similarities in hardware requirements, this computational architecture can be used as near-term proving grounds for technologies that would enable room-temperature quantum computation: in Ref. [51] a multimode cavity is used to encode information in the occupation numbers of each frequency mode; programmable unitary operations including entangling gates can be executed by a three-wave-mixing process with an optimally controlled pump pulse. Moreover, to implement universal quantum computing and measurement free error correction, the only hardware required by Ref. [51] is a pair of multimode resonators. The architecture proposed in this paper encodes information in the complex amplitudes of frequency modes of a resonator. Programmable linear transformations (as well as the nonlinear activation) are performed by time-dependant four-wave mixing. Finally, to implement the entire neural network processor, we need only a set of cascaded multimode resonators. These similarities suggest that the proposed design is a potentially less demanding version of a future room-temperature, nonlinear optics-based quantum computational device.

Experimentally realizing such all-optical hardware would undoubtedly be challenging. However, our estimates for hardware parameters such as pump powers, quality factors, and effective nonlinear susceptibility lie within the range of what has been achieved experimentally. The high degree of control required gives rise to fabrication and system integration-based complications. The generation of the pumps will also factor into the power budget and space requirements of the device. Off-the-shelf multichannel electro-optic or acousto-optic modulators have been used to generate such time-dependent signals. Typically, the V_π rating of these modulators is of the order of several volts, and the driving current required is several milliamperes. This places the power requirements to generate a single pump in the range of 1–10 mW [101,102]. A full system with hundreds of pumps of the order of 1 W, which is orders of magnitude lower than traditional digital electronic architectures such as graphical processing units. These challenges in realizing the proposed architecture are, however, not fundamental roadblocks. With the prospects of self-learning machines, reversible computing, simulations in synthetic dimensions, and room-temperature quantum computing that can be explored with such a device, investments to experimentally realize such a device would enable returns in a multitude of computational domains.

APPENDIX A: METHOD OF PASSIVE COUPLING FOR PROGRAMMABLE LINEAR TRANSFORMATIONS

The main text of this paper covers the method of *active* coupling. Using this scheme, it is possible to execute a single linear layer of the neural network using a single microring resonator. However, it requires active control of the coupling coefficient $\gamma(t)$, and places stringent control and fabrication requirements on the hardware. Similar dynamics can be realized using an alternative scheme, which we term as the method of *passive* coupling. In this method, the coupling between the resonator and the waveguide is constant γ, with the neurons and pumps being allowed to propagate past the resonators. Following absorption into the microring resonators, the neuron and pump modes interact via FWM. We consider similar conditions for the FWM as done in the case of active coupling, i.e., the pumps are much stronger than the neurons, are nondepletive, and obey the energy matching conditions. To understand the time dynamics of the neuron modes, we consider four modes—two neural modes and two pumps modes. The Hamiltonian of the interaction is given as $\hat{H} = \hbar\chi\left(\hat{p}_1\hat{p}_2^\dagger\hat{a}_1\hat{a}_2^\dagger\right) + \text{H.c.}$. Using coupled mode theory, the Heisenberg equations of motion can be written as

$$\dot{P}_i = 0 = -\frac{\Gamma}{2}P_i - \sqrt{\gamma}S_{\text{in},P_i} \tag{A1}$$

$$\frac{\mathrm{d}A_1}{\mathrm{d}t} = \left(-\frac{\Gamma}{2} + i\chi|P_1|^2 + i\chi|P_2|^2\right)A_1 + \left(\chi P_1P_2^*\right)A_2 - \sqrt{\gamma}S_{\text{in},1}, \tag{A2}$$

$$\frac{\mathrm{d}A_2}{\mathrm{d}t} = \left(-\frac{\Gamma}{2} + i\chi|P_1|^2 + i\chi|P_2|^2\right)A_2 - \left(\chi P_1^*P_2\right)A_1 - \sqrt{\gamma}S_{\text{in},2}, \tag{A3}$$

$$S_{\text{out},i} = S_{\text{in},i} + \sqrt{\gamma}A_i, \tag{A4}$$

where just as in the main text, A_i and P_i are the amplitudes of the neurons and pumps inside the resonator. The encoded data is introduced into the system via the input waveguide mode, denoted by $S_{\text{in},i}$ and the output neuron modes after interacting with the resonators are denoted by $S_{\text{out},i}$.

This formalism can be extended to N neurons, and just as in the case of active coupling, pumps, and neurons with the same frequency difference can give rise to cross-coupling effects. The updated coupled amplitude equation for the neuron modes is now

$$\frac{\mathrm{d}A_i}{\mathrm{d}t} = \left(-\frac{\Gamma}{2} + i\chi\sum_{m=1}^{N}|P_m|^2\right)A_i - \chi\left[\sum_{j>i}^{N}\sum_{k=1}^{j-1}\left(P_kP^*_{k+j-i}\right)A_j - \sum_{j<i}^{i-1}\sum_{k=1}^{j}\left(P^*_kP_{k+i-j}\right)A_j\right] - \sqrt{\gamma}S_{\text{in},i}. \tag{A5}$$

If we make the assumption that the first pump P_1 is much stronger than the other pumps, the cross-coupling terms can be neglected, leading to

$$\frac{\mathrm{d}A_i}{\mathrm{d}t} = \left(-\frac{\Gamma}{2} + i\chi|P_1|^2\right)A_i - \chi\left[\sum_{j>i}^{N}\left(P_1P_j^*\right)A_j - \sum_{j<i}^{i-1}\left(P_1^*P_j\right)A_j\right] - \sqrt{\gamma}S_{\text{in},i}, \tag{A6}$$

which together with Eq. (A4) lets us rewrite the system of coupled mode equations in a matrix form:

$$\vec{S}_{\text{out}} = \vec{S}_{\text{in}} + \sqrt{\gamma}\left[\mathbf{P}^{-1}\left(\dot{\vec{A}} + \sqrt{\gamma}\vec{S}_{\text{in}}\right)\right]. \tag{A7}$$

Here the matrix **P** is the same Toeplitz matrix of pump amplitudes defined by Eq. (4). A similar procedure can be followed to derive the coupled amplitude Eq. (4) in the case of active coupling. A deep neural network would typically consist of several layers, which in our case would be implemented by cascading multiple microring resonators consecutively. To enable repeated application of such a transformation, we need to ensure that the temporal envelope of the pulse does not vary significantly as it undergoes transformations through FWM. Assuming the S_{in} pulses to have a Gaussian temporal envelope, we can preserve the Gaussian shape of the output pulses S_{out} if the pulses are much longer than $1/\gamma$. For pulses with a large enough duration, we can make the adiabatic elimination $\dot{\vec{A}} = 0$, allowing us to work in the steady-state regime. We illustrate this approximation in Fig. 7 by comparing the solution of the steady-state model with the solution of the full dynamics. As the length of the input pulses increases, the steady-state model begins to closely resemble the model of the full dynamics. This approximation allows us to simplify Eq. (A7) into $\vec{S}_{\text{out}} = \left(\mathbb{I}_N + \gamma\mathbf{P}^{-1}\right)\vec{S}_{\text{in}} = \mathbf{T}\vec{S}_{\text{in}}$, where $\mathbb{I}_N$ is the N-dimensional identity matrix.

