

# All-optical spiking neurosynaptic networks with self-learning capabilities

J. Feldmann<sup>1</sup>, N. Youngblood<sup>2</sup>, C. D. Wright<sup>3</sup>, H. Bhaskaran<sup>2</sup> & W. H. P. Pernice<sup>1\*</sup>

Software implementations of brain-inspired computing underlie many important computational tasks, from image processing to speech recognition, artificial intelligence and deep learning applications. Yet, unlike real neural tissue, traditional computing architectures physically separate the core computing functions of **memory and processing**, making **fast, efficient and low-energy computing** difficult to achieve. To overcome such limitations, an attractive alternative is to design hardware that mimics neurons and synapses. Such hardware, when connected in networks or neuromorphic systems, processes information in a way more **analogous to brains**. Here we present an all-optical version of such a neurosynaptic system, capable of supervised and unsupervised learning. We exploit **wavelength division multiplexing techniques** to implement a scalable circuit architecture for photonic neural networks, successfully demonstrating pattern recognition directly in the optical domain. Such photonic neurosynaptic networks **promise access to** the high speed and high bandwidth inherent to optical systems, thus enabling the direct processing of optical telecommunication and visual data.

In our everyday lives, artificial neural networks are already extensively used, for example, for carrying out tasks such as face and speech recognition, which are frequently performed on our mobile phones<sup>1</sup>. In more complex applications, such as medical diagnostics<sup>2</sup> and autonomous driving<sup>1,3</sup>, high-speed data analysis will become ever more important. However, fulfilling this demand for fast and efficient processing using traditional computation techniques is problematic, owing to **speed and energy inefficiencies**<sup>4</sup>. Traditional computers are built following the von Neumann architecture, with two separate units for memory and processor operating in a sequential way, one command at a time. When compared to the massively parallel signal processing of the brain, it is clear why simulating a neural network in software on a machine based on the von Neumann architecture and limited by the transfer of data between memory and processor cannot be efficient<sup>5</sup>. The more radical approach of neuromorphic computing seeks to overcome the limitations of carrying out brain-like processing using conventional computers by developing hardware mimics of the basic building blocks of biological brains—that is, **neurons and synapses**—and combining these into suitably scaled networks and arrays. Such an approach could, for example, enable the efficient processing and analysis of data in parallel directly on-chip, thus finding widespread utility in power-critical situations such as for mobile devices and so-called ‘edge computing’ applications<sup>6</sup>.

Recently, a number of different concepts for realizing hardware (that is, neuromorphic) implementations for artificial intelligence have been proposed in the electrical domain<sup>7,8</sup> but optical approaches are very much in their infancy<sup>9–13</sup>. However, a very promising approach for photonic neuromorphic computing is based around **phase-change materials (PCMs)** and devices, because these have been shown to exhibit an intrinsic ability to provide in hardware the basic **integrate-and-fire functionality** of neurons and the plastic weighting operation of synapses<sup>14–18</sup>. A fully optical, integrated and scalable neuromorphic framework for implementing spiking neural networks using PCMs has—to the best of our knowledge—not yet been demonstrated.

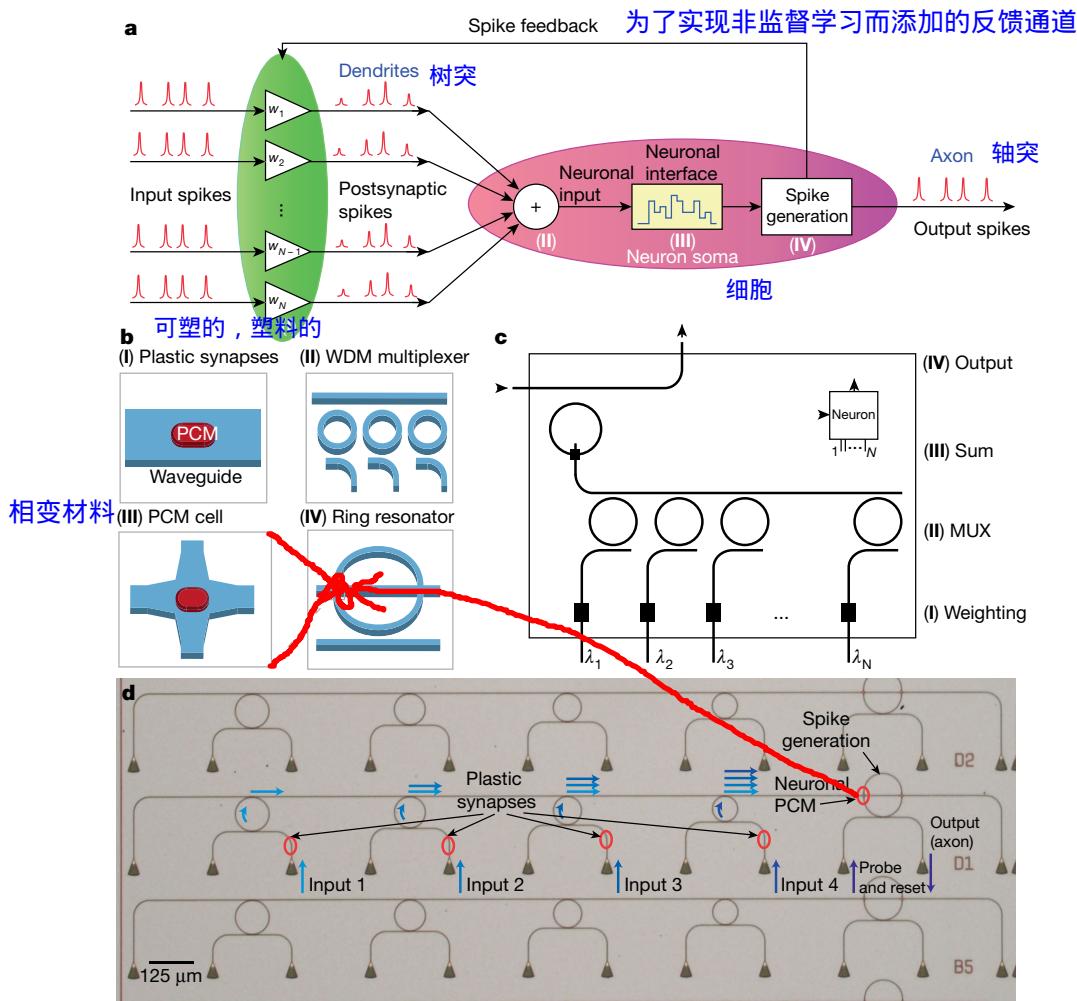
Here we propose and fabricate an **all-optical spiking neuron circuit**, with integrated all-optical synapses, and demonstrate that such a system is capable of the prototypical artificial intelligence task of pattern recognition. Moreover, training and learning in the system is implemented in both a supervised and an unsupervised way, both cases having a wide range of applications but relying on different learning rules. In the first case, where training sets with pairs of known inputs and outputs are present, supervised learning rules can be applied, such as the well-known backpropagation algorithm<sup>19,20</sup>. The second case requires unsupervised learning, meaning that the network adapts on its own to specific repeating features and patterns that are unknown beforehand<sup>21</sup>.

