OPTICS

Event-driven adaptive optical neural network

Frank Brückerhoff-Plückelmann[1], Ivonne Bente[1], Marlon Becker[2], Niklas Vollmar[3], Nikolaos Farmakidis[4], Emma Lomonte[1], Francesco Lenzini[1], C. David Wright[5], Harish Bhaskaran[4], Martin Salinga[3], Benjamin Risse[2], Wolfram H. P. Pernice[1,6]*

We present an adaptive optical neural network based on a large-scale event-driven architecture. In addition to changing the synaptic weights (synaptic plasticity), the optical neural network's structure can also be reconfigured enabling various functionalities (structural plasticity). Key building blocks are wavelength-addressable artificial neurons with embedded phase-change materials that implement nonlinear activation functions and nonvolatile memory. Using multimode focusing, the activation function features both excitatory and inhibitory responses and shows a reversible switching contrast of 3.2 decibels. We train the neural network to distinguish between English and German text samples via an evolutionary algorithm. We investigate both the synaptic and structural plasticity during the training process. On the basis of this concept, we realize a large-scale network consisting of 736 subnetworks with 16 phase-change material neurons each. Overall, 8398 neurons are functional, highlighting the scalability of the photonic architecture.

INTRODUCTION

The rise of artificial intelligence and the end of Moore's law require novel computation methods to fulfill the ever-growing demand for computational power (*1*). Brain-inspired or neuromorphic approaches have been one promising avenue because of gains in energy efficiency when implementing artificial neural networks (NNs). Neuromorphic devices mimic NNs on a hardware level instead of simulating them on conventional computers. From an operational viewpoint, this translates into highly parallel operation, event-driven in-memory computing, and stochasticity (*2*). From a hardware perspective, a neuromorphic system needs to be inherently scalable and requires "artificial neurons" that perform nonlinear operations and "artificial synapses" that store the linear connections between the neurons. Typically, the neurons in biological brains are relatively sparsely connected compared to the overall number of nodes in the network. This is favorable from a hardware perspective because interconnectivity in planar devices, as commonly used in optical and electrical approaches, is challenging. Furthermore, the neuromorphic device ideally needs to be trainable and thus adaptive in the sense that both the connection strength can be changed (synaptic plasticity) and the connection itself can be rewired (structural plasticity). Commercial electronic devices such as the Google tensor processing unit that are inspired by the principles of neuromorphic computing show a gain of factor 10 in computation speed while consuming 100 times less energy than conventional hardware for computing NNs (*3*). Going one step further, large-scale event-driven neuromorphic research prototypes like BrainScaleS (*4*), Loihi (*5*), or TrueNorth (*6*) consist of integrated electrical spiking NNs, featuring up to 1 million neurons on-chip. These event-driven architectures promise even higher energy efficiency in comparison to conventional layered structures since inactive neurons do not contribute to the computation and thus do not require energy.

Besides electronic approaches, photonic architectures are emerging to complement integrated circuit implementations. Photonic neuromorphic computing systems aim to outperform their electronic counterpart by exploiting the inherently low latency of optical data transfer, the large usable optical bandwidth of several terahertz, and additional degrees of freedom for parallelization based, for example, on wavelength division multiplexing (WDM) (*7*). WDM offers a direct way for parallel data processing and passive on-chip routing and is compatible with broadband optical approaches. In particular, WDM enables parallel processing with a single photonic computation unit (*8*), broadcast-and-weight networks (*9*), and temporal encoding in combination with dispersive waveguides (*10*). By integrating functional materials, photonic platforms further offer the possibility to implement reconfigurable elements. Among the available options, phase-change materials (PCMs) are particularly attractive because they provide strong optical contrast and can be reversibly switched for many cycles. When embedded in a waveguide framework, optically coupled PCMs offer a compact way of integrating nonlinearities and memory into the photonic circuit (*11*). Thus, PCMs are deployed as reprogrammable synapses (*12*), activation functions (*13*), and in-memory computation units (*14*) among other applications (*15*). Apart from hardware accelerators for designated linear computation tasks, exemplary optical NNs (ONNs) also including nonlinearities have been implemented. A basic nonlinear behavior can be realized by nonlinear optical effects (*16*, *17*), hybrid optical-electronic schemes (*18*), and functional materials (*13*). Going one step further, optical spiking neurons based on vertical-cavity surface-emitting lasers have been demonstrated (*19*, *20*). While small-scale feed-forward networks with a fixed structure have been realized on-chip (*13*, *21*, *22*) and also larger recurrent networks off-chip (*23*), operative event-driven architectures featuring both synaptic and structural plasticity have been elusive.

Here, we present an ONN that also exhibits structural plasticity in addition to synaptic plasticity. While many implementations of ONNs are limited to feed-forward connectivity and single-use of

[1]Physical Institute, University of Münster, Heisenbergstraße 11, 48149 Münster, Germany. [2]Institute for Geoinformatics, University of Münster, Heisenbergstraße 2, 48149 Münster, Germany. [3]Institute of Materials Physics, University of Münster, Wilhelm-Klemm-Straße 10, 48149 Münster, Germany. [4]Department of Material, University of Oxford, Parks Road, Oxford OX1 3PH, UK. [5]Department of Engineering, University of Exeter, North Park Road, Exeter EX4 4QF, UK. [6]Kirchhoff-Institute for Physics, University of Heidelberg, Im Neuenheimer Feld 227, 69120 Heidelberg, Germany.
*Corresponding author. Email: wolfram.pernice@kip.uni-heidelberg.de

neurons, our proposed event-driven architecture computes activations and neuron updates iteratively, allowing multiuse of neurons, nested connectivity, as well as sequential input and output data. Although related to recurrent NNs (RNNs), our algorithm is, to the best of our knowledge, unique, not only in the optical domain but also in the NN literature in general. We realize in-memory computing by using $Ge_2Sb_2Te_5$ (GST) nanocells embedded in multimode interference (MMI) devices. While the GST nanocells emulate artificial neurons, the artificial synapses are encoded in the wavelength and power of optical pulses driving the ONN. To implement WDM capability, we use a self-aligning combination of ring resonators and reflective Bragg mirrors for wavelength-selective addressing of individual neurons. The use of programmable PCMs and programmable optical pulses enables both synaptic and structural plasticity. To demonstrate the capability of the approach, we fabricate networks of interconnected wavelength-addressable GST neurons and characterize their individual performance. Exploiting on-chip plasticity, we train the neurons to distinguish between English and German text samples via an evolutionary algorithm. During the training, we investigated the structural changes of the NN. Since our approach is inherently scalable, we fabricate a large-scale network consisting of 736 subnetworks with 11,776 neurons in total to showcase the flexibility and excellent future scalability potential of the architecture.