In this case as well, we see that the linear transformation matrix **T** has only N degrees of freedom, as opposed to N^2

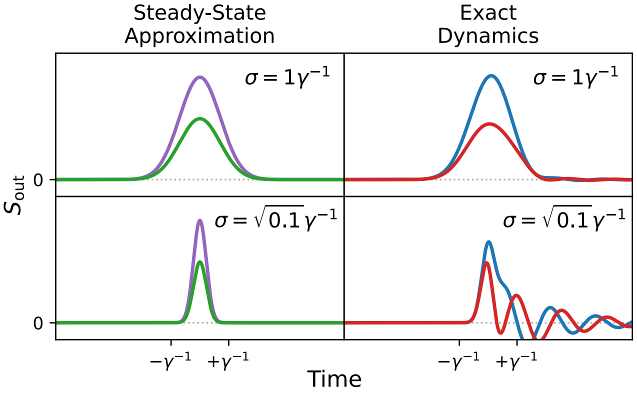

FIG. 7. Comparison of the steady-state model and the full model for the pulse with the Gaussian envelope of different durations. The right column illustrates the correct profile of S_{out}, while the left column show the profile as predicted by a steady-state model. For pulses much shorter than $1/\gamma$, we see the breakdown of the steady-state model.

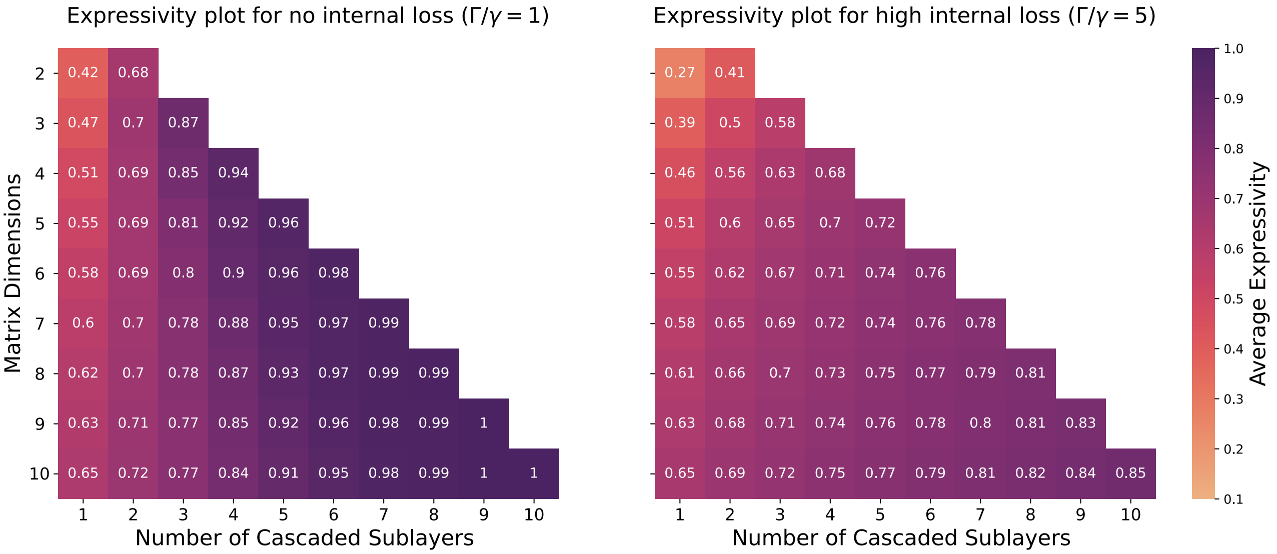

FIG. 8. Quantifying the expressivity of the transformation $\prod \mathbf{T}$ in different parameter regimes of the *passive coupling scheme*. Each plot displays the average expressivity as we vary the number of sublayers (the horizontal axis) for a given matrix dimension (the vertical axis). On the left, the expressivity at no internal loss ($\Gamma/\gamma = 1$) reaches unity at sufficiently many sublayers. On the right, the expressivity at high loss ($\Gamma/\gamma = 5$) is consistently lower.

degrees of freedom that the weights of a fully connected deep neural network would have. Just as in the case of active capture, a single instance of FWM would be limited in its expressivity. To solve this problem, we introduce the concept of sublayers, i.e., a layer would be implemented by several noncommuting cascaded matrices of the form $\mathbf{T}$. Each sublayer can be understood as performing the same function as a single timestep in the case of active coupling, i.e., an instance of FWM. Physically, this would require multiple subsequent ring resonators, one per sublayer. By cascading multiple sublayers, the number of free parameters increases and, therefore, would be expected to span larger groups of unitary operations. To quantify the group of operations that can be spanned by matrices of the form $\mathbf{T}$, we perform expressivity calculations as done in the case of active capture, with the expressivity measure defined by Eq. (5). Figure 8 illustrates that in the case where there is not internal loss ($\Gamma/\gamma = 1$, i.e., $\gamma_H = 0$), upon cascading a sufficient number of sublayers, we can span the entire group of unitary matrices. In the more pessimistic case of high internal loss ($\Gamma/\gamma = 5$, i.e., $\gamma_H = 4\gamma$), the expressivity saturates under 1, thus spanning only a fraction of the unitary group.

This method of passive capture shares the same architectural benefits as the method of active capture—the speed of the computation scales with the power of the pumps, and the FWM across neuron modes is still parallel. The lower bound of computational speed applies in this case as well, because the pumps need to be strong enough for the FWM to occur before the neurons leak out of the resonators. Interestingly, too large of a Q factor would also pose limitations in the pulse-capture efficiency, due to the resonator's inherent decoupling from the environment. This is not an issue in the case of active capture due to the tunable coupling coefficient.

APPENDIX B: EXPRESSIVITY OF THE REALIZED TRANSFORMATIONS

In Secs. II A and II B we studied the expressivity of the linear transformations realized in our proposed hardware. We do this by randomly sampling M unitary matrices, each

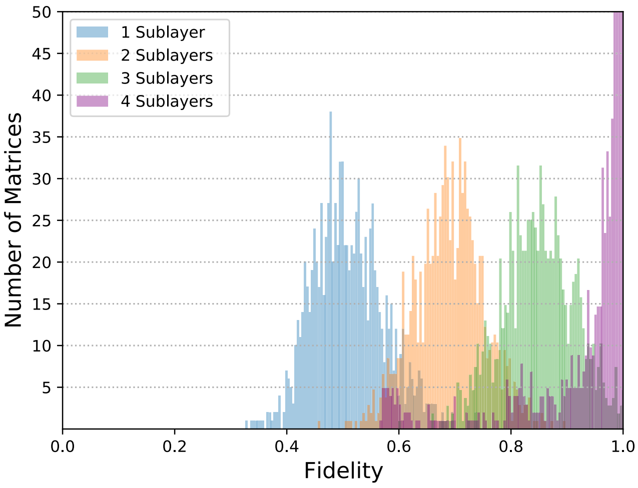

FIG. 9. Distribution of the fidelity of the optimized "passive coupling" transformations depending on the number of sublayers utilized.