We devise a **scalable, layered architecture**, based on wavelength division multiplexing (WDM), to realize such a complex integrated photonic system, presenting a photonic neural network consisting of four neurons and sixty synapses (and 140 optical elements in total) that is able to successfully **recognize letters** presented to it. By implementing an all-optical spiking neural network on a nanophotonic chip, we provide a step towards optical neuromorphic systems, which will benefit from the high bandwidth and fast signalling properties that come with operating fully in the optical domain<sup>22,23</sup>.

## Photonic implementation of an artificial neuron

A sketch of our optical spiking neuron circuit, also incorporating all-optical synapses, showing how it is integrated onto a nanophotonic platform, is shown in Fig. 1a, b. The neuron represents a system comprising **N input (pre-synaptic) neurons**, **one output (post-synaptic) neuron** and **N interconnecting synapses**. Each connection between the pre-synaptic neurons and the post-synaptic neuron has a certain weight  $w_i$ . In the configuration of Fig. 1a, optical pulses from the pre-synaptic neurons are fed from the left into the connecting synapses, and thence to the **post-synaptic neuron itself**. The synapses are built of optical waveguides (see Fig. 1b) and weighting is achieved via PCM cells (here cells have an area of  $3.6 \mu\text{m}^2$  and are shown as red squares) integrated on top of the waveguides, which can modify the propagating optical

<sup>1</sup>Institute of Physics, University of Münster, Münster, Germany. <sup>2</sup>Department of Materials, University of Oxford, Oxford, UK. <sup>3</sup>Department of Engineering, University of Exeter, Exeter, UK.  
\*e-mail: wolfram.pernice@uni-muenster.de



**Fig. 1 | All-optical spiking neuronal circuits.** **a, b**, Schematic of the network realized in this study, consisting of several pre-synaptic input neurons and one post-synaptic output neuron connected via **PCM synapses**. The input spikes are weighted using PCM cells and summed up using a WDM multiplexer (MUX). If the integrated power of the post-synaptic spikes surpasses a certain threshold, the PCM cell on the ring resonator switches and an **output pulse** (neuronal spike) is generated. **c**, Photonic circuit diagram of an integrated optical neuron with symbol

block shown in the inset (top right). Several of these blocks can be connected to larger networks using the wavelength inputs and outputs (see Fig. 5). **d**, Optical micrograph of three fabricated neurons (B5, D1 and D2), showing four input ports. The four small ring resonators on the left are used to couple light of different wavelengths from the inputs to a single waveguide, which then leads to the PCM cell at the crossing point with the large ring. The triangular structures on the bottom are grating couplers used to couple light onto and off the chip.

mode in a controlled manner. PCMs are frequently used in re-writable optical disk technologies, such as Blu-ray RE, and exhibit a large contrast in the absorption of light between their amorphous and crystalline states (phases) of matter<sup>24–26</sup>.

#### 不规则的

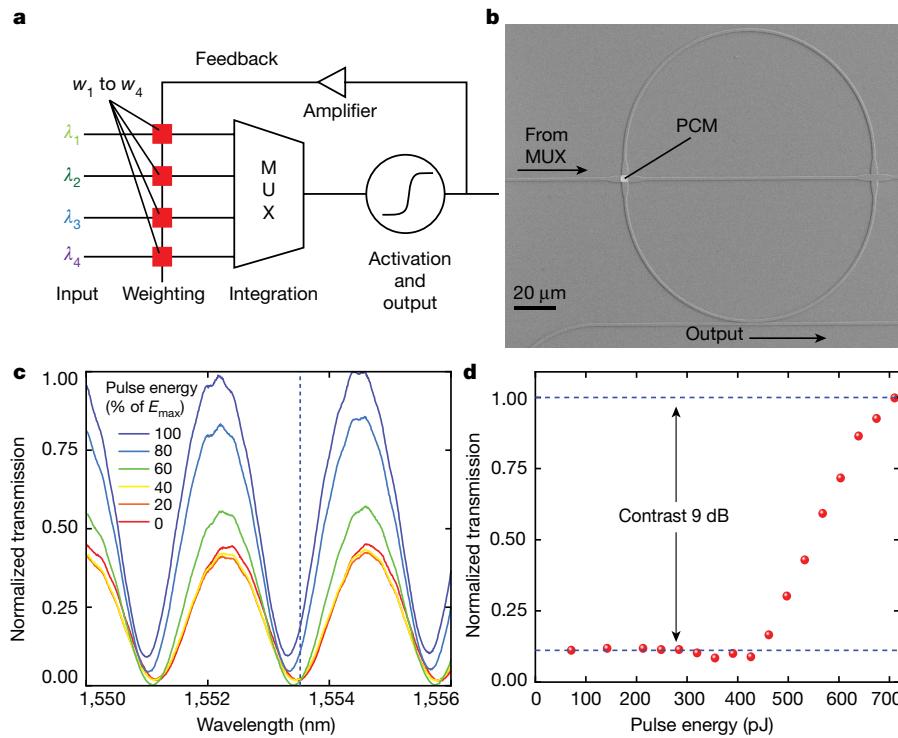
With the PCM cell in the **amorphous** state, the synaptic waveguide is highly transmissive, representing a strong connection between two neurons. In the **crystalline** 颗粒的 state, however, most of the light is absorbed, leading to a weak connection. After the input pulses have been weighted, they are combined into a single waveguide using wavelength division multiplexing (WDM)<sup>27,28</sup> and guided to the (output) spiking neuron circuit. This is composed of a ring resonator with its own integrated PCM cell that can be switched (between the crystalline and amorphous states) by the incoming combined pulses. Switching the neuronal PCM cell in turn changes the optical resonance condition of the ring and its propagation loss.