Event-driven optical computation architecture

Our architecture is based on the concept illustrated in Fig. 1, which consists of artificial photonic neurons on-chip and artificial synapses which are encoded in optical pulses via wavelength and amplitude. A large number of individually wavelength-selectable artificial neurons is provided on a fabricated chip. The desired network configuration is implemented by connecting PCM-based neurons with suitable optical pulse links as sketched in Fig. 1A. This approach enables storing the internal state of the network in the neurons, while the network structure is solely defined through routing suitable pulse sequences to a desired subset of the NN. As the routing is wavelength-based, we directly achieve structural plasticity by adapting the pulse sequence, in particular the wavelength of the pulses driving the ONN. The photonic circuit itself does not need to be modified during operation, greatly simplifying the overall system.

The main building blocks of the adaptive ONN are wavelength-addressable PCM neurons, connected via a single bus waveguide. Each PCM neuron consists of a wavelength-selective optical element and a waveguide-coupled PCM cell. The PCM cell implements multilevel nonvolatile memory and thus enables storing of the internal state of the neuron. By wavelength multiplexing, we can address multiple neurons at the same time by sending several optical pulses of different wavelengths through the bus waveguide. To realize the on-chip neurons, we use the nonvolatile PCM GST which is highly absorptive in its crystalline state and just barely absorbs propagating light in the amorphous state. We reversibly switch between the structural phase states of the GST cell with high-power optical pulses sent through the waveguide (*12*, *13*, *24*, *25*). Low-power optical pulses heat the GST cell so moderately that its phase state is not altered by them. In contrast, the phase state and hence the transmission through the GST cell change if the optical pulse power is sufficient for crystallization or amorphization. The change in transmission through the GST cell is highly nonlinear with respect to the applied pulse power (*13*). Therefore, the GST serves as a controllable nonlinear activation function and simultaneously as a memory for low-power optical read pulses (*26*).

Our optical system consisting of multiple PCM neurons (Fig. 1A) is driven by a queue of events as shown in Fig. 1B. The system operates asynchronously as the event is immediately processed and the nonvolatility of the GST cells does not impose further time restrictions. Each event consists of multiple optical pulses defining the weighted connections between the neurons via the pulse wavelengths (synaptic connections) and intensities (synaptic strengths). Therefore, we can use the same photonic structure to compute different NNs by changing the event queue as sketched in Fig. 1A. Apart from the flexible utilization of the computation architecture, it directly enables structural plasticity during the training by changing the pulse wavelengths. Besides enabling adaptive ONNs, this approach greatly improves durability and tolerance, since single faulty neurons can be replaced with unused ones, which only reduces the number of available neurons but does not affect the overall computation.

Figure 1B explains the working principle of the architecture in depth. Desired synaptic weights are encoded in the intensity of optical pulses. These pulses serve as inputs to desired neurons which are selected by wavelength using the multiplexing approach described in further detail below. In general, the state of the "postsynaptic" neuron is computed by multiplying the activations of the "presynaptic" neurons (saved in the GST cells) with the synaptic weight of the respective connection between pre- and postsynaptic neuron (encoded the intensity of an optical pulse), accumulating those weighted activations (by letting the attenuated pulses of different wavelengths superimpose on a photodetector), and applying the activation function of the postsynaptic neuron for the given weighted sum. A single event, which consists of simultaneous pulses with chosen wavelengths and intensities, directly implements all those steps. Each event updates the internal state (GST phase) of one neuron, the postsynaptic neuron. After choosing this neuron, several low-power optical pulses are sent to the system simultaneously, each pulse representing one synaptic connection from a presynaptic neuron. For each pulse, the wavelength corresponds to the wavelength address of the presynaptic neuron and the pulse power corresponds to the synaptic strength. Next, the pulses are propagated through the photonic circuit via waveguides and are attenuated by the respective PCM neuron where the exact attenuation is determined by the PCM′s state. Afterward, the thus-processed optical pulses are collected at the output of the circuit and are summed by a photodetector. To finish the event, a high-power optical pulse is sent to the ONN with a pulse power proportional to the measured photodetector voltage and a wavelength corresponding to the desired postsynaptic neuron′s wavelength address. In this way, the postsynaptic neuron′s state is set, determining the activation response of this neuron for the next event according to the switching properties of the GST cell. We limit the switching pulse power to a maximum value to avoid damaging the PCM of the postsynaptic neuron. Afterward, we repeat this sequence for the next event in the event queue. If required, for example, for the final classification, we read out the state of a neuron with a low-power read pulse with a wavelength corresponding to the neuron′s address. Figure 1C shows the equivalent steps carried out during the physical operation in the ONN compared to the framework of artificial NNs (NN). The steps in the NN are indicated in orange in the top row,