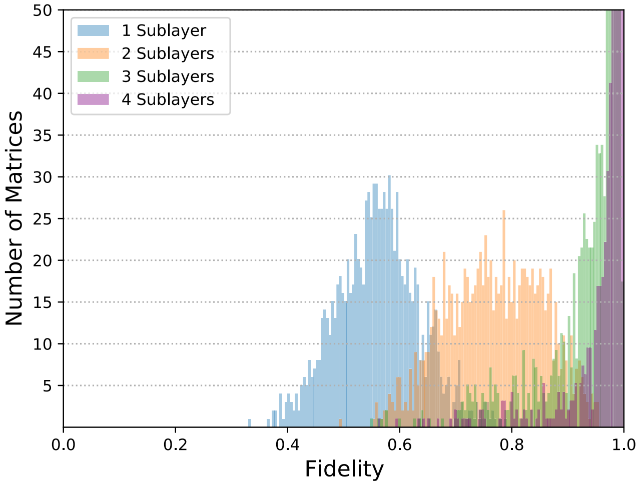

FIG. 10. Distribution of the fidelity of the optimized "active coupling" transformations depending on the number of sublayers utilized.

denoted by U_i, and for each one of them attempting to realize it in our hardware. The average overlap between target and realization, also known as fidelity, also known as expressivity is defined by Eq. (5). Here we present more detailed statistics over the sample of $M = 1000$ matrices, by giving a histogram of the single sample (single U_i) fidelities, instead of just their averages seen in Figs. 8 and 2.

For the case of passive coupling, transformations are given by matrices of the form $\prod \mathbf{T}$. Figure 9 shows the distribution of the number of matrices with the average fidelities increasing as the number of sublayers are increased. We specifically chose small 4 × 4 matrices, as the behavior is easier to depict at that scale. Similar results can be seen for transformations of the type $\prod e^{\Delta t \mathbf{P}}$ realised by the active coupling method, as seen in Fig. 10.

APPENDIX C: DETERMINISTIC ACTIVE CAPTURE OF A PULSE INTO A RESONATOR

The dynamics of a incoming pulse S_{in} being captured into a resonator mode A is described by Eqs. (13) and (14), where S_{out} is the outgoing pulse envelope. To ensure the entirety of the pulse is captured, we can rearrange the equations as

$$S_{\mathrm{out}} = 0 \implies \sqrt{\gamma(t)} = \frac{S_{\mathrm{in}}}{A}, \tag{C1}$$

$$\frac{\mathrm{d}A}{\mathrm{d}t} = -\frac{\gamma_{\mathrm{H}}}{2}A + \frac{S_i^2}{2A}. \tag{C2}$$

We can solve this differential equation for an arbitrary incoming envelope. The analytical solution for an incoming pulse with a Gaussian envelope $S_{\mathrm{in}} = S_0 e^{-(((t-t_0)^2)/2w^2)}$ is $\sqrt{\gamma(t)} = ((S_0 e^{-(((t-t_0)^2)/2w^2)})/A)$, where

$$A(t) = S_0 \left(\sqrt{\pi} w e^{\frac{\gamma_{\mathrm{H}}}{2}\left(\frac{\gamma_{\mathrm{H}} w^2}{2} + 2t_0 - 2t\right)} \times \frac{\left(1 + \operatorname{erf}\left(-\frac{\gamma_{\mathrm{H}} w^2}{2} - t_0 + t\right)\right)}{2} \right)^{\frac{1}{2}}. \tag{C3}$$

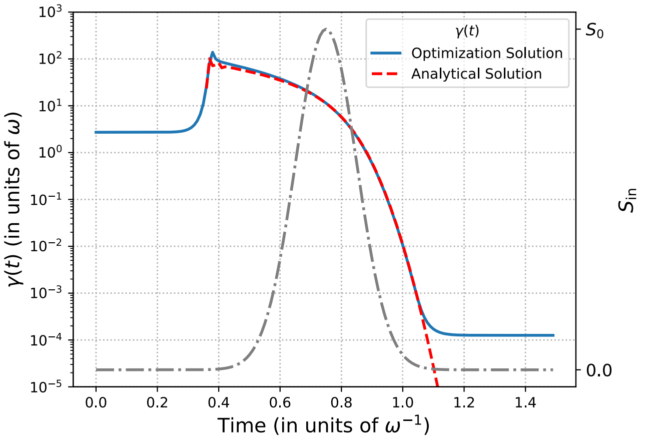

FIG. 11. The controllable coupling coefficient $\gamma(t)$ for the capture of an incoming pulse of a Gaussian envelope (in dashed grey line, on the right vertical axis).

For other envelopes, a numerical solution, either through solving the differential equation, or through an optimization problem minimizing S_{out} is also possible. Figure 11 illustrates the agreement between the analytical solution for $\gamma(t)$ in Eq. (C3) and that obtained via a generic numerical optimization.

APPENDIX D: NONLINEAR ACTIVATION FUNCTION WITH SECOND-HARMONIC PUMP USED FOR AMPLIFICATION

The proposed nonlinear activation function in the main text uses a subharmonic mode that interacts with the neuron modes. This subharmonic mode operates at frequencies that are at half of the frequencies of the neuron modes. However, constraints such as the transparency of the material and the availability of high-efficiency sources at required frequencies could present experimental difficulties. Hence, we discuss an alternative nonlinear activation function, where we allow the neuron modes to interact with pumps that are at the second harmonic. Similarly to the main text, the interaction can be modeled by the same system of partial differential equations, in which we permute the neuron and pump modes:

$$\frac{\partial E_{\mathrm{sec}}}{\partial z} + \frac{\eta}{c}\frac{\partial E_{\mathrm{sec}}}{\partial t} = -\kappa E_{\mathrm{n}}^2 - \alpha E_{\mathrm{sec}}, \tag{D1}$$