When the neuronal PCM cell is in the crystalline state, a suitable probe pulse sent along the ‘output’ waveguide couples strongly into the ring resonator and so no output pulse (spike) will be observed. However, if the instantaneous combined power of the weighted input pulses from the pre-synaptic neurons is high enough to switch the neuronal PCM cell to its amorphous state, the probe pulse is no longer on resonance with the ring and will be transmitted past the ring, thus

generating an output neural spike. As the switching of the PCM cell only occurs above a certain threshold power, the neuron only generates an output pulse (spike) if the weighted sum of the input power exceeds this threshold. Thus, the system naturally emulates the basic integrate-and-fire functionality of a biological neuron, with the distinction that the artificial neuron integrates over the optical power at a fixed time, as opposed to its biological counterpart, which integrates incoming pulses over time. This artificial neuron, shown in photonic circuit form in Fig. 1c, serves as a building block in layered photonic networks (described below) suited to the scalable implementation of neurosynaptic systems.

The neurosynaptic system described so far provides the basic structure needed for supervised learning tasks, where the weights of the inputs are set by an external supervisor. However, in order to be capable of unsupervised learning too, we add a **feedback waveguide channelling** parts of the neuron’s output spike back to the synaptic PCM cells. In this way (and as described in more detail below), the connections from all **inputs that contributed to a particular output spike will be enhanced, while those that did not contribute will be weakened—or, in Hebbian terms, “neurons that fire together, wire together”**<sup>29</sup>.

Figure 1d shows an optical micrograph of the actual implementation ( $\times 3$ ) of a single-neuron neurosynaptic system, fabricated via



**Fig. 2 | Spike generation and operation of the artificial neuron.**

a, Schematic of the photonic implementation of a phase-change neuron circuit. Light of different wavelengths is weighted by phase-change elements  $w_1$  to  $w_4$  and summed up by a multiplexer to a single waveguide. If this activation energy surpasses a threshold, an output pulse is generated, and the weights are updated. b, Scanning electron micrograph of a ring resonator used to implement the activation function. By switching the PCM cell on top of the waveguide crossing, the resonance condition of the resonator can be tuned. The waveguide on the bottom of the ring is used to probe the resonance and generate an output pulse.

c, Normalized transmission measurement (divided by the measured maximum transmission value) of the device in b and its dependence on pulse energy. The resonance shifts towards shorter wavelength with increasing pulse energy sent to the PCM cell on the ring. At the same time, the transmission increases because of reduced absorption in the PCM cell and thus changes the coupling between ring and waveguide. d, Normalized transmission to the output at a fixed wavelength (dashed line in c), showing the activation function used to define the firing threshold of the neuron.

## 氮化物

electron-beam lithography on a platform consisting of a layer of silicon nitride on a layer of silicon oxide. Several input waveguides each with a synaptic PCM cell on top (indicated by red ovals) are fed to the upper waveguide using small ring resonators as a simple multiplexing device (the ring resonators have four different radii increasing linearly from 40  $\mu\text{m}$  to 55  $\mu\text{m}$ , an optical quality factor ( $Q$  factor) value of around 10,000 and provide an insertion loss of 1.5 dB; see Supplementary Information section 5 for more details). This upper waveguide then leads the light to the neuronal (large) ring resonator (radius 60  $\mu\text{m}$ ) with its own integrated PCM cell (area 9  $\mu\text{m}^2$ ). Probe pulses sent to the waveguide lying below the neuronal ring resonator either couple to it or generate an output spike depending, as described previously, on whether the neuronal PCM cell is in the crystalline or amorphous state. The light is coupled onto and off the chip using grating couplers, which provide access to multiple optical fibres in the measurement setup.

## Optical performance of a single neuron device

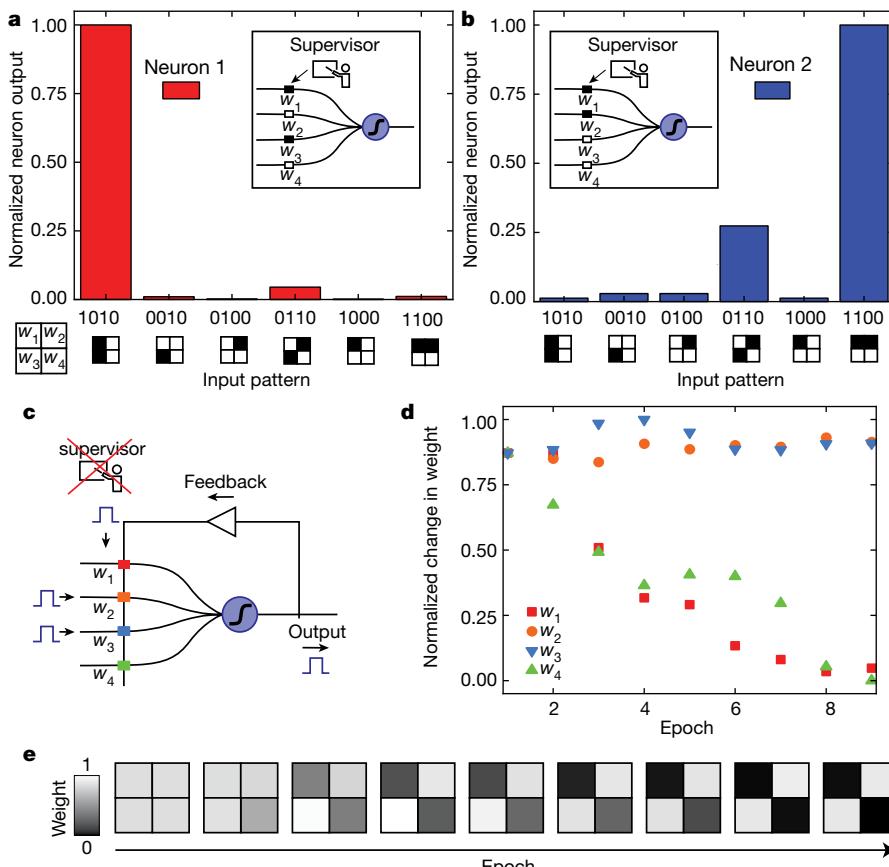
Figure 2a shows in detail the photonic signal processing and optical operation of a single neuron. Each input pulse is sent on a different wavelength  $\lambda_i$  and is first partly absorbed (weighted) by the relevant synaptic PCM cell. After weighting, the individual input waveguides are combined with a multiplexer to a single waveguide, summing up the input powers. If this power is high enough to switch the neuronal PCM cell of the large ring resonator (see Fig. 2b), an output spike is generated and in the case of unsupervised learning the synaptic weights are adjusted using the feedback loop. In this specific device example an output pulse is generated if the summed power here exceeds 430 pJ (see Supplementary Information section 9). In Fig. 2b, a more detailed scanning electron micrograph of the ring resonator used to deliver the spiking neuron function can be seen. The neuronal PCM cell used

to tune the resonance condition is deposited on top of a waveguide crossing specially designed for low optical losses<sup>30</sup> (0.23 dB). The second waveguide crossing (without a PCM cell) is used only for testing purposes, and allows a bias input.