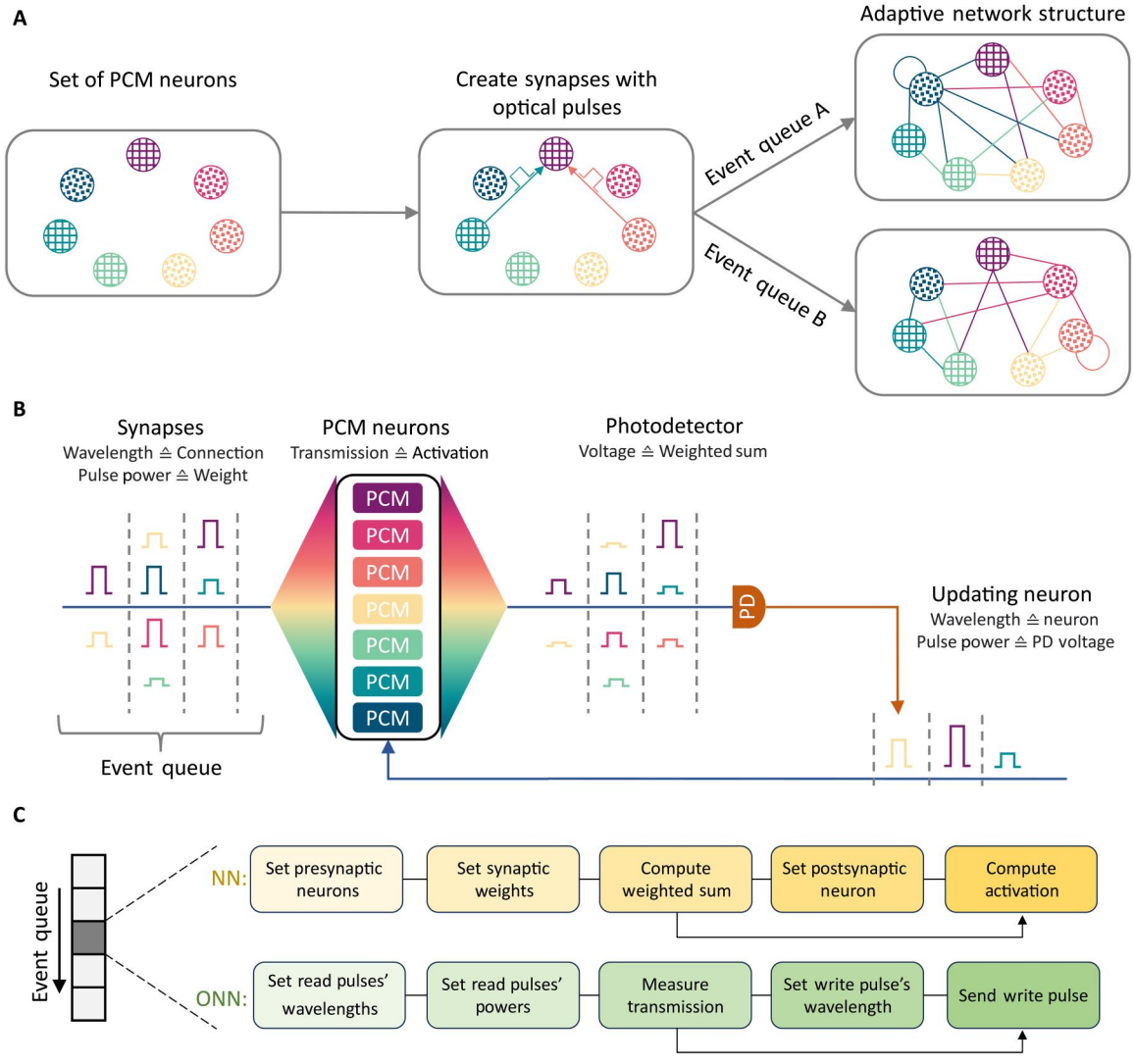

Fig. 1. Computing neural networks on the event-driven architecture. (**A**) The system consists of waveguide-coupled PCM cells that emulate the functionality of artificial neurons. We address every PCM neuron on-chip with a distinct wavelength. To plastically connect the ONN, we encode the synapses in the event queue which is driving the system. This way, we can adapt and change the structure of the whole NN without having to modify the photonic circuit itself. (**B**) We operate the system sequentially, executing one event after the other. For each event, we send a set of input pulses (the synaptic configuration information) to the PCM neurons simultaneously. The wavelengths of the pulses correspond to the wavelength addresses of the presynaptic neurons and the pulse powers to the synaptic weights. The pulse powers are multiplied with the respective presynaptic PCM neuron activation and are afterward summed by a photodetector (PD). To conclude the event, a high-power write pulse is sent to the ONN, with a wavelength corresponding to the address of the postsynaptic neuron and a pulse power proportional to the output voltage of the PD. In this way, the state of the postsynaptic neuron is set. (**C**) Correspondence between neural nets and their optical implementation. Each physical operation performed by the ONN (green, lower row) can be directly mapped to the computation steps in an artificial NN (orange, upper row).

while the corresponding steps in the ONN are shown in green in the row below.

RESULTS

To implement the conceptual architecture outlined above, we fabricate the system in the silicon nitride platform. Several artificial neurons are linked to one bus waveguide which is used as input and output by operating the neurons in reflection mode. This collection of neurons is then repeated many times on the overall chip to arrive at a large number of usable neurons for implementing desired ONNs. Wavelength selectivity is implemented using microring resonators with varying radii.

Figure 2A shows an optical microscope image of the photonic circuit consisting of the bus waveguide and 16 wavelength-addressable PCM neurons in a false-color overlay, which illustrates the different resonance wavelengths. The varying resonance wavelength is indicated by the different colors, corresponding to rings with different radii. By operating the rings in reflection mode with a Bragg reflector provided in the drop waveguide, we achieve perfectly matched demultiplexing and multiplexing. This way, we use the ring add-drop filters as wavelength-selective devices with pulses propagating in both directions. As a result, also the output signals

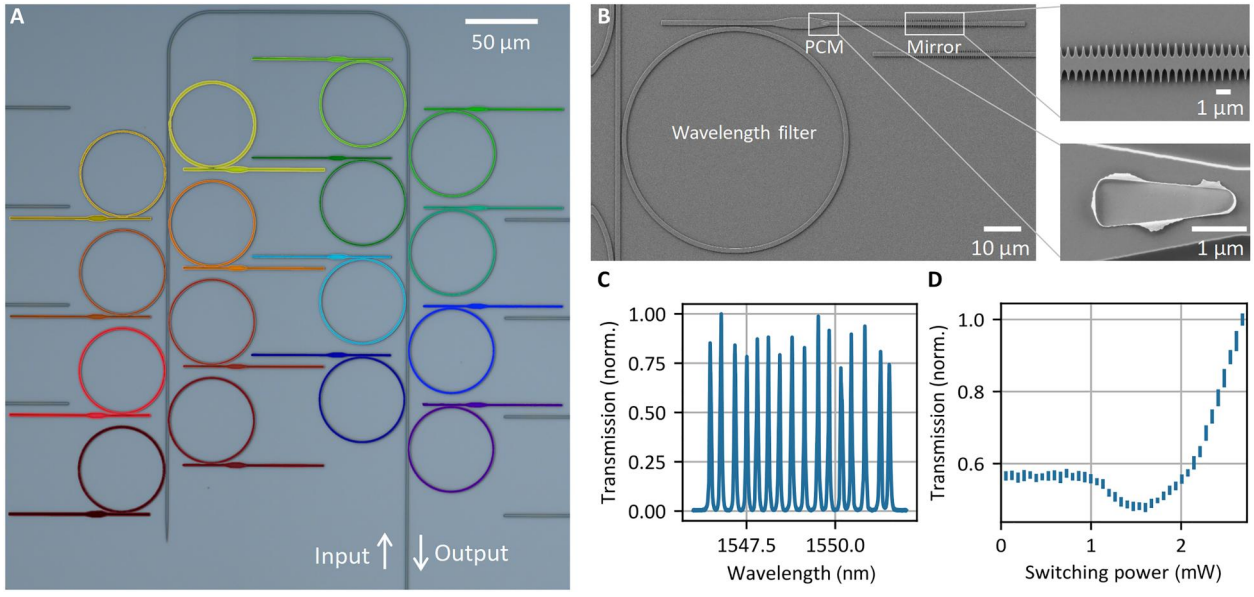

Fig. 2. Photonic circuit implementing a subset of 16 artificial neurons. (**A**) Input pulses are sent to the PCM neurons via a common bus waveguide. Microring resonators select a pulse of a certain wavelength from the bus waveguide, let the pulse interact with the PCM cell and couple it back to the bus waveguide in reverse direction. (**B**) The scanning electron microscope shows a single optical neuron. It consists of a critically coupled ring resonator for wavelength addressing, a PCM cell on top of an MMI focusing structure and a compact Bragg mirror to reflect the light back to the bus waveguide. The insets show a zoom of the Bragg mirror and the PCM cell. (**C**) The transmission spectrum of the photonic circuit in (A) shows 16 clearly distinguishable resonance peaks, each one corresponding to a different optical PCM neuron. (**D**) We obtain both excitatory and inhibitory behavior of the neuron by defining a partially amorphous state as the ground state in our neural network and switch the cell depending on the input pulse power. The bar length shows the SD of the transmission for each PCM state when randomly switching between the states.