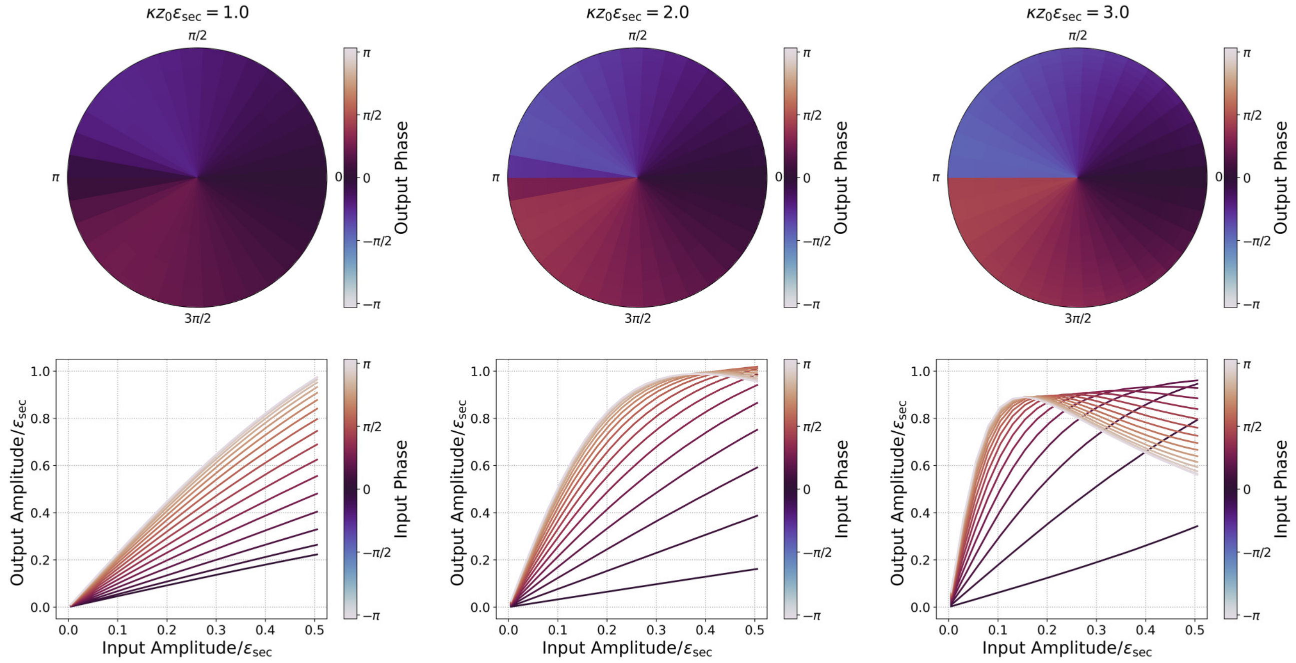

FIG. 12. The alternative nonlinear activation function based on a nonlinear interaction with a pump at the second harmonic. This activation function provides reliable (nonlinear) amplification over the majority of input phases and amplitudes, circumventing losses that might have been experienced elsewhere in the circuit.

$$\frac{\partial E_{\rm n}}{\partial z} + \frac{\eta}{c}\frac{\partial E_{\rm n}}{\partial t} = \kappa E_{\rm sec} E_{\rm n}^* - \alpha E_{\rm n}. \tag{D2}$$

Here, $E_{\rm n}$ is the neuron mode and $E_{\rm sec}$ is the second-harmonic pump mode.

A set of numerical solutions can be seen in Fig. 12. These nonlinear activation functions can be used in order to amplify the neuron modes and circumvent losses.

APPENDIX E: EQUATIONS OF MOTION IN A WAVEGUIDE WITH THREE-WAVE MIXING

Here we explicitly derive the equations of motion used in the main text. Consider the “neuron” and “subharmonic” fields:

$$\mathbf{E}_n = \mathbf{f}_n^p(x,y) E_n(z,t) e^{i(\omega t - kz)} + \text{c.c.}, \tag{E1}$$

$$\mathbf{E}_s = \mathbf{f}_s^p(x,y) E_s(z,t) e^{i\frac{1}{2}(\omega t - kz)} + \text{c.c.}. \tag{E2}$$

Here $\mathbf{f}_*^p(x,y)$ describes the profile of the waveguide mode and $\mathbf{E}_*(z,t)$ describes the shape of the wave packet. Of note is that we keep track of the complex conjugate part as we have nonlinear processes. Given Maxwell’s equations in matter and the typical parameterization of nonlinear susceptibility we have the nonlinear wave equation

$$\left(\nabla^2 - \frac{n^2}{c^2}\right)\mathbf{E} = \frac{1}{\varepsilon_0 c^2}\partial_t^2 \mathbf{P}_{\rm NL} = \frac{1}{c^2}\chi^{(2)}\mathbf{E}\mathbf{E}, \tag{E3}$$

where $\mathbf{E} = \mathbf{E}_n + \mathbf{E}_s$ and we have approximated $\nabla \cdot \mathbf{E} = 0$.

The linear version of the above equation provides an eigenvalue problem defining the shape of the waveguide modes:

$$\left(\partial_x^2 + \partial_y^2\right)\mathbf{f}_n^p(x,y) = \left(-(ik)^2 + (i\omega)^2\frac{n^2}{c^2}\right)\mathbf{f}_n^p(x,y), \tag{E4}$$

$$\left(\partial_x^2 + \partial_y^2\right)\mathbf{f}_s^p(x,y) = \left(-\left(i\frac{k}{2}\right)^2 + \left(i\frac{\omega}{2}\right)^2\frac{n^2}{c^2}\right)\mathbf{f}_s^p(x,y). \tag{E5}$$

The nonlinear perturbation leads to the following equation of motions for the wave packet envelopes. In its derivation we take into account that they are slowly varying functions for which $\partial_z \ll k$ and $\partial_t \ll \omega$. Therefore, we have the following coupled equations of motion:

$$\left(\partial_z + \frac{n}{c}\partial_t\right)E_n(z,t) = -\kappa E_s^2(z,t), \tag{E6}$$

$$\left(\partial_z + \frac{n}{c}\partial_t\right)E_s(z,t) = \kappa E_n(z,t)E_s^*(z,t), \tag{E7}$$

where κ is

$$\kappa\frac{c}{\omega} = \frac{\int \chi^{(2)}\mathbf{f}_n^{p*}\mathbf{f}_s^p\mathbf{f}_s^p\,\mathrm{d}x\mathrm{d}y}{\int \mathbf{f}_n^{p*}\mathbf{f}_n^p\,\mathrm{d}x\mathrm{d}y} = -\frac{\int \chi^{(2)}\mathbf{f}_s^{p*}\mathbf{f}_n\mathbf{f}_s^{p*}\,\mathrm{d}x\mathrm{d}y}{\int \mathbf{f}_s^{p*}\mathbf{f}_s^p\,\mathrm{d}x\mathrm{d}y}. \tag{E8}$$

These two expressions for κ have to be equal for energy to be conserved in the equations of motion. For mode overlap of the order of unity, we have $\kappa(c/\omega) \approx \chi^{(2)}$.

APPENDIX F: AMPLITUDE SCALES

Throughout the main text we treated the physics of the two components of the neural network architecture independently. One one hand, we have the physics of the matrix multiplication enabled through four-wave mixing. While the control pump powers in that setup were extremely important (they set the time scales for the multiplication operations and they were the main source of heating in the hardware), there were no significant constraints of the powers (or amplitudes) of the neural modes. The neural modes need only be much weaker than the pump powers in order to ensure we can assume quasistatic pumps (and this assumption is not of fundamental importance, rather it is done to simplify the modeling).

On the other hand, we have the physics of the elementwise activation functions realized during the propagation in a waveguide. The amplitude scales of both the neural modes and the control pumps need to be carefully calibrated in order to perform the desired activation function. Realistic values for these scales are presented in the main text.