To characterize the integrate-and-fire type response (or activation function) of our neuron circuit, several transmission spectra for the resonator were obtained after sending pulses of different energy to the neuronal PCM cell via the crossing waveguide (Fig. 2c; see Methods). By plotting the transmission of the ring resonator after different excitation pulses versus the pulse energy at a fixed wavelength, an activation function as shown in Fig. 2d was obtained. Depending on which wavelength is chosen, the contrast and maximum transmission level of the output function can be adjusted. Figure 2d shows the activation at 1,553.4 nm (that is, at the dashed line in Fig. 2c), representing the operation with the highest contrast (of 9 dB) between the output states. It can clearly be seen that a substantial output was generated only above a threshold energy of 60% of the maximum pulse energy. This non-linear response, resembling the rectified linear unit (ReLU) function, is crucial for neural activation functions, because it projects complex input data to higher dimensions, enabling linear separation by the output neurons<sup>31</sup>.

## Supervised and unsupervised learning

Having realized an all-optical spiking neuron, supervised learning tests were carried out. In this case, the synaptic weights of the network are set by an external supervisor (as, for example, done in software-based artificial neural networks using the backpropagation algorithm). Here, a training set of data consisting of pairs of input patterns and the expected output is shown to the network. Depending on the deviation between the expected and actual output, the synaptic weights are adjusted in an



**Fig. 3 | Supervised and unsupervised learning with phase-change all-optical neurons.** **a** and **b** show the neuron output of two individual neurons when presented with different input patterns. Neuron 1 learned to recognize pattern '1010', while neuron 2 generates an output signal when '1100' is shown. In this example, the eight weights of the neural network were set by an external supervisor. **c**, Schematic illustrating the unsupervised learning mechanism in an all-optical neuron. If an output spike is generated, the synaptic weights for which the input and the feedback pulses overlap in time are potentiated (the weight is increased),

optimization process until the solution is best approximated and the network is trained.

The experimental neural network used was composed of two single neurosynaptic systems (of the type shown in Fig. 1d), each consisting of **four input** (pre-synaptic) neurons connected to one output (post-synaptic) neuron by four PCM synapses. The weights of the first neuron were set to the pattern '**1010**', meaning that the first and third PCM synapses were in a high transmission state (contributing substantially to the activation energy) and the second and fourth were in a low transmission state (contributing less to the activation energy). The second neuron was trained in the same way to the pattern '**1100**'.

In Fig. 3a, b, the post-synaptic neuron output is plotted as a function of the input pattern. It can clearly be seen that in both cases (that is, for both input patterns) the neurons were trained successfully and, from the neuron's output it is easy to determine which pattern was presented to the network. Using only two output neurons on the same set of input neurons, our all-optical neuromorphic system can already solve simple image recognition tasks. By increasing the number of inputs per neuron and the number of neurons, more complex images can be processed and more difficult tasks, such as letter (or digit) recognition or language identification can be solved using the same basic approach, as we show below (in the experiments for letter recognition, and by simulation for digit recognition and language identification).

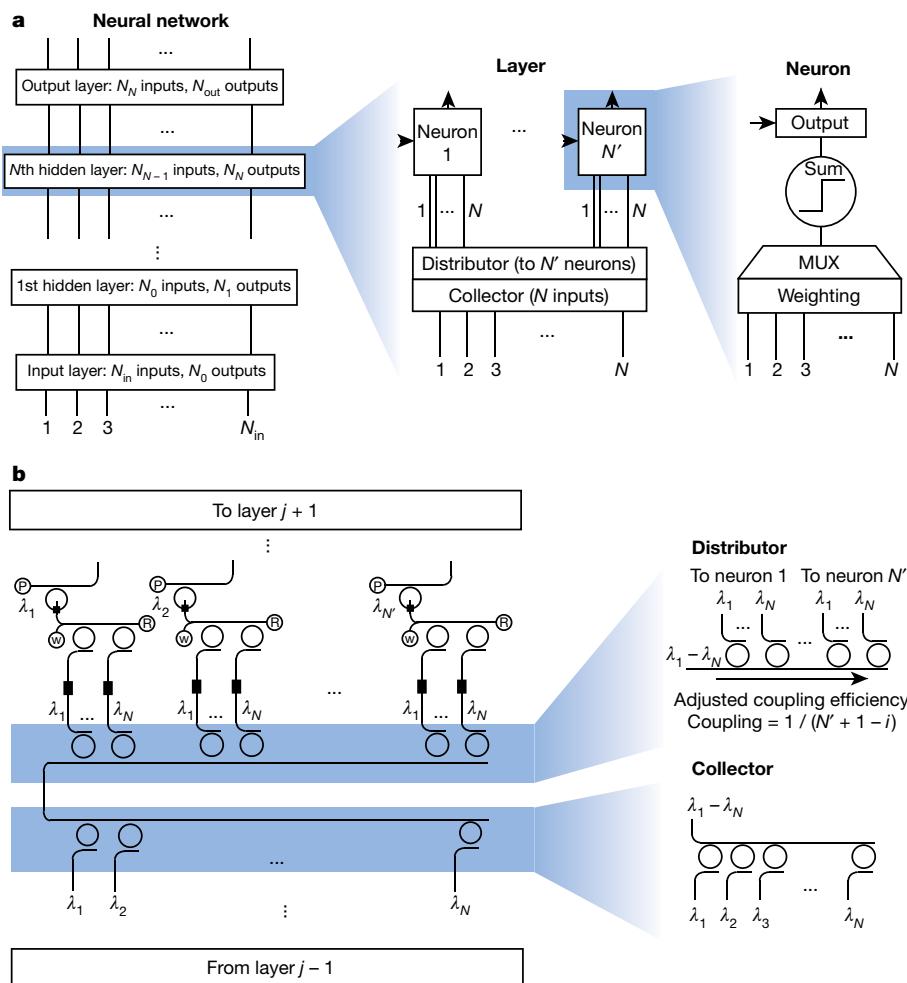
We have thus illustrated an example of supervised learning. This learning technique is feasible for many tasks but has the limitation that a training set with tuples of input patterns and expected outputs

whereas the weights that are only hit by the single feedback pulse are depressed. **d**, Changes in the four synaptic weights over time when the pattern '0110' is repeatedly shown, starting from fully amorphous (high transmitting) weights. The weights for which the input and feedback pulse overlap remain almost constant over several epochs. The other weights, for which only the feedback pulse is shown, decrease continuously. **e**, Changes in the weights over time, clarifying that the information of the pattern is encoded in the weights.

must be present. If the output is unknown, for example, if an unknown repeating pattern must be discovered in a data stream, then supervised learning is not applicable and unsupervised learning procedures are necessary. In an unsupervised approach, the network updates its weights on its own and in this way adapts to a certain pattern over time, without the need for an external supervisor.