are propagating in the same bus waveguide as the input signals, but in the opposite direction. Figure 2B shows a single PCM neuron unit in detail. The ring resonator is coupled to the bus waveguide and implements the wavelength filter. Because of the ultralow loss of the silicon nitride platform, we achieve critical coupling from the bus waveguide to the ring for equal gaps between the ring and bus waveguide as well as the ring and neuron waveguide (see fig. S2). The symmetric design ensures that a small deviation from the designed coupling strength only changes the resonance width but not the peak transmission. The radius of the ring resonator affects the resonance wavelength and is different for each ring to ensure spectrally distinguishable PCM neurons. The actual neuron is placed in the drop waveguide within a carefully designed MMI structure (Fig. 2B, insets). To implement the neuron, we use a 12-nm-thick GST cell covered by 5-nm ZnS-SiO_2 placed in the evanescent near field of the MMI. The focusing MMI structure improves the coupling to the GST cell by smoothening the absorption profile and thus distributes the optical power over the complete area of the GST cell (see fig. S4). At the output of the MMI, we use a compact apodized Bragg mirror (see fig. S3) to reflect the signal back into the ring. Figure 2C shows the transmission profile of a full device. As designed, there are 16 distinguishable transmission peaks with an average spacing of 336 ± 47 pm, each corresponding to one of the PCM neurons. Each single peak arises from the transfer function of a single add-drop ring which couples the light from the bus waveguide and back (see fig. S2B) and includes the reflection spectrum of the Bragg mirror (see fig. S2C) and the PCM cell itself. Overall, the cross-talk between the PCM neurons is below −23.9 ± 3.5 dB. The symmetric multiplexing and demultiplexing design ensures a low variation in the peak transmission below 0.39 dB even if the resonances are closely spaced in wavelength. Therefore, each neuron receives similar optical power during addressing. Figure 2D shows the resulting activation function of such a PCM neuron unit depending on the input power of a pulse train of 18 pulses with 200-ns length, starting from a partial amorphous state set by a single pulse of 2-mW power. For pulse powers below approximately 1 mW, the activation does not change since the GST is not heated above its crystallization temperature. For higher pulse powers, the PCM starts to crystallize further and thus the transmission decreases. For even higher pulse powers, the GST is partially amorphized and the transmission increases again. Overall, the contrast between the high and low activation states is 3.2 dB. To determine the repeatability of the process, the PCM neuron is randomly switched between the states 1000 times. The mean transmission SD for each state is 1.4%. From Fig. 2D, note that the activation function has an inhibitory area where an increased input signal decreases the activation and an excitatory area where an increasing input power increases the activation, similar to the biological counterpart (*27*). As the architecture only supports positive weights, excitatory-only activation functions could only propagate excitatory signals between the neurons but could not attenuate a neuron′s activation. By deploying an activation function exhibiting both an inhibitory and excitatory area, this functionality is provided even in the presence of the restriction to only positive synaptic weights and activations in the NN.

To study the performance of the system, we implement an exemplary NN which is trained on-chip to distinguish between English and German text samples based on the distribution of the vowels. We use six of the PCM neurons in the photonic circuit shown in Fig. 2A and, for each sample, we encode the distribution of

vowels in five of those by setting their starting state with a corresponding high-power pulse. Subsequently, we send the event queue to the network. Then, we perform the classification of the text based on reading out the activation of the sixth neuron, where a high transmission corresponds to an English text and a low transmission to a German one. Because the structure of the event queue is, in general, not differentiable, we choose an evolutionary algorithm (*28*) for training, instead of using backpropagation. We fix the event queue to three events and send two input pulses corresponding to two synapses per event. In addition, we introduce a sanity check for the event queues to make sure that only input neurons or neurons that have been activated before can serve as presynaptic neurons. This check can be adapted if the internal state of the network must be remembered for the next event queue, for example, when processing temporal input data.

Figure 3A illustrates how the evolutionary algorithm finds an event queue that is suitable for the classification task. The algorithm starts by randomly generating two event queues, each event consisting of the synaptic connections (wavelength address), the synaptic strength (pulse power), and the selection of the postsynaptic neuron (wavelength address). The event queues consist of three events with two presynaptic neurons in each event. Then, the actual evolution process consisting of three individual steps is performed. First, we generate a "child" event queue from two "parent" event queues. Here, we randomly recombine the events from both queues, for example, taking the first two from "parent 1" and the last one from "parent 2." Next, we mutate the parent event queues by varying each event individually. Mutation is achieved by adding Gaussian noise to the optical pulse power which sets the synaptic weight. During the mutation, we also potentially rewire the synapse by changing the wavelength of the pulse. Moreover, we add a new random event queue to the population, to minimize the risk of converging into a local minimum. In the third step, we check all queues for sanity and test them on random text samples. If the stopping criterion, sufficient performance, or maximum number of evolutions is reached, then the best-performing queue is selected, and the algorithm stops. Otherwise, the population is reduced to the two best-performing queues which then serve as the parent queues for the next evolution step. A detailed experimental description of the training process is provided in the Supplementary Materials.