Lastly, there is the question of deriving the time-dependent coupling coefficient $\gamma(t)$ necessary for the complete capture of traveling pulses into the resonators. As seen in the previous section, that coefficient is independent of the strength of the pulses.

Nonetheless, it would be instructive to know how to convert between the two scales (of stationary modes inside the resonators and of traveling modes in the waveguides). The simplest way to do that is to relate these scales to absolute energy carried by the given mode.

1. Inside the resonators

In the main text it was never necessary to specify an absolute scale for the amplitudes of the neural modes while being inside the FWM resonators. Any such scales canceled out in all dynamical equations as all time scales were governed only by pump powers (thanks to the nondepleting pump assumption we made). We had only specified the pump powers. If necessary, a convenient way to parameterize the neuron modes, i.e., the standing waves inside of the resonators, would be

$$\mathbf{E}_{\text{ring}} = \mathbf{f}_{\text{ring}}(x,y) A \sin(m\theta)\, e^{i\omega t} + \text{c.c}, \tag{F1}$$

where $\mathbf{f}_{\text{ring}}(x,y)$ is a unitless mode profile as given by the mode eigenvalue problem (with some prescribed normalization), and A is the mode amplitude (measured in field strength units). For the given ring resonator, the cross-sectional coordinates are x and y, the polar angle coordinate is θ, and m is the order of the standing wave. The corresponding mode energy U_{ring} or average number of photons ($U_{\text{ring}}/\hbar\omega$) can be found by integrating the energy density:

$$U_{\text{ring}} = \varepsilon_{\text{ring}} |A|^2 2\pi R \int \mathbf{f}_{\text{ring}} \text{d}x\text{d}y, \tag{F2}$$

where R is the radius of the ring, and the integral is over the cross section of the ring, and $\varepsilon_{\text{ring}}$ is the ring permittivity.

2. Free propagation in the waveguide

In the main text and the rest of this appendix we have used the following form for a propagating Gaussian pulse in the waveguide:

$$\mathbf{E}_{\text{wg}} = \mathbf{f}_{\text{wg}}(x,y) E(z,t) e^{i(\omega t - kz)} + \text{c.c.}. \tag{F3}$$

Here $E = \epsilon e^{-(((-(z/c)+t)^2)/2w^2)}$ is the wave packet envelope, $\mathbf{f}_{\text{wg}}$ is a unitless mode profile (with some prescribed normalization), z is the coordinate along the length of the waveguide, w specifies the duration of the packet, and ϵ (measured in units of field strength) is the amplitude by which we parameterize the propagating modes. Integrating the energy density gives us the energy of such a pulse:

$$U_{\text{wg}} = \varepsilon_{\text{wg}} |\epsilon|^2 c w \sqrt{\pi} \int \mathbf{f}_{\text{wg}} \text{d}x\text{d}y. \tag{F4}$$

Here ε_{wg} is the permittivity of the waveguide material and c is the speed of light in that material.

These two expressions, together with conservation of energy, and the previously presented pulse-capture dynamics permit us to connect the parameterizations of all stages of our architecture.

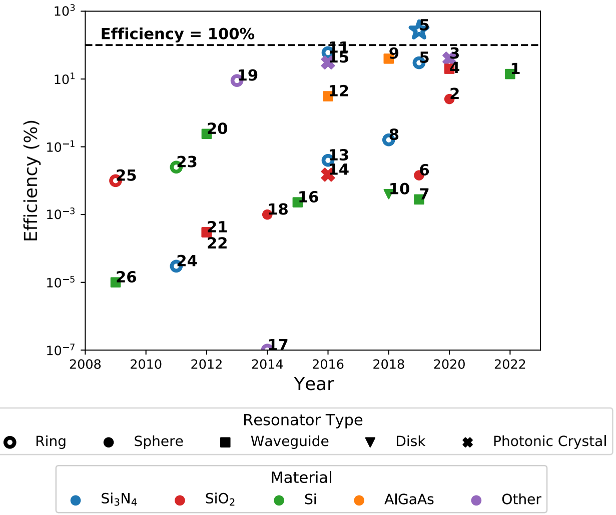

FIG. 13. Recent progress in third-harmonic generation and four-wave mixing experiments. Numerical labels are linked to the corresponding publications listed in the table below.

	Year	Publication
1	2022	High-efficiency four-wave mixing in low-loss silicon photonic spiral waveguides beyond the singlemode regime
2	2020	Third-harmonic generation enhancement in an ITO nanoparticle-coated microresonator
3	2020	High efficiency cascaded third-harmonic generation in a quasi-periodically poled $KTiOPO_4$ crystal
4	2020	Coherently enhanced third-harmonic generation in cascaded microfibers
5	2019	Efficient telecom-to-visible spectral translation through ultralow power nonlinear nanophotonics
6	2019	Microcavity nonlinear optics with an organically functionalized surface
7	2019	Efficient, broadband third-harmonic generation in silicon nanophotonic waveguides spectrally shaped by nonlinear propagation
8	2018	Efficient third-harmonic generation in composite aluminum nitride/silicon nitride microrings
9	2018	Travelling-wave resonant four-wave mixing breaks the limits of cavity-enhanced all-optical wavelength conversion
10	2018	Boosting third-harmonic generation by a mirror-enhanced anapole resonator
11	2016	Efficient and low-noise single-photon-level frequency conversion interfaces using silicon nanophotonics
12	2016	Low-power continuous-wave four-wave mixing wavelength conversion in AlGaAs-nanowaveguide microresonators
13	2016	Frequency comb generation in the green using silicon nitride microresonators
14	2016	Phase-matched third-harmonic generation via doubly resonant optical surface modes in 1D photonic crystals
15	2016	Cascaded third-harmonic generation with one KDP crystal
16	2015	Coherent visible-light-generation enhancement in silicon-based nanoplasmonic waveguides via third-harmonic conversion
17	2014	Green, red, and IR frequency comb line generation from single IR pump in AlN microring resonator
18	2014	Optical frequency conversion in silica-whispering-gallery-mode microspherical resonators
19	2013	New CMOS-compatible platforms based on silicon nitride and Hydex for nonlinear optics
20	2012	Bridging the mid-infrared-to-telecom gap with silicon nanophotonic spectral translation
21	2012	Nonlinear microfiber loop resonators for resonantly enhanced third harmonic generation
22	2012	Broadband third harmonic generation in tapered silica fibres
23	2011	Travelling-wave resonant four-wave mixing breaks the limits of cavity-enhanced all-optical wavelength conversion
24	2011	Harmonic generation in silicon nitride ring resonators
25	2009	Low power four wave mixing in an integrated, micro-ring resonator with $Q = 1.2$ million
26	2009	Green light emission in silicon through slow-light enhanced third-harmonic generation in photonic-crystal waveguides

APPENDIX G: EXPERIMENTAL PROGRESS

In Fig. 13 we track recent experimental progress in the creation of high-efficiency FWM hardware.

[1] T. Young, D. Hazarika, S. Poria, and E. Cambria, Recent trends in deep learning based natural language processing [review article], IEEE Comput. Intell. Mag. **13**, 55 (2018).

