OPTICA

Monadic Pavlovian associative learning in a backpropagation-free photonic network

James Y. S. Tan,[1,†] Zengguang Cheng,[1,2,†] Johannes Feldmann,[1] Xuan Li,[1]
Nathan Youngblood,[1,3] Utku E. Ali,[1] C. David Wright,[4] Wolfram H. P. Pernice,[5,6]
and Harish Bhaskaran[1,*]

[1] Department of Materials, University of Oxford, Parks Road, Oxford, OX1 3PH, UK
[2] Current address: State Key Laboratory of ASIC and System, School of Microelectronics, Fudan University, Shanghai, 200433, China
[3] Current address: Department of Electrical and Computer Engineering, University of Pittsburgh, 3700 O'Hara St., Pittsburgh, Pennsylvania 15261, USA
[4] Department of Engineering, University of Exeter, Exeter, EX4 4QF, UK
[5] Institute of Physics, University of Muenster, 48149, Muenster, Germany
[6] Center for Soft Nanoscience, University of Muenster, 48149, Muenster, Germany
*Corresponding author: harish.bhaskaran@materials.ox.ac.uk

Received 11 February 2022; revised 13 May 2022; accepted 22 May 2022; published 14 July 2022

Over a century ago, Ivan P. Pavlov, in a classic experiment, demonstrated how dogs can learn to associate a ringing bell with food, thereby causing a ring to result in salivation. Today, it is rare to find the use of Pavlovian type associative learning for artificial intelligence applications even though other learning concepts, in particular, backpropagation on artificial neural networks (ANNs), have flourished. However, training using the backpropagation method on "conventional" ANNs, especially in the form of modern deep neural networks, is computationally and energy intensive. Here, we experimentally demonstrate a form of backpropagation-free learning using a single (or monadic) associative hardware element. We realize this on an integrated photonic platform using phase-change materials combined with on-chip cascaded directional couplers. We then develop a scaled-up circuit network using our monadic Pavlovian photonic hardware that delivers a distinct machine learning framework based on single-element associations and, importantly, using backpropagation-free architectures to address general learning tasks. Our approach reduces the computational burden imposed by learning in conventional neural network approaches, thereby increasing speed while also offering a higher bandwidth inherent to our photonic implementation.

https://doi.org/10.1364/OPTICA.455864

1. INTRODUCTION

The ability to decipher non-trivial patterns in data using computational techniques has led to the development of sophisticated machine intelligence approaches with a plethora of scientifically and technologically important applications [1–4]. Such approaches have predominantly been performed on general-purpose digital electronic processors (i.e., GPUs and CPUs), but this can introduce unwanted and deleterious computational latency and limitations to data throughput. Thus, special-purpose hardware accelerators designed intentionally for use in machine learning applications are an essential development [5–9]. Harnessing the wavelength-multiplexing capabilities of photonics to carry out parallel operations simultaneously in special-purpose accelerators can greatly increase the capacity of intelligent information processing [10–13].

Practical associative learning hardware accelerators require a hardware device structure that can associate inputs to a device, but current implementations using electronic [14–26], optoelectronic [27], and synthetic biological [28] approaches are limited at the device level. Specifically, the ability to monadically associate at least two inputs together is distinctly absent at a device level. In this paper, we experimentally demonstrate such a single associative learning element, one that exploits the ultrahigh bandwidth capabilities of photonics in a readily scalable architecture with the potential to deliver future artificial neural networks with very significantly faster and lower energy cost training as compared to conventional approaches.

2. ASSOCIATIVE MONADIC LEARNING ELEMENT CONCEPT

In biological systems, a fundamental associative learning process—classical conditioning—can be described using the neural circuit in Fig. 1(a) [29,30]. For a motor neuron to generate an action potential, it must receive a sensory signal. Pavlov in his experiment showed that salivation in a dog can be "stimulated" by associating the ringing of a bell with food [31], i.e., the two sensory stimuli were associated so as to generate an identical response. The process