We generate German and English training data by using ChatGPT to compose short paragraphs of 1000 words in both languages. During the training of the ONN, we start at a random point in the chat and determine the distribution of the vowels in the next 200 words. In each evolution, we test each NN with three English text samples and three German text samples. During training, we minimize the ratio between the average output activations $A_{\text{out,German}}$ and $A_{\text{out,English}}$. Figure 3 (B and C) shows how the network's structure and performance change over the number of evolutions. In the first evolution, the network cannot clearly differentiate between the text samples. Already in the next evolution, the network finds a working event queue that clearly distinguishes between both languages. Also, the structure of the NN changes in this step. For the other evolutions, only the synaptic weights are adapted. For the last four evolutions, the best-performing event queue remains the same, successfully classifying 24 text samples. For evaluating the network performance, we use a different set of text samples in each evolution, which results in a slightly different output for an "English" or "German" label for the same network. Apart from extracting a suitable network for the classification task during the training process, we can also learn from the structure of the network. Once the network can successfully differentiate between languages, only the number of "a" and "o" in the

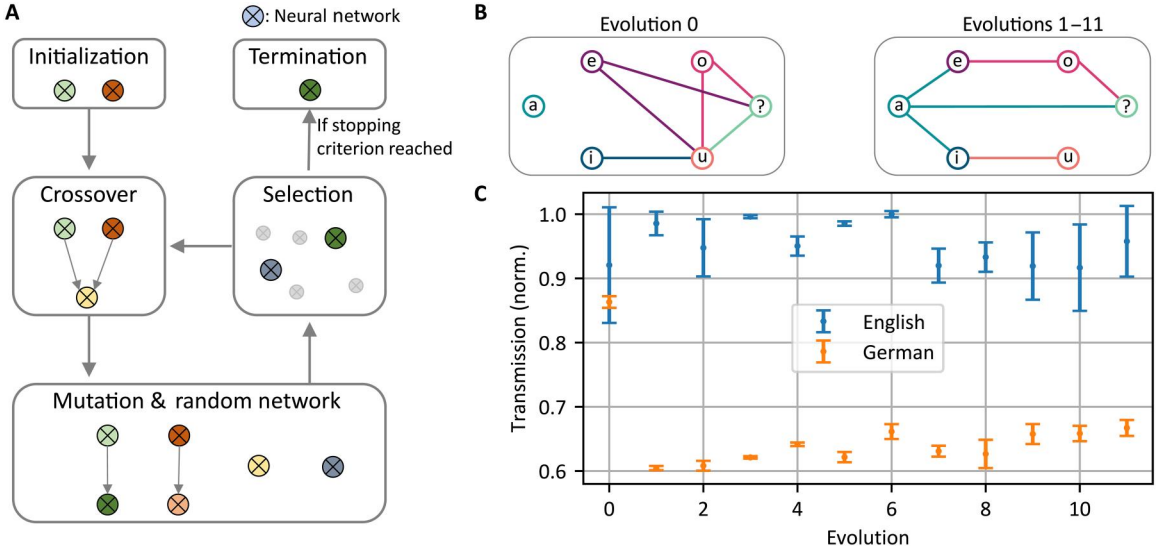

Fig. 3. Evolutionary learning with an adaptive ONN. (**A**) We use an evolutionary algorithm to learn an event queue for distinguishing between German and English text samples. First, we randomly initialize a set of neural networks. Next, we perform a crossover between the neural networks, combining synapses from one neural network with those of another one. Then, the parent event queues are randomly mutated, and a random queue is added to the population. Last, we choose a subset of all event queues based on their performance. If the performance is sufficient, then we stop the learning algorithm; otherwise, we repeat the mating process on the selected queues. (**B**) We use an evolutionary algorithm to train the ONN on-chip. Apart from changes in the synaptic weights, the structure of the network is also modified. (**C**) Already after the second evolution, the network can clearly distinguish between the text samples.

distribution of vowels is necessary to distinguish between the text samples. Consequently, the other vowels do not need to be considered, further improving the efficiency of the network.

Last, we characterize the large-scale system shown in Fig. 4A consisting of 738 subnetworks with 11,776 PCM neurons integrated on an area of approximately 235 mm^2. By connecting each subnetwork with a neighboring one, an indirect long-range interaction between the neurons can be created. Figure 4B shows the distribution of the neurons' transmissions in the as-sputtered amorphous state and crystalline state. For the amorphous state, the neurons follow a Gaussian distribution with a mean transmission of −18.25 dB and an SD of only 0.95 dB, highlighting the robustness of the photonic design. Around −14 dB of the total insertion loss (−18.25 dB) is attributed to the grating couplers (−8 dB total assuming −4 dB per coupler) which are used to couple the light to the chip, and to the 50:50 directional coupler (−6 dB total, −3 dB per pass in forward and backward direction) separating the back reflected light from the input light. In addition, there is absorption in the amorphous state of the PCM cell, coupling loss to and from the ring resonator, reflection losses at the Bragg mirror, and a waveguide propagation loss (waveguide length of up to 2 cm). Overall, 9854 neurons of the chip show reasonable transmission above −21 dB in the amorphous state. Also, for the crystalline state, the transmission follows a Gaussian distribution with a mean of −33.99 dB and an SD of 2.54 dB. The increase in the SD is mainly caused by slightly varying sizes of PCM patches during fabrication and causes a large deviation of the transmission in the crystalline state due to the higher absorptance. Since the GST cell is comparatively small with a tapered width from 825 to 422 nm over a length of 2.4 μm, and placed on a focusing MMI structure, deviations in the patch size have a large impact on the transmission. In total, 9002 neurons show a transmission between −40 and −27 dB in the crystalline state. Figure 4C shows a heatmap of functional neurons which show at least a contrast of −10 dB between the amorphous and crystalline state. Typically, either most neurons in a subnetwork are functional or the whole subnetwork is broken because of a fabrication error like a broken grating coupler. In total, 8398 neurons are functioning here, resulting in a fabrication yield of 71.3%.

DISCUSSION

ONNs (*13*, *21*, *29*, *30*) aim to emulate the structure and thus efficiency of biological NNs. Photonic implementations especially excel at calculating the linear connections (synapses) between the neurons, whereas electronic devices such as photodiodes and modulators or special material properties like those of PCMs are better suited to compute the needed nonlinearities. All optical solutions based on nonlinear effects strongly increase the complexity of the photonic circuit and require higher energies to trigger corresponding effects (*31*). Current integrated ONN prototypes have in common that they emulate the full layered structure of feed-forward networks. Consequently, the photonic circuits needed to realize complex networks become quite large which limits the network size and decreases the fabrication error tolerance. In contrast, the event-driven ONN architecture described here is easily scalable due to the separation of neurons and their activations from the synapses. This is possible because the structural information is stored in the wavelength of the optical pulses instead of being hard-wired into the circuit. The main advantage of the concept is that the structure and type of the network are not predetermined by the photonic circuit anymore but solely depend on the events driving the network. Consequently, instead of rendering the functionality of the whole photonic circuit useless, faulty neurons can simply be replaced by changing the structure of the network. Furthermore, the event-driven approach allows one to compute only the synapses that are active during a given event and thus does not require one to update irrelevant neurons in each step.