[2] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, Improved protein structure prediction using potentials from deep learning, Nature **577**, 706 (2020).

[3] J. Chrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver, Mastering Atari, Go, chess and shogi by planning with a learned model, Nature **588**, 604 (2020).

[4] V. Almeida, C. Barrios, and R. E. A. Panepucci, All-optical control of light on a silicon chip, Nature **431**, 1081 (2004).

[5] J. Leuthold, C. Koos, and W. Freude, Nonlinear silicon photonics, Nat. Photonics **4**, 535 (2010).

[6] L. Bernstein, A. Sludds, R. Hamerly, V. Sze, J. Emer, and D. Englund, Freely scalable and reconfigurable optical hardware for deep learning, Sci. Rep. **11**, 1 (2021).

[7] L. Bernstein, A. Sludds, C. Panuski, S. Trajtenberg-Mills, R. Hamerly, and D. Englund, Single-shot optical neural network, Sci. Adv. **9** (25), eadg7904 (2023).

[8] T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon, An optical neural network using less than 1 photon per multiplication, Nat. Commun. **13**, 1 (2022).

[9] T. Wang, M. M. Sohoni, L. G. Wright, M. M. Stein, S.-Y. Ma, T. Onodera, M. G. Anderson, and P. L. McMahon, Image sensing with multilayer, nonlinear optical neural networks, Nat. Photonics **17** (5), 408 (2023).

[10] Z. Chen, A. Sludds, R. Davis, I. Christen, L. Bernstein, L. Ateshian, T. Heuser, N. Heermeier, J. A. Lott, S. Reitzenstein *et al.*, Deep learning with coherent vcsel neural networks, Nat. Photonics **17** (8), 723 (2023).

[11] T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, and Q. Dai, Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit, Nat. Photonics **15**, 367 (2021).

[12] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, All-optical machine learning using diffractive deep neural networks, Science **361**, 1004 (2018).

[13] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, Large-scale optical neural networks based on photoelectric multiplication, Phys. Rev. X **9**, 021032 (2019).

[14] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, and D. Englund, Deep learning with coherent nanophotonic circuits, Nat. Photonics **11**, 441 (2017).

[15] J. R. Basani, S. K. Vadlamani, S. Bandyopadhyay, D. R. Englund, and R. Hamerly, A self-similar sine–cosine fractal architecture for multiport interferometers, Nanophotonics **12**, 975 (2023).

[16] R. Hamerly, S. Bandyopadhyay, and D. Englund, Asymptotically fault-tolerant programmable photonics, Nat. Commun. **13**, 6831 (2022).

[17] D. P. López, Programmable integrated silicon photonics waveguide meshes: Optimized designs and control algorithms, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2019).

[18] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, Optimal design for universal multiport interferometers, Optica **3**, 1460 (2016).

[19] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, Experimental realization of any discrete unitary operator, Phys. Rev. Lett. **73**, 58 (1994).

[20] S. Bandyopadhyay, A. Sludds, S. Krastanov, R. Hamerly, N. Harris, D. Bunandar, M. Streshinsky, M. Hochberg, and D. Englund, Single chip photonic deep neural network with accelerated training, arXiv:2208.01623.

[21] F. Ashtiani, A. J. Geers, and F. Aflatouni, An on-chip photonic deep neural network for image classification, Nature **606**, 1 (2022).

[22] X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti *et al.*, 11 TOPS photonic convolutional accelerator for optical neural networks, Nature **589**, 44 (2021).

[23] J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja *et al.*, Parallel convolutional processing using an integrated photonic tensor core, Nature **589**, 52 (2021).

[24] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Broadcast and weight: An integrated network for scalable photonic spike processing, J. Lightwave Technol. **32**, 4029 (2014).

[25] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. Pernice, All-optical spiking neurosynaptic networks with self-learning capabilities, Nature **569**, 208 (2019).

[26] A. Jha, C. Huang, H.-T. Peng, B. Shastri, and P. R. Prucnal, Photonic spiking neural networks and cmos-compatible graphene-on-silicon spiking neurons, arXiv:2109.13797.

[27] K. Vandoorne, W. Dierckx, B. Schrauwen, D. Verstraeten, R. Baets, P. Bienstman, and J. Van Campenhout, Toward optical signal processing using photonic reservoir computing, Opt. Express **16**, 11182 (2008).

[28] F. Duport, B. Schneider, A. Smerieri, M. Haelterman, and S. Massar, All-optical reservoir computing, Opt. Express **20**, 22783 (2012).

[29] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, Experimental demonstration of reservoir computing on a silicon photonics chip, Nat. Commun. **5**, 1 (2014).

[30] M. Rafayelyan, J. Dong, Y. Tan, F. Krzakala, and S. Gigan, Large-scale optical reservoir computing for spatiotemporal chaotic systems prediction, Phys. Rev. X **10**, 041037 (2020).

[31] A. N. Tait, T. F. Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Neuromorphic photonic networks using silicon photonic weight banks, Sci. Rep. **7**, 7430 (2017).

[32] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. Miller, and D. Psaltis, Inference in artificial intelligence with deep optics and photonics, Nature **588**, 39 (2020).

[33] M. Miscuglio, A. Mehrabian, Z. Hu, S. I. Azzam, J. George, A. V. Kildishev, M. Pelton, and V. J. Sorger, All-optical nonlinear activation function for photonic neural networks, Opt. Mater. Express **8**, 3851 (2018).

[34] B. Shi, N. Calabretta, D. Bunandar, D. Englund, and R. Stabile, in *2018 Photonics in Switching and Computing (PSC)* (IEEE, Limassol, Cyprus, 2018), p. 1.

[35] Y. Zuo, B. Li, Y. Zhao, Y. C. Chen, G. B. Jo, J. Liu, and S. Du, All-optical neural network with nonlinear activation functions, Optica **6**, 1132 (2019).

[36] M. Borghi, A. Trenti, and L. Pavesi, Four wave mixing control in a photonic molecule made by silicon microring resonators, Sci. Rep. **9**, 1 (2019).

[37] G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos, and N. Pleros, An all-optical neuron with sigmoid activation function, Opt. Express **27**, 9620 (2019).

[38] M. T. Hill, E. E. Frietman, H. de Waardt, G.-d. Khoe, and H. J. Dorren, All fiber-optic neural network using coupled SOA based ring lasers, IEEE Trans. Neural Netw. **13**, 1504 (2002).

[39] D. Rosenbluth, K. Kravtsov, M. P. Fok, and P. R. Prucnal, A high performance photonic pulse processing device, Opt. Express **17**, 22767 (2009).

[40] A. N. Tait, T. F. De Lima, M. A. Nahmias, H. B. Miller, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, Silicon photonic modulator neuron, Phys. Rev. Appl. **11**, 064043 (2019).

[41] R. Amin, J. George, S. Sun, T. Ferreira de Lima, A. N. Tait, J. Khurgin, M. Miscuglio, B. J. Shastri, P. R. Prucnal, T. El-Ghazawi *et al.*, ITO-based electro-absorption modulator for photonic neural activation function, APL Mater. **7**, 081112 (2019).