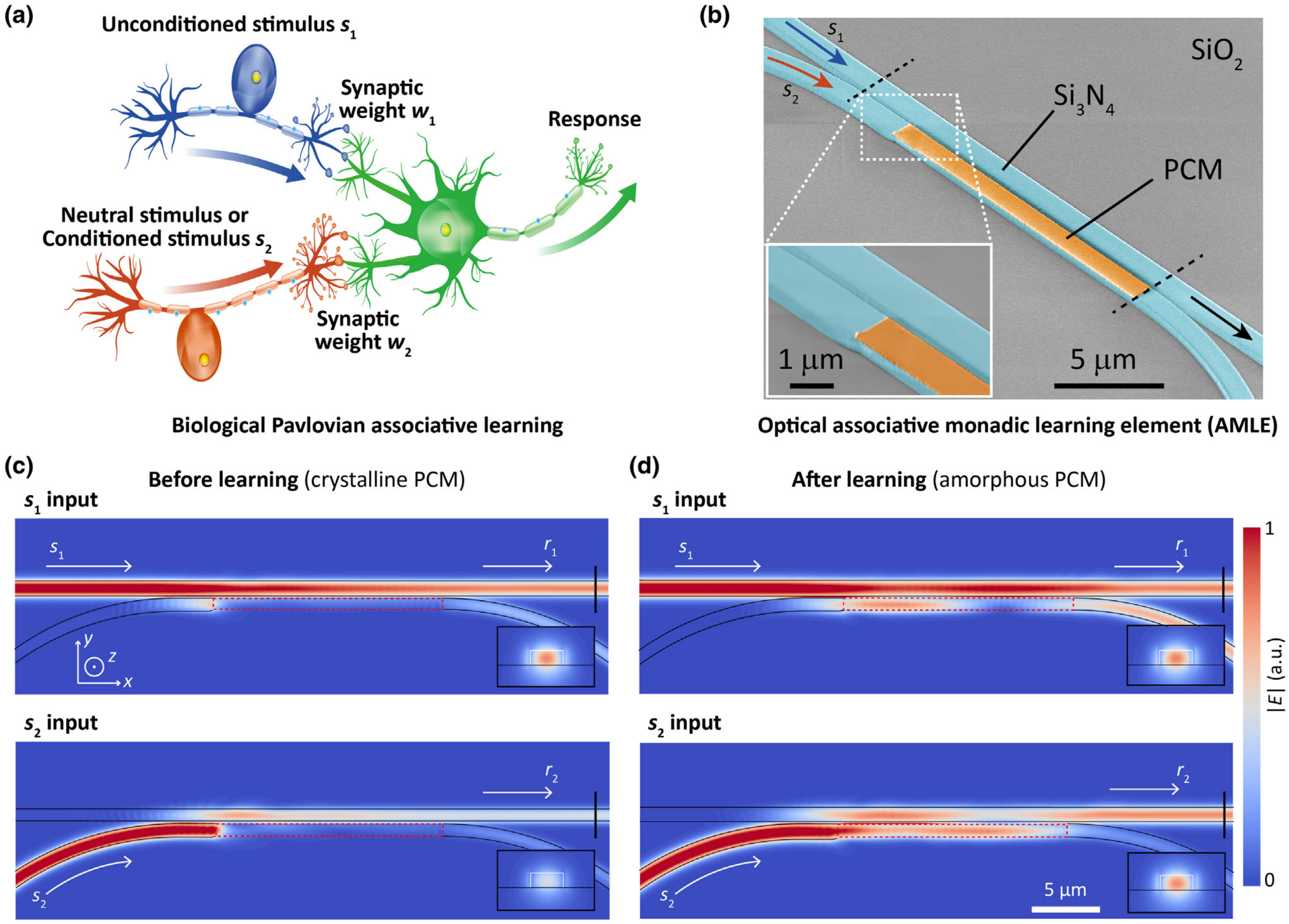
(a)
Unconditioned stimulus s_1
Synaptic weight w_1
Response
Neutral stimulus or Conditioned stimulus s_2
Synaptic weight w_2
Biological Pavlovian associative learning
(b)
s_1
s_2
SiO$_2$
Si$_3$N$_4$
PCM
1 µm
5 µm
Optical associative monadic learning element (AMLE)
(c)
Before learning (crystalline PCM)
s_1 input
s_1
r_1
y
z
x
s_2 input
r_2
s_2
(d)
After learning (amorphous PCM)
s_1 input
s_1
r_1
s_2 input
r_2
s_2
5 µm
1
|E| (a.u.)
0

Fig. 1. Optical associative monadic learning element (AMLE). (a) Simplified illustration of the neural circuitry for associative learning. After stimulus s_2 is associated with stimulus s_1, both elicit the same response. (b) SEM image (false colored) of a fabricated AMLE, consisting of Si$_3$N$_4$ directional couplers (cyan), SiO$_2$ undercladding (gray), and GST cell (marigold). Inset shows detail of coupling area with GST. (c), (d) Corresponding electric field profiles of AMLE (c) before and (d) after learning, with s_1 and s_2 inputs. Inset: output cross-section optical field at locations denoted by black vertical bars. The color bar is normalized.

of associating stimulus s_2 from a sensory neuron with natural stimulus s_1 is the process of association, and this is a learning mechanism. Once the association between the two signals is established, the response is triggered when either s_1 or s_2 is sent to the motor neuron through a synaptic weight [w_1 or w_2 in Fig. 1(a)]—the network has now learned this response. Thus, this simplified neural circuitry has two main roles: to converge and associate the two inputs, and to store memories of these associations *in situ*.