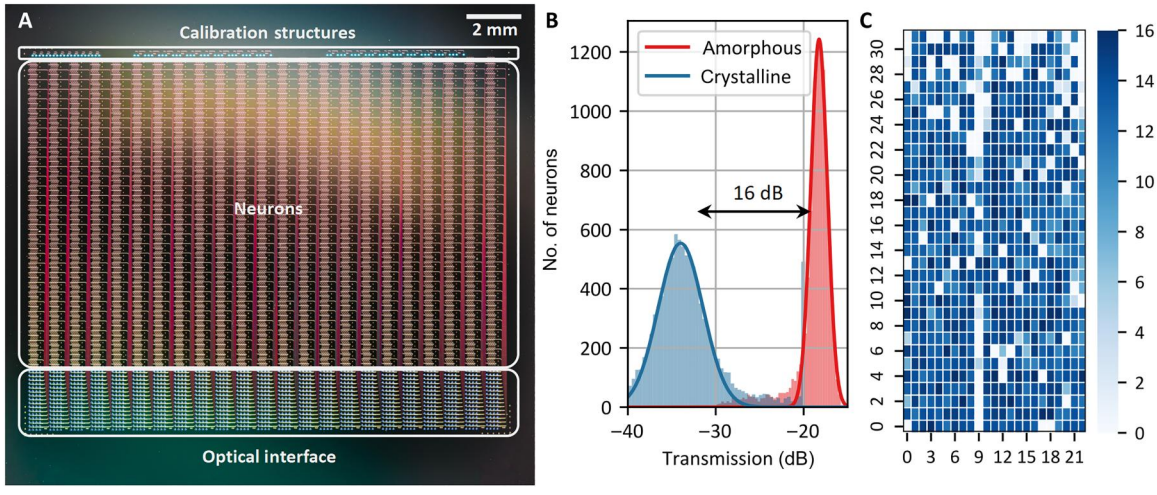

Fig. 4. Scalability of the adaptive ONN. (**A**) We characterize the performance of an ONN consisting of 11,776 different PCM neurons arranged in a 32 × 23 array of separate devices. Each device contains 16 neurons. (**B**) For each PCM neuron, we measure the overall transmission in the as-sputtered amorphous and fully crystalline state. We obtain a switching contrast of 16 dB between the amorphous (−18.25 ± 0.95 dB) and crystalline state (−33.99 ± 2.54 dB). Some GST cells are not deposited at all, resulting in the blue peak of around 500 nominally "crystallized" neurons at −20 dB due to a saturated detector. (**C**) Typically, the neurons in a subnetwork are either mostly functional and show a contrast between the amorphous and crystalline state of at least 10 dB, or completely broken due to a fabrication error in the coupling or routing. In total, 8398 neurons are functional.

Using PCMs to implement artificial neurons comes with the benefit of a strong nonlinear effect while also providing a nonvolatile memory. In particular, the nonvolatility strongly decreases the timing requirements of subsequent pulses reading the neuron activation and enables the event-based structure in the first place. Furthermore, the inhibitory and excitatory parts of the activation function compensate for only positive weights in our network. Optically switching GST cells up to 100 million cycles has been already demonstrated (*32*) and reversible electrical switching for more than 2 billion cycles (*33*). Changing to monoatomic PCMs such as antimony is a promising route to further improve the performance (*34*).

By placing the GST on an MMI-like structure and using it in a double-pass configuration, we are able to achieve a high switching contrast of more than 3 dB, which removes any need to amplify the effect of GST by, for example, integrating it into a separate ring resonator (*13*). Consequently, the neuron′s wavelength address is always the wavelength of the highest transmission and thus can be easily used for a feedback loop compensating for drifts in the lasers or photonic circuit induced by temperature. As a future step, waveguide-coupled plasmonic antennas could be used to provide broadband field enhancement in the GST region, potentially increasing the switching contrast and thus the capability of the ONN while simultaneously reducing switching energies and switching times (*35*). By using a single add-drop ring in both directions for multiplexing instead of two different rings (*14*), we create a self-aligning demultiplexer-multiplexer structure that strongly improves the fabrication tolerance of the system.

In contrast to backpropagation-based training (*31*), evolutionary algorithms excel at training complex physical systems (*36*) and simplify the training process. The structural plasticity of our ONN is not differentiable and thus not directly compatible with backpropagation. Since all GST neurons behave slightly differently (see fig. S5), backpropagation would also require all neurons to be characterized individually before the training process. Also, the setup itself might have noise sources and offsets that had to be modeled. Therefore, directly training on-chip with an evolutionary algorithm is a straightforward option to train the event-driven ONN.

Overall, the ONN architecture described above is readily scalable and rather compact for a photonic circuit due to the small size of the PCM neurons. The largest part of the footprint results from the 30-μm-radius ring resonators. By fabricating the neurons in a photonic platform with higher refractive index contrast such as silicon on an insulator, the bending radius can potentially be reduced by a factor of 10, leading to a 100 times higher integration density and up to 1 million neurons on a single 20 mm–by–20 mm chip. Making use of vertical integration of the silicon and silicon nitride waveguides (*37*) the mostly straight large-scale optical routing can still be done in silicon nitride, ensuring an overall low propagation loss. The self-aligning multiplexer-demultiplexing structure ensures a reasonable fabrication tolerance for the photonic circuit.

To develop the approach further, an integrated driver module would ideally be needed to generate the events driving the ONN on-chip. During our experiments, we were limited by the off-chip laser sources. By deploying integrated lasers with electro-optic phase shifters for frequency tuning, subnanosecond wavelength switching times can be achieved (*38*) and the overall system integrated.

MATERIALS AND METHODS

We fabricate the ONN on a Rogue Valley Microdevices wafer with a layer stack consisting of 330-nm Si3N4 on 3300-nm SiO2 on 500-μm Si. We anneal the chip at 1100° for 4 hours before the fabrication of the photonic circuit to improve the silicon nitride film quality. Overall, we fabricate four different masks with a 100-kV electron beam lithography system (Raith, EBPG5150). First, we spin-coat, expose, and develop the positive tone resist polymethyl methacrylate (PMMA; Allresist, AR-P 672.045) and evaporate gold alignment markers afterward. Next, we spin-coat the negative tone resist AR-N 7520.12 (Allresist) and write the mask for the photonic circuit. We fully etch the mask into the silicon nitride via reactive ion etching with CHF3/O2 plasma and strip the remaining AR-N afterward. The etching process results in nearly rectangular waveguides with a width of about 1.18 μm. The deviation from the designed width of 1.2 μm is attributed to the shrinkage of the mask before the etching. We again use PMMA to fabricate the mask for the phase change material and its capping layer. Then, we deploy a magnetron sputtering system (PVD, AJA International Inc.) to radio frequency sputter 12-nm GST and 5-nm ZnS-SiO2. The ZnS-SiO2 protective cover prevents oxidation of the GST. Last, we spin-coat and expose an 800-nm cladding layer of HSQ 16% (Dow Corning) overall structures.