[42] J. K. George, A. Mehrabian, R. Amin, J. Meng, T. F. De Lima, A. N. Tait, B. J. Shastri, T. El-Ghazawi, P. R. Prucnal, and V. J. Sorger, Neuromorphic photonics with electro-absorption modulators, Opt. Express **27**, 5181 (2019).

[43] M. A. Nahmias, A. N. Tait, L. Tolias, M. P. Chang, T. Ferreira de Lima, B. J. Shastri, and P. R. Prucnal, An integrated analog O/E/O link for multi-channel laser neurons, Appl. Phys. Lett. **108**, 151106 (2016).

[44] I. A. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, Reprogrammable electro-optic nonlinear activation functions for optical neural networks, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2019).

[45] K. Harkhoe, G. Verschaffelt, A. Katumba, P. Bienstman, and G. Van der Sande, Demonstrating delay-based reservoir computing using a compact photonic integrated chip, Opt. Express **28**, 3086 (2020).

[46] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, Experimental demonstration of reservoir computing on a silicon photonics chip, Nat. Commun. **5**, 3541 (2014).

[47] S. Sunada and A. Uchida, Photonic neural field on a silicon chip: Large-scale, high-speed neuro-inspired computing and sensing, Optica **8**, 1388 (2021).

[48] A. Skalli, J. Robertson, D. Owen-Newns, M. Hejda, X. Porte, S. Reitzenstein, A. Hurtado, and D. Brunner, Photonic neuromorphic computing using vertical cavity semiconductor lasers, Opt. Mater. Express **12**, 2395 (2022).

[49] G. H. Li, C. R. Leefmans, J. Williams, R. M. Gray, M. Parto, and A. Marandi, Deep learning with photonic neural cellular automata, arXiv:2309.13186.

[50] V. Lopez-Pastor and F. Marquardt, Self-learning machines based on Hamiltonian echo backpropagation, Phys. Rev. X. **13** (3), 031020 (2023).

[51] S. Krastanov, M. Heuck, J. H. Shapiro, P. Narang, D. R. Englund, and K. Jacobs, Room-temperature photonic logical qubits via second-order nonlinearities, Nat. Commun. **12**, 1 (2021).

[52] L. Deng, The MNIST database of handwritten digit images for machine learning research, IEEE Signal Process. Mag. **29**, 141 (2012).

[53] I. Boikov, D. Brunner, and A. De Rossi, Direct coupling of nonlinear integrated cavities for all-optical reservoir computing, arXiv:2307.10950.

[54] M. Heuck, J. G. Koefoed, J. B. Christensen, Y. Ding, L. H. Frandsen, K. Rottwitt, and L. K. Oxenløwe, Unidirectional frequency conversion in microring resonators for on-chip frequency-multiplexed single-photon sources, New J. Phys. **21**, 033037 (2019).

[55] W. Suh, Z. Wang, and S. Fan, Temporal coupled-mode theory and the presence of non-orthogonal modes in lossless multimode cavities, IEEE J. Quantum Electron. **40**, 1511 (2004).

[56] A. E. Rastegin, Trace distance from the viewpoint of quantum operation techniques, J. Phys. A: Math. Theor. **40**, 9533 (2007).

[57] G. Cappellini and S. Trillo, Third-order three-wave mixing in single-mode fibers: Exact solutions and spatial instability effects, J. Opt. Soc. Am. B **8**, 824 (1991).

[58] M. Zhang, B. Buscaino, C. Wang, A. Shams-Ansari, C. Reimer, R. Zhu, J. M. Kahn, and M. Lončar, Broadband electro-optic frequency comb generation in a lithium niobate microring resonator, Nature **568**, 373 (2019).

[59] K. M. Kaini, H. M. Mbonde, H. C. Frankis, R. Mateman, A. Leinse, A. P. Knights, and J. D. B. Bradley, Four-wave mixing in high-Q tellurium-oxide-coated silicon nitride microring resonators, OSA Contin. **3**, 3497 (2020).

[60] Q. Nicolás and J. E. Sipe, Why you should not use the electric field to quantize in nonlinear optics, Opt. Lett. **42**, 3443 (2017).

[61] R. Paschotta, *Encyclopedia of Laser Physics and Technology* (Wiley-VCH, Zürich, Switzerland, 2008), Vol. 1.

[62] X. Ji, F. A. Barbosa, S. P. Roberts, A. Dutt, J. Cardenas, Y. Okawachi, A. Bryant, A. L. Gaeta, and M. Lipson, Ultra-low-loss on-chip resonators with sub-milliwatt parametric oscillation threshold, Optica **4**, 619 (2017).

[63] M. A. Nahmias, T. F. De Lima, A. N. Tait, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, Photonic multiply-accumulate operations for neural networks, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2019).

[64] M. J. Filipovich, Z. Guo, B. A. Marquez, H. D. Morison, and B. J. Shastri, in *2020 IEEE Photonics Conference (IPC)* (IEEE, Vancouver, BC, Canada, 2020), p. 1.

[65] M. Miscuglio, Z. Hu, S. Li, J. K. George, R. Capanna, H. Dalir, P. M. Bardet, P. Gupta, and V. J. Sorger, Massively parallel amplitude-only Fourier neural network, Optica **7**, 1812 (2020).

[66] B. Keller, R. Venkatesan, S. Dai, S. G. Tell, B. Zimmer, W. J. Dally, C. T. Gray, and B. Khailany, in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (IEEE, Honolulu, HI, USA, 2022), p. 16.

[67] H. Zhao, B. Li, H. Li, and M. Li, Enabling scalable optical computing in synthetic frequency dimension using integrated cavity acousto-optics, Nat. Commun. **13**, 1 (2022).

[68] C. Panuski, D. Englund, and R. Hamerly, Fundamental thermal noise limits for optical microcavities, Phys. Rev. X **10**, 041046 (2020).

[69] J. K. George, R. Amin, J. Meng, T. F. de Lima, A. N. Tait, B. J. Shastri, T. El-Ghazawi, P. R. Prucnal, and V. J. Sorger, Experimental realization of arbitrary activation functions for optical neural networks, Opt. Express **28**, 12138 (2020).

[70] D. Brunner, S. Reitzenstein, and I. Fischer, in *2016 IEEE International Conference on Rebooting Computing (ICRC)* (IEEE, San Diego, CA, USA, 2016), p. 1.

[71] J. George, R. Amin, A. Mehrabian, J. Khurgin, T. El-Ghazawi, P. R. Prucnal, and V. J. Sorger, in *Signal Processing in Photonic Communications* (Optical Society of America, Zürich, Switzerland, 2018), p. SpW4G.

[72] T. S. Rasmussen, Y. Yu, and J. Mork, All-optical nonlinear activation function for neuromorphic photonic computing using semiconductor Fano lasers, Opt. Lett. **45**, 3844 (2020).