To do this, **an update rule** needs to be defined. A common concept in unsupervised learning is spike-timing-dependent plasticity following **Hebb's postulate**<sup>29</sup>. Here the change in the synaptic weight after an output spike depends on the **relative timing between the input and the output spike of a neuron (the timing difference between pre- and post-synaptic neuron firings)**. If an input signal arrives just before an output spike was generated, that input signal is likely to have contributed to reaching the firing threshold and the corresponding weight will be increased. If the input pulse arrives after the output spike occurred, the synaptic weight will be decreased. The amount of potentiation (weight increase) or depression (weight decrease) is a function of the time difference between the input and output spikes, as described in a previous study<sup>32</sup>.

A similar but simplified learning rule is applied in our all-optical neuron approach. The timing between incoming pulses and output pulses in our case is **fixed** (because we operate the neuron in a clocked way, with one complete input pattern per time step; see Methods), so there is no varying time delay between input and output events. We therefore adopt a simplified learning rule, **increasing the synaptic weights of all inputs that contributed to a spike generation, and**



**Fig. 4 | Scaling architecture for all-optical neural networks.** **a**, The general neural network is composed of an input layer, an output layer and several hidden layers. Each of these layers consists of a collector gathering the information from the previous layer, a distributor that splits the signal equally to individual neurons and the neuronal and synaptic elements of the layer itself. Each neuron has a weighting unit and a multiplexer to calculate the weighted sum of the inputs. The sum is then fed to an activation unit which decides whether an output pulse is generated.

**decreasing the weights of all that did not.** Experimentally, we obtain this behaviour by overlapping (in time) the output pulses with the input pulses (see Methods).

Figure 3d, e shows the development over time of the four synaptic weights during unsupervised learning by a single neuron. Initially, all the PCM synapses are in the amorphous (high transmittance) state. When the input pattern '0110' is repeated, the neuron adapts to it over time, until finally the neuron has learned this pattern without any intervention from an external supervisor. The neuron is now specialized to recognize this particular pattern. From Fig. 3d it is clear that the weights w<sub>2</sub> and w<sub>3</sub>, corresponding to inputs three and four, stay almost constant over time, as the overlapping input and feedback pulses preserve their amorphous state. By contrast, weights w<sub>1</sub> and w<sub>4</sub> are depressed stepwise with each epoch.

### Scalability of the photonic artificial neural networks

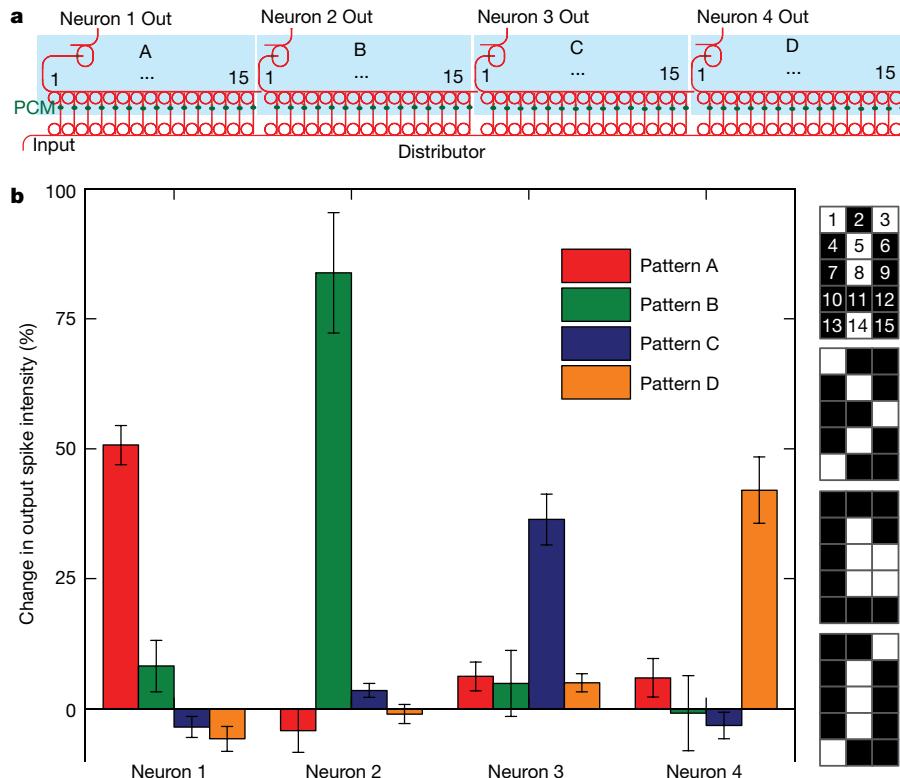
Having successfully demonstrated a single-neuron neurosynaptic system as a fundamental building block for photonic neural networks, we then developed a way of connecting these artificial neurons into larger networks. We implemented an architecture exploiting individually addressable, interlinked photonic layers, as shown schematically in Fig. 4a. The whole network consists of an input and an output layer, which are optically connected via N hidden layers. Each hidden layer

**b**, Photonic implementation of a single layer from the network. The collector unites the optical pulses from the previous layer using a WDM multiplexer. A distributor made from the same rings as the collector but with adjusted coupling efficiency distributes the input signal equally to the PCM synapses of each neuron. The letters P, W and R denote the input ports used to probe the output, set the weights and return the neuronal PCM to its initial state.

takes the output of the previous layer as an input and passes its outputs to the next layer. The input layer is the optical interface to the real world, taking the data to be processed and distributing it to the next level in the network.

A single layer of the network consists of a collector, a distributor and its neurosynaptic elements. The collector gathers all the outputs from the previous layer, which are then equally distributed to the N neurons within the layer (fully connected network) by the distributor. The photonic neurons themselves operate as described above: a phase-change synapse weights the inputs and a WDM multiplexer builds the sum, which is passed to the activation unit that decides whether a neuronal output pulse is transmitted. In this architecture, each layer is addressed optically with its own waveguide for generating the probe signal. Therefore, the optical power in the layer is not limited by the transmitted optical response from a previous layer.