For characterizing the photonic circuit and training the NN, we deploy three tunable lasers (two Santec TSL-710 and one New Focus 6427). We generate desired optical pulses from the continuous wave laser with an Optilab electro-optic modulator (IML-1550-40-PM-V) driven by a pulse generator (Hewlett Packard, 8131A) and amplify them with a PriTel LNHPFA-30. For the transmission measurements, we use Thorlabs RXM10AF and New Focus Model 2011 detectors. The whole setup is sketched in detail in the fig. S1.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S6
References

REFERENCES AND NOTES

1. T. N. Theis, H.-S. P. Wong, The end of Moore's law: A new beginning for information technology. *Comput. Sci. Eng.* **19**, 41–50 (2017).
2. C. D. Schuman, S. R. Kulkarni, M. Parsa, J. P. Mitchell, P. Date, B. Kay, Opportunities for neuromorphic computing algorithms and applications. *Nat. Comput. Sci.* **2**, 10–19 (2022).
3. N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, R. Boyle, P. L. Cantin, C. Chao, C. Clark, J. Coriell, M. Daley, M. Dau, J. Dean, B. Gelb, T. V. Ghaemmaghami, R. Gottipati, W. Gulland, R. Hagmann, C. Richard Ho, D. Hogberg, J. Hu, R. Hundt, D. Hurt, J. Ibarz, A. Jaffey, A. Jaworski, A. Kaplan, H. Khaitan, D. Killebrew, A. Koch, N. Kumar, S. Lacy, J. Laudon, J. Law, D. Le, C. Leary, Z. Liu, K. Lucke, A. Lundin, G. MacKean, A. Maggiore, M. Mahony, K. Miller, R. Nagarajan, R. Narayanaswami, R. Ni, K. Nix, T. Norrie, M. Omernick, N. Penukonda, A. Phelps, J. Ross, M. Ross, A. Salek, E. Samadiani, C. Severn, G. Sizikov, M. Snelham, J. Souter, D. Steinberg, A. Swing, M. Tan, G. Thorson, B. Tian, H. Toma, E. Tuttle, V. Vasudevan, R. Walter, W. Wang, E. Wilcox, D. H. Yoon, In-datacenter performance analysis of a tensor processing unit. *Proc. - Int. Symp. Comput. Archit.* **45**, 1–12 (2017).
4. C. Pehle, S. Billaudelle, B. Cramer, J. Kaiser, K. Schreiber, Y. Stradmann, J. Weis, A. Leibfried, E. Müller, J. Schemmel, The BrainScaleS-2 accelerated neuromorphic system with hybrid plasticity. *Front. Neurosci.* **16**, 1–21 (2022).
5. M. Davies, N. Srinivasa, T. H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, Y. Liao, C. K. Lin, A. Lines, R. Liu, D. Mathaikutty, S. McCoy, A. Paul, J. Tse, G. Venkataramanan, Y. H. Weng, A. Wild, Y. Yang, H. Wang, Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro.* **38**, 82–99 (2018).