[73] E. Mos, J. Schleipen, and H. De Waardt, Optical-mode neural network by use of the nonlinear response of a laser diode to external optical feedback, Appl. Opt. **36**, 6654 (1997).

[74] A. Dejonckheere, F. Duport, A. Smerieri, L. Fang, J. L. Oudar, J. Haelterman, and S. Massar, All-optical reservoir computer based on saturation of absorption, Opt. Express **22**, 10868 (2014).

[75] Z. Cheng, H. K. Tsang, X. Wang, K. Xu, and J. Xu, In-plane optical absorption and free carrier absorption in graphene-on-silicon waveguides, IEEE J. Sel. Top. Quantum Electron. **20**, 43 (2014).

[76] G. H. Li, R. Sekine, R. Nehra, R. M. Gray, L. Ledezma, Q. Guo, and A. Marandi, All-optical ultrafast ReLU function for energy-efficient nanophotonic deep learning, Nanophotonics **12**, 847 (2022).

[77] Q. Guo, R. Sekine, L. Ledezma, R. Nehra, D. J. Dean, A. Roy, R. M. Gray, S. Jahani, and A. Marandi, Femtojoule femtosecond all-optical switching in lithium niobate nanophotonics, Nat. Photonics **16**, 625 (2022).

[78] M. Jankowski, C. Langrock, B. Desiatov, A. Marandi, C. Wang, M. Zhang, C. R. Phillips, M. Lončar, and M. Fejer, Ultrabroadband nonlinear optics in nanophotonic periodically poled lithium niobate waveguides, Optica **7**, 40 (2020).

[79] J. Upham, Y. Tanaka, Y. Kawamoto, Y. Sato, T. Nakamura, B. S. Song, T. Asano, and S. Noda, Time-resolved catch and release of an optical pulse from a dynamic photonic crystal nanocavity, Opt. Express **19**, 23377 (2011).

[80] H. I. Nurdin, M. R. James, and N. Yamamoto, Perfectly capturing traveling single photons of arbitrary temporal wavepackets with a single tunable device, arXiv:1609.05643.

[81] E. Dulkeith, F. Xia, L. Schares, W. M. Green, and Y. A. Vlasov, Group index and group velocity dispersion in silicon-on-insulator photonic wires, Opt. Express **14**, 3853 (2006).

[82] J. Liu, G. Huang, R. N. Wang, J. He, A. S. Raja, T. Liu, N. J. Engelsen, and T. J. Kippenberg, High-yield, wafer-scale fabrication of ultralow-loss, dispersion-engineered silicon nitride photonic circuits, Nat. Commun. **12**, 2236 (2021).

[83] K. D. Shaw, in *Solid State Lasers and Nonlinear Crystals* (SPIE, San Jose, CA, United States, 1995), Vol. 2379, p. 365.

[84] S. Krastanov, Sigmoid Activation Function for Neural Networks based on Nonlinear Optics (2022), https://pluto.krastanov.org/nonlin-opt-sigmoid.html.

[85] M. Heuck, P. T. Kristensen, Y. Elesin, and J. Mørk, Improved switching using Fano resonances in photonic crystal structures, Opt. Lett. **38**, 2466 (2003).

[86] P. T. Kristensen, J. R. de Lasson, M. Heuck, N. Gregersen, and J. Mørk, On the theory of coupled modes in optical cavity-waveguide structures, J. Lightwave Technol. **35**, 4247 (2017).

[87] S. Bergfeld and W. Daum, Second-harmonic generation in GaAs: Experiment versus theoretical predictions of $\chi^{(2)}_{xyz}$, Phys. Rev. Lett. **90**, 036801 (2003).

[88] Z. Yan, H. He, H. Liu, M. Iu, O. Ahmed, E. Chen, P. Blakey, Y. Akasaka, T. Ikeuchi, and A. S. Helmy, χ 2-based AlGaAs phase sensitive amplifier with record gain, noise, and sensitivity, Optica **9**, 56 (2022).

[89] H. Sato, M. Abe, I. Shoji, J. Suda, and T. Kondo, Accurate measurements of second-order nonlinear optical coefficients of 6H and 4H silicon carbide, JOSA B **26**, 1892 (2009).

[90] J. Liu, G. Huang, R. N. Wang, J. He, A. S. Raja, T. Liu, N. J. Engelsen, and T. J. Kippenberg, High-yield, wafer-scale fabrication of ultralow-loss, dispersion-engineered silicon nitride photonic circuits, Nat. Commun. **12**, 1 (2021).

[91] L. Chang, A. Boes, P. Pintus, J. D. Peters, M. Kennedy, W. Jin, X.-W. Guo, S.-P. Yu, S. B. Papp, and J. E. Bowers, in *CLEO: Science and Innovations* (Optical Society of America, San Jose, CA, United States, 2019), p. SF2I.

[92] A. N. R. Ahmed, A. Mercante, S. Shi, P. Yao, and D. W. Prather, Vertical mode transition in hybrid lithium niobate and silicon nitride-based photonic integrated circuit structures, Opt. Lett. **43**, 4140 (2018).

[93] T. Tiecke, K. Nayak, J. D. Thompson, T. Peyronel, N. P. de Leon, V. Vuletić, and M. Lukin, Efficient fiber-optical interface for nanophotonic devices, Optica **2**, 70 (2015).

[94] H. Larocque, M. A. Buyukkaya, C. Errando-Herranz, S. Harper, J. Carolan, C.-M. Lee, C. J. Richardson, G. L. Leake, D. J. Coleman, M. L. Fanto *et al.*, Tunable quantum emitters on large-scale foundry silicon photonics, arXiv:2306.06460.

[95] S. Pai, I. A. Williamson, T. W. Hughes, M. Minkov, O. Solgaard, S. Fan, and D. A. Miller, Parallel programming of an arbitrary feedforward photonic network, IEEE J. Sel. Top. Quantum Electron. **26**, 1 (2020).

[96] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[97] S. Bandyopadhyay, R. Hamerly, and D. Englund, Hardware error correction for programmable photonics, Optica **8**, 1247 (2021).

[98] S. K. Vadlamani, D. Englund, and R. Hamerly, Transferable learning on analog hardware, Sci. Adv. **9**, eadh3436 (2023).

[99] X. Guo, T. D. Barrett, Z. M. Wang, and A. Lvovsky, Backpropagation through nonlinear units for the all-optical training of neural networks, Photonics Res. **9**, B71 (2021).

[100] M. Prabhu, C. Roques-Carmes, Y. Shen, N. Harris, L. Jing, J. Carolan, R. Hamerly, T. Baehr-Jones, M. Hochberg, V. Čeperić *et al.*, Accelerating recurrent Ising machines in photonic integrated circuits, Optica **7**, 551 (2020).

[101] R. Wang, A. Malik, I. Šimonytė, A. Vizbaras, K. Vizbaras, and G. Roelkens, Compact GaSb/silicon-on-insulator 2.0 × µm widely tunable external cavity lasers, Opt. Express **24**, 28977 (2016).

[102] U. Keller, Recent developments in compact ultrafast lasers, Nature **424**, 831 (2003).