Figure 4b describes how the constituent parts of a layer translate into the actual photonic circuit. The outputs from a previous layer are multiplexed onto a single waveguide using ring resonators (thus building the collector). This signal is then equally distributed to the neurons within this layer, again using ring resonators for demultiplexing (thus building the distributor). By choosing the gap between the feeding waveguide and ring resonator, the coupling efficiency can be tuned (see Supplementary Fig. 8). Following the formula for the coupling



**Fig. 5 | Experimental realization of a single-layer spiking neural network.** **a**, The device consists of four photonic neurons, each with 15 synapses. Each synapse corresponds to a pixel in a  $3 \times 5$  image (see **b**) and is encoded in the wavelengths corresponding to the ring multiplexers (see numbering in **b**). The full device comprises an integrated photonic circuit

built up from 140 optical components. **b**, The change in output spike intensity is shown for the four trained patterns illustrated on the right-hand side. The neural network successfully recognizes the four patterns as each neuron only responds (spikes) to one of the patterns. The error bars denote the standard deviation for  $n = 5$ .

efficiency  $c_{\text{eff},i} = 1/(N' + 1 - i)$  with  $N'$  neurons on the layer and the neuron position  $i$  then, for example, in a layer with four neurons, this means that  $1/(4 + 1 - 1) = 1/4$  of the light is coupled to the first neuron,  $1/3$  of the remaining light is passed to the second neuron,  $1/2$  to the third and the residual light to the fourth neuron (that is, each of the four neurons receives  $1/4$  of the light). The circuit diagram of the actual photonic neuron, the neuro-synaptic system, is shown in Fig. 1c and is the same as used in the experiments described above. The output pulses of a layer can then be connected to the collector of the next layer.

We note that when using the above approach no waveguide crossings are needed for distributing the signal to the neurons, thus preventing crosstalk and losses. Because the output pulses are generated for each layer individually, there is also no accumulation of errors and signal contamination over subsequent layers. This fact also simplifies the timing of the network as each layer can be processed step by step. First, the input pulses are sent, and the activation units are switched where appropriate. Second, the output probe pulses are sent and transmitted to the next layer (if the threshold for neuron switching was reached). In a final step, the PCM cells on the rings must be returned to their initial state.

### Realization of a single-layer neurosynaptic system

Figure 5 shows the experimental implementation of a full layer of the proposed neural network design, consisting of four neurons with 15 synapses each. The full device is composed of more than 140 optical elements; optical micrographs of the photonic circuit are presented in Supplementary Information section 2. This network is capable, for example, of differentiating between four 15-pixel images, here representing the four letters A, B, C and D. In this system, the neurons are optically fed via an integrated WDM distributor with 15 ring resonators per neuron, while the collector is implemented off-chip using fibre-based WDM components (see Methods). As desired, all four output

neurons are only activated when the learned pattern is shown (Fig. 5b): neuron 1 fires only if pattern 'A' is shown, neuron 2 reacts only to pattern 'B', and so forth. Thus, the network is able to successfully classify the four 15-pixel images. (A more complex exemplar task of language recognition, using a larger network with the same architecture, is discussed in the Supplementary Information.) We note that in the all-optical implementation of this architecture, all artificial neurons need to be recrystallized after each spiking event. Therefore the number of operation cycles is eventually limited by the endurance of the PCM cells. Although individual PCM devices in endurance experiments have already shown  $10^{12}$  switching cycles<sup>16</sup>, further improvements in material design and device engineering are needed for high-speed and long-term switching operation.

Integrated phase-change photonic networks, designed and implemented as described above, are capable of simple pattern recognition tasks and can adapt to specific patterns. When operated with a waveguide feedback loop, they are capable of learning without an operator, and can do this in a non-volatile fashion using PCMs. The large contrast in absorption of light between their amorphous and crystalline states of matter makes PCMs attractive for use as a synaptic weighting mechanism. Compared to conventional computers that can only simulate the parallelism of neural networks, our all-optical neurons are intrinsically suited for mimicking biological neural networks. Our proposed neurons could operate several orders of magnitude faster than biological neural networks, thus enabling large amounts of data to be handled in a short amount of time. Working exclusively in the optical domain, the spiking neurosynaptic network benefits from the high bandwidth and fast data transfer rates intrinsic to light. Moreover, using a layered circuit architecture, our network could be scaled (with foundry processing) to more complex systems. In this way off-chip components (used here for experimental expediency only) such as laser sources, optical amplifiers and modulators can also be integrated into

a full system. Our integrated design holds promise for realizing all-optical neural networks capable of addressing the upcoming challenges of big data and deep learning.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-019-1157-8>.

Received: 6 May 2018; Accepted: 21 March 2019;

Published online 8 May 2019.