In our optical device, we embed the above functionality in an associative monadic learning element (AMLE) [Fig. 1(b)]. The device has two coupled waveguides and a thin film of phase-change material Ge$_2$Sb$_2$Te$_5$ (GST) on the lower waveguide that effectively modulates the coupling between the waveguides. The GST exists in two states (amorphous or crystalline). These two states (and the fractional volumes between the two states) govern the amount of coupling between the waveguides. As shown in Fig. 1(c), when the material is crystalline, there is no association between inputs s_1 and s_2. However, when the two inputs (learning pulses) arrive at the same time, the material has sufficient absorption to amorphize the GST, changing the coupling between the waveguides. As the material amorphizes, inputs s_1 and s_2 begin to "associate" as shown in Fig. 1(d). As the number of learning pulses increases, a larger volume of material switches from the crystalline to the amorphous state, until reaching a point when the two inputs result in an output that is nearly indistinguishable—we set this level and term this the learning threshold. Unlike previous non-volatile phase-change material photonic memories that relied primarily on optical amplitudes for their operation [11], we here employ the optical phase difference between inputs s_1 and s_2 to precisely control the phase state of the GST cell. This enables us to establish the precise extent of association between the inputs (details in Supplement 1). Our phase-change material GST is known to have an ultrafast structural phase transition time (sub-ns amorphization and few-ns crystallization time [32]), high cycling endurance (~10^{12} cycles [33,34]), and long retention time (>10 years at room temperature [34]). A thin capping layer of indium tin oxide (ITO) is deposited on the GST to prevent oxidation, and to help localize optically induced heat to enable low-power phase switching [11,35].

The association between inputs s_1 and s_2 during the learning process occurs only when the two inputs are paired at a specific optical phase delay $\Delta\varphi$. This results in a change in the synaptic weight Δw between stimulus and response signals. In our AMLE, the optical delay $\Delta\varphi$ is introduced by the optical phase difference between s_1 and s_2 inputs. We found that the slightest vibration and/or temperature change in the measurement environment can cause the optical phases to vary erratically. We mitigated this by using an on-chip layout (details in Supplement 1 Fig. S2.1) that greatly reduced effects of environmental disturbances on phase control.

The use of directional couplers ensures that the design is applicable over a broad optical wavelength range. Our simulations of the design show the natural response, as outlined in Fig. 1(c) (prior to associative learning taking place) and Fig. 1(d) (after the association). Before learning, only s_1 leads to a high transmission response, whereas s_2 does not. After learning, both s_1 and s_2 produce high transmission responses, which indicates that the two inputs are now associated, i.e., the system has "learned" to associate the two inputs such that both trigger the same response.

3. OBSERVATION OF PHOTONIC ASSOCIATIVE LEARNING

We now experimentally characterize the dynamic response of the AMLE. To achieve this, it is necessary to control the input signal combinations (s_1 only, s_2 only, and both s_1 and s_2) to the AMLE. We achieve this through the use of wavelength-selective, critically coupled ring resonators to each of the inputs leading to the AMLE (details in Supplement 1). This allows us to characterize the real-time dynamics of the associative learning process.

The starting point of the AMLE for our experiments is its crystalline state. We probe output transmission r_1 and r_2 of the AMLE in real-time using wavelengths λ_1 and λ_2. As shown in Fig. 2(a), transmission readouts r_1 and r_2 remained the same for single input pulses at 1.45 nJ (pulse widths $\tau = 100$ ns) as expected (for events 1 to 4). However, when inputs s_1 and s_2 were sent simultaneously at pump wavelength λ_0 with a fixed phase delay of $\pi/2$ at 0.66 nJ ($\tau = 100$ ns) each in event 5, transmission changes Δr_1 and Δr_2 for s_1 and s_2 probe readouts are $\sim -4\%$ and $\sim +4\%$ respectively. As the input pump pulse power was increased from 0.87 nJ ($\tau = 100$ ns) to 1.45 nJ ($\tau = 100$ ns) in events 6 to 8, the probe readouts changed by approximately -7% and $+7\%$, respectively, both of which are well above our output transmission threshold of $r_{\mathrm{th}} \sim 5\%$. Details of optical pulses are provided in Appendix A. Effectively, these experiments show that the two inputs can be "taught" to associate with each other such that either triggers the response, i.e., the associations are learned.