6. M. V. Debole, B. Taba, A. Amir, F. Akopyan, A. Andreopoulos, W. P. Risk, J. Kusnitz, C. O. Otero, T. K. Nayak, R. Appuswamy, P. J. Carlson, A. S. Cassidy, P. Datta, S. K. Esser, G. J. Garreau, K. L. Holland, S. Lekuch, M. Mastro, J. Mckinstry, C. Di Nolfo, J. Sawada, B. Paulovicks, K. Schleupen, B. G. Shaw, J. L. Klamo, M. D. Flickner, J. V. Arthur, D. S. Modha, TrueNorth: Accelerating from zero to 64 million neurons in 10 years. *Computer (Long. Beach. Calif).* **52**, 20–29 (2019).
7. F. Brückerhoff-Plückelmann, J. Feldmann, H. Gehring, W. Zhou, C. D. Wright, H. Bhaskaran, W. Pernice, Broadband photonic tensor core with integrated ultra-low crosstalk wavelength multiplexers. *Nanophotonics* **11**, 4063–4072 (2022).
8. J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, H. Bhaskaran, Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
9. A. N. Tait, M. A. Nahmias, B. J. Shastri, P. R. Prucnal, Broadcast and weight: An integrated network for scalable photonic spike processing. *J. Light. Technol.* **32**, 4029–4041 (2014).
10. X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, D. J. Moss, 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
11. F. Brückerhoff-Plückelmann, J. Feldmann, C. D. Wright, H. Bhaskaran, W. H. P. Pernice, Chalcogenide phase-change devices for neuromorphic photonic computing. *J. Appl. Phys.* **129**, 151103 (2021).
12. Z. Cheng, C. Ríos, W. H. P. Pernice, C. David Wright, H. Bhaskaran, On-chip photonic synapse. *Sci. Adv.* **3**, 1–7 (2017).
13. J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, W. H. P. Pernice, All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **569**, 208–214 (2019).
14. S. G. Sarwat, F. Brückerhoff-Plückelmann, S. G. C. Carrillo, E. Gemo, J. Feldmann, H. Bhaskaran, C. D. Wright, W. H. P. Pernice, A. Sebastian, An integrated photonics engine for unsupervised correlation detection. *Sci. Adv.* **8**, 1–10 (2022).
15. M. Wuttig, H. Bhaskaran, T. Taubner, Phase-change materials for non-volatile photonic applications. *Nat. Photonics.* **11**, 465–476 (2017).
16. S. Bhattacharya, S. N. Patra, S. Mukhopadhyaya, An all optical prototype neuron based on optical Kerr material. *Optik (Stuttg).* **126**, 13–18 (2015).
17. J. R. Basani, S. Krastanov, M. Heuck, D. R. Englund, All-photonic artificial neural network processor via nonlinear optics, in *2022 Conference on Lasers and Electro-Optics*, San Jose CA, 15 to 20 May 2022 (Optica Publishing Group, 2022), pp. 1–16.
18. A. N. Tait, T. Ferreira De Lima, M. A. Nahmias, H. B. Miller, H. T. Peng, B. J. Shastri, P. R. Prucnal, Silicon photonic modulator neuron. *Phys. Rev. Appl.* **11**, 1 (2019).
19. M. A. Nahmias, B. J. Shastri, A. N. Tait, P. R. Prucnal, A leaky integrate-and-fire laser neuron for ultrafast cognitive computing. *IEEE J. Sel. Top. Quantum Electron.* **19**, 1–12 (2013).
20. J. Robertson, T. Deng, J. Javaloyes, A. Hurtado, Controlled inhibition of spiking dynamics in VCSELs for neuromorphic photonics: Theory and experiments. *Opt. Lett.* **42**, 1560–1563 (2017).
21. S. Bandyopadhyay, A. Sludds, S. Krastanov, R. Hamerly, N. Harris, D. Bunandar, M. Streshinsky, M. Hochberg, D. Englund, Single chip photonic deep neural network with accelerated training. arXiv.2208.01623 (2022). https://doi.org/10.48550/arXiv.2208.01623.
22. F. Ashtiani, A. J. Geers, F. Aflatouni, An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
23. J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, D. Brunner, Reinforcement learning in a large-scale photonic recurrent neural network. *Optica.* **5**, 756 (2018).
24. C. Ríos, N. Youngblood, Z. Cheng, M. Le Gallo, W. H. P. Pernice, C. D. Wright, A. Sebastian, H. Bhaskaran, In-memory computing on a photonic platform. *Sci. Adv.* **5**, 1–10 (2019).
25. X. Li, N. Youngblood, C. Ríos, Z. Cheng, C. D. Wright, W. H. Pernice, H. Bhaskaran, Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell. *Optica.* **6**, 1 (2019).
26. C. Rios, M. Stegmaier, P. Hosseini, D. Wang, T. Scherer, C. D. Wright, H. Bhaskaran, W. H. P. Pernice, Integrated all-photonic non-volatile multi-level memory. *Nat. Photonics.* **9**, 725–732 (2015).
27. O. K. Swanson, A. Maffei, From hiring to firing: Activation of inhibitory neurons and their recruitment in behavior. *Front. Mol. Neurosci.* **12**, 1–9 (2019).
28. A. Slowik, H. Kwasnicka, Evolutionary algorithms and their applications to engineering problems. *Neural Comput. Appl.* **32**, 12363–12379 (2020).
29. Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, S. Du, All-optical neural network with nonlinear activation functions. *Optica.* **6**, 1132 (2019).
30. B. Shi, N. Calabretta, R. Stabile, Deep neural network through an InP SOA-based photonic integrated cross-connect. *IEEE J. Sel. Top. Quantum Electron.* **26**, 1–11 (2020).
31. L. G. Wright, T. Onodera, M. M. Stein, T. Wang, D. T. Schachter, Z. Hu, P. L. McMahon, Deep physical neural networks trained with backpropagation. *Nature* **601**, 549–555 (2022).
32. W. Zhou, X. Li, N. Farmakidis, J. Tan, J. Feldmann, F. Brückerhoff-Plückelmann, C. D. Wright, W. H. P. Pernice, H. Bhaskaran, Demonstration of over 10^8 cycling endurance in the nonvolatile photonic memory cells. *Eur. Phase-Change Ovonic Symp.* **1**, 8–9 (2021).
33. W. Kim, M. Brightsky, T. Masuda, N. Sosa, S. Kim, R. Bruce, F. Carta, G. Fraczak, H. Y. Cheng, A. Ray, Y. Zhu, H. L. Lung, K. Suu, C. Lam, paper presented at Technical Digest - International Electron Devices Meeting (3 to 7 December 2016, San Francisco, CA), pp. 2–4.
34. M. Salinga, B. Kersting, I. Ronneberger, V. P. Jonnalagadda, X. T. Vu, M. Le Gallo, I. Giannopoulos, O. Cojocaru-mirédin, R. Mazzarello, A. Sebastian, Monatomic phase change memory. *Nat. Mater.* **17**, 681–685 (2018).
35. E. Gemo, S. G.-C. Carrillo, C. R. De Galarreta, A. Baldycheva, H. Hayat, N. Youngblood, H. Bhaskaran, W. H. Pernice, C. D. Wright, Plasmonically-enhanced all-optical integrated phase-change memory. *Opt. Express* **27**, 24724–24737 (2019).
36. H. Zhang, J. Thompson, M. Gu, X. D. Jiang, H. Cai, P. Y. Liu, Y. Shi, Y. Zhang, M. F. Karim, G. Q. Lo, X. Luo, B. Dong, L. C. Kwek, A. Q. Liu, Efficient on-chip training of optical neural networks using genetic algorithm. *ACS Photonics.* **8**, 1662–1672 (2021).
37. M. L. Davenport, J. E. Bowers, Efficient and broad band coupling between silicon and ultra-low-loss silicon nitride waveguides, in *2016 IEEE Photonics Conference (IPC)*, Waikoloa, HI, 2 to 6 October 2016 (IEEE, 2017), pp. 631–632.
38. Y. Ueda, Y. Saito, T. Shindo, S. Kanazawa, M. Ishikawa, High-speed tunable laser based on electro-optic effect for wavelength switching. *NTT Tech. Rev.* **20**, 65–73 (2022).
39. W. Bogaerts, P. de Heyn, T. van Vaerenbergh, K. de Vos, S. Kumar Selvaraja, T. Claes, P. Dumon, P. Bienstman, D. van Thourhout, R. Baets, Silicon microring resonators. *Laser Photonics Rev.* **6**, 47–73 (2012).
40. C. Rios, M. Stegmaier, Z. Cheng, N. Youngblood, C. D. Wright, W. H. P. Pernice, H. Bhaskaran, Controlled switching of phase-change materials by evanescent-field coupling in integrated photonics [invited]. *Opt. Mater. Express.* **8**, 2455–2470 (2018).

Acknowledgments: We thank J. Schütte for assistance with the picture in Fig. 4. **Funding:** This work was supported by Deutsche Forschungsgemeinschaft grant CRC 1459 and EXC 2181/1 – 390900948 [the Heidelberg STRUCTURES Excellence Cluster (to I.B., M.B., N.V., and W.H.P.)]; EU H2020 grants 780848, Fun-COMP and 101017237, PHOENICS (to H.B., C.D.W., and W.H.P.); and UKRI grant numbers EP/T023899/1, EP/R001677/1, and EP/W022931/1 (N.F. and H.B.). **Author contributions:** Conceptualization: F.B.-P., H.B., and W.H.P. Methodology: F.B.-P., I.B., M.B., N.V., N.F., and E.L. Investigation: F.B.-P., I.B., M.B., N.V., N.F., E.L., and F.L. Visualization: F.B.-P., I.B., M.B., N.V., and N.F. Supervision: W.H.P., H.B., C.D.W., M.S., and B.R. Writing—original draft: F.B.-P., I.B., and M.B. Writing—review and editing: All authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 25 May 2023
Accepted 19 September 2023
Published 20 October 2023
10.1126/sciadv.adi9127