- Lane, N. D. et al. Squeezing deep learning into mobile and embedded devices. *IEEE Pervasive Comput.* **16**, 82–88 (2017).
- Amato, F. et al. Artificial neural networks in medical diagnosis. *J. Appl. Biomed.* **11**, 47–58 (2013).
- Nawrocki, R. A., Voyles, R. M. & Shaheen, S. E. A mini review of neuromorphic architectures and implementations. *IEEE Trans. Electron Dev.* **63**, 3819–3829 (2016).
- Preissl, R. et al. Compass: a scalable simulator for an architecture for cognitive computing. In *Proc. Int. Conf. on High Performance Computing, Networking, Storage and Analysis (SC '12)* <https://doi.org/10.1109/SC.2012.34> (IEEE, 2012).
- von Neumann, J. *The Computer and the Brain* (Yale Univ. Press, 1958).
- Wu, H., Yao, P., Gao, B. & Qian, H. Multiplication on the edge. *Nat. Electron.* **1**, 8–9 (2018).
- Furber, S. Bio-inspired massively-parallel computation. In *Advances in Parallel Computing Vol. 27 Parallel Computing: On the Road to Exascale* (IOS Press, 2018).
- Schmitt, S. et al. Neuromorphic hardware in the loop: training a deep spiking network on the BrainScaleS wafer-scale system. In *Proc. Int. Joint Conf. on Neural Networks 2227–2234* <https://ieeexplore.ieee.org/document/7966125> (2017).
- Shen, Y. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photon.* **11**, 441–446 (2017).
- Vinckier, Q. et al. High-performance photonic reservoir computer based on a coherently driven passive cavity. *Optica* **2**, 438–446 (2015).
- Brunner, D., Soriano, M. C., Mirasso, C. R. & Fischer, I. Parallel photonic information processing at gigabyte per second data rates using transient states. *Nat. Commun.* **4**, 1364–1367 (2013).
- Ferreira de Lima, T., Shastri, B. J., Tait, A. N., Nahmias, M. A. & Prucnal, P. R. Progress in neuromorphic photonics. *Nanophotonics* **6**, 577–599 (2017).
- Vandoorne, K. et al. Experimental demonstration of reservoir computing on a silicon photonics chip. *Nat. Commun.* **5**, 3541 (2014).
- Cheng, Z., Ríos, C., Pernice, W. H. P., Wright, C. D. & Bhaskaran, H. On-chip photonic synapse. *Sci. Adv.* **2**, e1700160 (2017).
- Kim, S. et al. NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning. In *Proc. International Electron Devices Meeting (IEDM) 17.1.1–17.1.4* (IEEE, 2015).
- Kuzum, D., Jeyasingh, R. G. D., Lee, B. & Wong, H. P. Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing. *Nano Lett.* **12**, 2179–2186 (2012).
- Wright, C. D., Liu, Y., Kohary, K. I., Aziz, M. M. & Hicken, R. J. Arithmetic and biologically-inspired computing using phase-change materials. *Adv. Mater.* **23**, 3408–3413 (2011).
- Pantazi, A., Woźniak, S., Tuma, T. & Eleftheriou, E. All-memristive neuromorphic computing with level-tuned neurons. *Nanotechnology* **27**, 355205 (2016).
- Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
- Burr, G. W. et al. Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. *IEEE Trans. Electron Dev.* **62**, 3498–3507 (2015).
- Le, Q. V. et al. Building high-level features using large scale unsupervised learning. In *29th Int. Conf. on Machine Learning* <https://dl.acm.org/citation.cfm?id=3042641> (2012).
- Alduino, A. & Paniccia, M. Interconnects: wiring electronics with light. *Nat. Photon.* **1**, 153–155 (2007).
- Sun, C. et al. Single-chip microprocessor that communicates directly using light. *Nature* **528**, 534–538 (2015).
- Wuttig, M. & Yamada, N. Phase-change materials for rewriteable data storage. *Nat. Mater.* **6**, 824–832 (2007).
- Raoux, S., Xiong, F., Wuttig, M. & Pop, E. Phase change materials and phase change memory. *MRS Bull.* **39**, 703–710 (2014).
- Burr, G. W. et al. Recent progress in phase-change memory technology. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **6**, 146–162 (2016).
- Tait, A. N., Nahmias, M. A., Shastri, B. J. & Prucnal, P. R. Broadcast and weight: an integrated network for scalable photonic spike processing. *J. Lightwave Technol.* **32**, 3427–3439 (2014).
- Tait, A. N. et al. Neuromorphic silicon photonic networks. *Sci. Rep.* **7**, 7430 (2016).
- Hebb, D. *The Organization of Behaviour* (Wiley, 1949).
- Feldmann, J. et al. Calculating with light using a chip-scale all-optical abacus. *Nat. Commun.* **8**, 1256 (2017).
- Hahnloser, R. H. R., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J. & Seung, H. S. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* **405**, 947–951 (2000).
- Bi, G. & Poo, M. Synaptic modification by correlated activity: Hebb's postulate revisited. *Annu. Rev. Neurosci.* **24**, 139–166 (2001).

**Acknowledgements** This research was supported by EPSRC via grants EP/J018694/1, EP/M015173/1 and EP/M015130/1 in the UK and the Deutsche Forschungsgemeinschaft (DFG) grant PE 1832/5-1 in Germany. W.H.P.P. acknowledges support by the European Research Council through grant 724707. We acknowledge funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 780848 (Fun-COMP).

**Reviewer information** *Nature* thanks Geoffrey Burr, Ingo Fischer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

**Author contributions** W.H.P.P., H.B. and C.D.W. conceived the experiment. J.F. fabricated the devices (with assistance from N.Y.). N.Y. performed the deposition of the GST material. J.F. implemented the measurement setup and carried out the measurements (with help from N.Y.). All authors discussed the data and wrote the manuscript together.

**Competing interests** The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-019-1157-8>.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

**Correspondence and requests for materials** should be addressed to W.H.P.P.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

## METHODS

**Device fabrication.** The nanophotonic circuits used in this study are realized using electron-beam lithography with a 100-kV system (Raith EBPG 5150). In a first step, opening windows for lift-off processing of alignment markers made from gold are exposed in the positive tone resist polymethylmethacrylate (PMMA) on a silicon wafer (Rogue Valley Microdevices) with a 3,300-nm silicon oxide and 344-nm silicon nitride layer on top. After development in 1:3 methyl isobutyl ketone (MIBK):isopropanol for 2 min, a stack of 5-nm-thick chromium, 120-nm-thick gold and 5-nm-thick chromium layers are evaporated via electron-beam physical vapour deposition. The lift-off step to remove the PMMA is performed in acetone, leaving the gold markers for the alignment in the next electron-beam lithography steps.

In the second lithography step the photonic structures are defined. TI Prime is used as an adhesion agent for the negative-tone electron-beam resist ma-N 2403. The photonic circuitry is developed in MF-319 for 60 s and afterwards placed on a hotplate at 105 °C for 2 min of reflow processing to reduce surface roughness. By reactive ion etching in a CHF<sub>3</sub>/O<sub>2</sub> plasma, the resist mask is transferred into the sample until the silicon nitride is fully etched. The remaining resist on the structures is removed in an oxygen plasma for 10 min.

The last electron-beam lithography step consists of writing windows for the deposition of the PCM and is executed in the same way as defining the marker windows. After development, 10 nm of the PCM GST are sputter-deposited and covered by a 10-nm-thick film of indium tin oxide to prevent oxidation of the GST. The GST and indium tin oxide capping layers were deposited using radio-frequency sputtering with an argon plasma (5 mtorr working pressure, 15 standard cubic centimetres per minute (sccm) Ar, 30 W radio-frequency power, and a base pressure of  $2 \times 10^{-6}$  torr). Finally, the GST is crystallized on a hot plate for about 10 min at 210 °C. The photonic circuits are composed of single mode waveguides at 1,550 nm with a width of 1.2 μm.

**Measurement setup.** The experimental setup used to operate the all-optical neurons comprises pattern generation and read-out of the individual weights, as sketched in the Supplementary Information. The optical read-out is achieved via transmission measurement using a continuous-wave laser (Santec, TSL 510) and four low-noise photodetectors D1 to D4 (New Focus, Model 2011) that are monitored on a computer. To couple light efficiently onto the chip, an optical fibre array is aligned with respect to the on-chip grating couplers to provide multi-port input. For optimal coupling efficiency to the chip, the polarization is optimized with a set of polarization controllers.

Pattern generation as input for the on-chip neuron is accomplished with four continuous-wave lasers set to different wavelengths matching the on-chip multiplexer. Off-chip the four light paths are combined using a fibre multiplexer and desired optical pulses are created using an electro-optical modulator and a computer-controlled electrical pulse generator (Agilent, HP 8131A). After amplifying the pulses with an erbium-doped fibre amplifier (Pritel) the pulses are de-multiplexed again and guided to the on-chip device. Using circulators, the pump and probe light is counter-propagating through the device, enabling efficient separation of the beam paths. Arbitrary input patterns are then selected by switching the shutters of the pump lasers on and off.

Similar to the input pulses, the output pulses are also created from a continuous-wave laser in combination with an electro-optical modulator and amplified by an erbium-doped fibre amplifier. A small portion of the output is measured with detector D0, while the remaining light is amplified and sent to ports F1 to F4 as a feedback pulse for weight adjustments. Turning off the feedback amplifier puts the device in a supervised learning mode. The setup used for the four-neuron network is shown in Supplementary Fig. 2.

**Activation unit.** To obtain the ReLU function shown in Fig. 2d, pulses with different energies up to approximately 700 pJ were sent to the ring resonator resembling the activation unit. Because of changes introduced in the structure of the PCM on the waveguide crossing, the transmission spectrum undergoes a considerable change between the initial crystalline state (zero pulse energy) and the final amorphous state (maximum pulse energy). This is partially explained by the resonance wavelength (at which the light is coupled into the ring and therefore the transmission is minimal) slightly shifting to shorter wavelength with increasing amount of amorphization (which causes a change in the real part of the refractive index of the PCM, leading to a slightly shifted resonance condition; see Supplementary Information). However, we also observe that the minimum transmission after amorphizing the PCM cell is much higher. This is a combined effect of the change in the imaginary part of the refractive index (absorption) and a change in the extinction ratio of the resonator. The lower absorption in the amorphous phase clearly leads to higher transmission but, equally importantly, changing the loss per round trip in the ring affects the coupling between ring and waveguide and therefore alters the extinction ratio.

**Estimation of the energy balance.** The maximum pulse energy used to switch the PCM cell in this experiment was 710 pJ, employing optical pulses of width

200 ns. For setting the weights, similar energies are used. In the experiments we operate our neurons with relatively long optical pulses in order to implement two-pulse switching<sup>30</sup>, which relies on overlapping pulses in the time domain. As the output pulse was amplified off-chip and the optical path was therefore relatively long (compared to on-chip waveguides), longer pulses had to be applied to ensure the overlap. We note, however, that lower switching energies can be employed by moving towards picosecond pulses<sup>30,33</sup>. It is also important to note that this operation scheme does not require continuous energy input to maintain the state of the PCM weights, owing to their non-volatile response. Therefore, the energy budget per neuron to perform one operation is given by the switching energy for the ring resonator plus the energy required to return the PCM element to its original fully crystalline state.

**Update of weights in unsupervised learning.** Via a feedback waveguide, the neuron's output spike is guided back to the synaptic PCM cells (see Fig. 3c). If the input pulses are now long enough in time such that they overlap with the feedback pulse at the waveguide crossing where the synaptic PCM cell is located, then the overlapping pulses have enough energy to switch the synapse into its low-absorbing amorphous state (weights  $w_2$  and  $w_3$  in Fig. 3c). A feedback pulse that encounters a synapse without an input pulse will partly crystallize the corresponding PCM cell because of the lower pulse energy, and therefore decrease the weight (weights  $w_1$  and  $w_4$ ). Owing to the properties of the PCM used, the PCM cells can only be amorphized in a single step<sup>33</sup>, meaning that if the neuron fires, all contributing inputs will always be potentiated completely. By contrast, full crystallization can be achieved in several steps<sup>30</sup> and the weights can correspondingly be decreased stepwise. Successful unsupervised learning can be accomplished using such weight update rules, as we show experimentally in Fig. 3 for small-scale systems and in Supplementary Information section 10 by simulation for larger-scale systems.

**Image encoding and implementation of the WDM distributor and collector.** To feed a certain pattern to the neural network, it has to be encoded in optical pulse patterns, which are presented to the on-chip network. The images shown in Fig. 5b (corresponding to the letters A to D) are encoded in the following way: each pixel corresponds to the resonance wavelength of one of the ring resonators within a neuron, as indicated by the numbers superimposed on the pixels in Fig. 5b. These wavelengths are aligned to WDM channels 27 to 41 in the telecommunication C-band. In the experiment we present the 'white' pixels to the network such that the pulse pattern corresponding to an 'A' is, for example, represented by an optical pulse consisting of wavelengths 1, 3, 5, 8 and 14. These wavelengths are multiplexed onto the input waveguide (as described in more detail in the Supplementary Information) and equally split to the synapses of the four neurons by the distributor. After adjusting the synapses (PCM cells) corresponding to the patterns 'A', 'B', 'C' and 'D' using optical pulses, as described previously for the single neuron (Fig. 3), the four different pulse patterns are sent to the input of the device. Subsequently, the change in the output spike intensity is observed for all four neurons, as shown in Fig. 5b.

**Simulation of language identification with a two-layer network.** Using the scalable architecture, we further simulate the performance of a scaled-up, multilayer version of the network of Fig. 5a in order to carry out a much more complex task of language identification. The network for this task consists (as shown in Supplementary Fig. 16a, b) of four input neurons, three hidden-layer neurons and two output neurons. The network is assembled using the scaling architecture described in Fig. 4 and built up using model representations of photonic neurons according to the measured experimental data (activation function shown in Fig. 2d)). This particular network is then used to detect whether the language of a given input text is either English or German (the sample texts are taken from ref.<sup>34</sup>). In a first step the ratio of each vowel (a, e, i, o and u) and the total number of characters in the input text is calculated (preprocessing). In a second step the five ratios obtained are fed to the inputs of the neural network and the outputs are computed. Each neuron in the simulated network uses the measured optical response from the on-chip neurons (see Supplementary Information section 10). With a count of about 35 words in the input text, an accuracy above 90% for the language detection is already attained. With 150 words the accuracy reaches 99.6%.

## Data availability

All data used in this study are available from the corresponding author upon reasonable request.

33. Ríos, C. et al. Integrated all-photonic non-volatile multi-level memory. *Nat. Photon.* **9**, 725–732 (2015).
34. Goldhahn, D., Eckart, T. & Quasthoff, U. Building large monolingual dictionaries at the Leipzig Corpora Collection: from 100 to 200 languages. In *Proc. 8th Int. Conf. on Language Resources and Evaluation (LREC'12)* [http://www.lrec-conf.org/proceedings/lrec2012/pdf/327\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf) (2012).