We then show in Fig. 2(a) (bottom chart) that these learned associations can also be reversed. A set of pulses at 0.43 nJ ($\tau = 100$ ns) in event 9, followed by 0.19 nJ pulses ($\tau = 100$ ns) in event 10 resulted in the "forgetting" process, where the readouts r reverted to the baselines ($r_1 \sim 0.14$ for s_1 input probe and $r_2 \sim 0$ for s_2 input probe). For our measurements, readout above a threshold $r_{th} \sim 0.05$ is designated as the learned state.

Figure 2(b) shows a single cycle of the real-time output readout of associative learning in events 5 to 8 and the forgetting process in events 9 to 11 of Fig. 2(a). To test the repeatability of our associative learning and forgetting processes, we subjected the AMLE through 80 learning cycles, examined over a period of 50 minutes. After

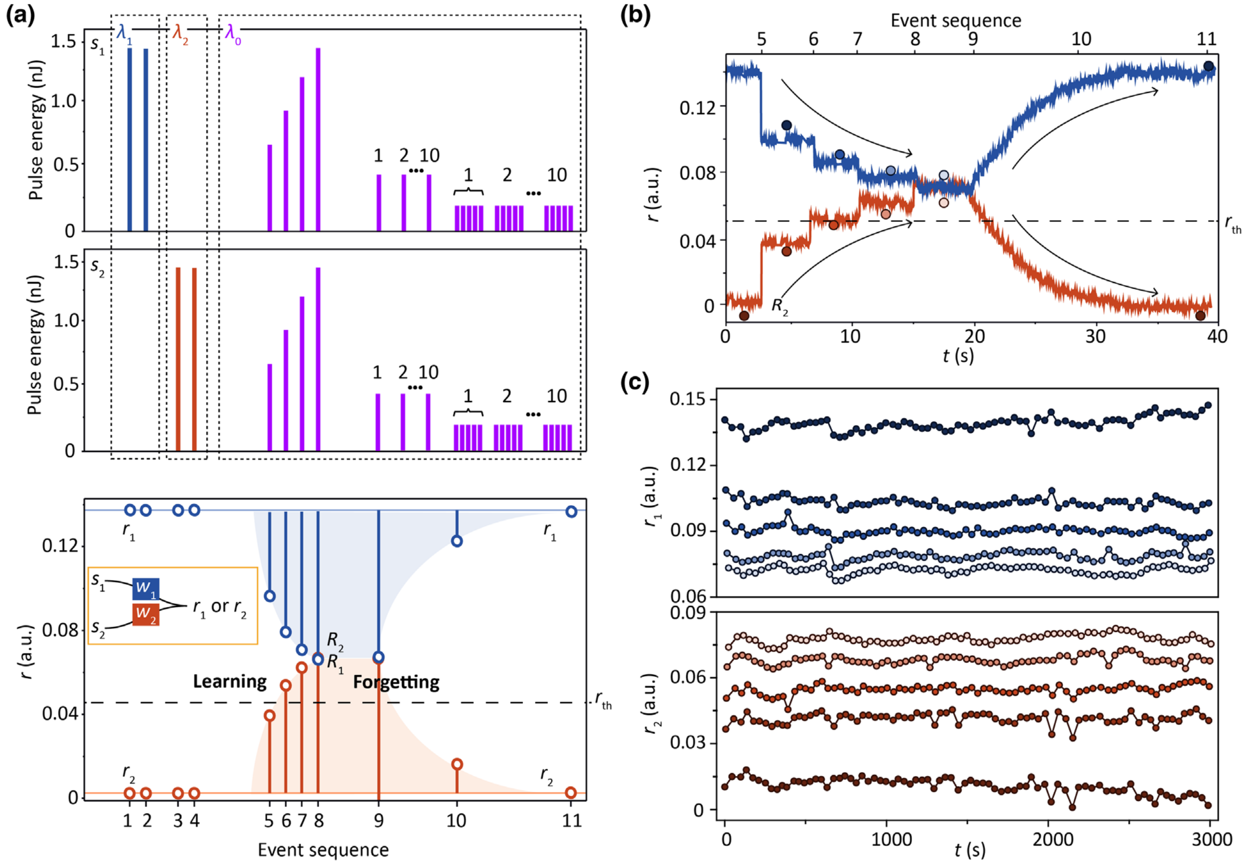

Fig. 2. Photonic Pavlovian learning process. (a) Input–output relation of AMLE. s_1 and s_2 inputs denote "food" and "bell" inputs; the corresponding output transmission r (bottom chart) represents transmission responses r_1 and r_2. Blue, red, and purple bars of the top and middle charts denote s_1, s_2, and both $s_1 - s_2$ input incidences, respectively (λ_1, λ_2, and λ_0 are the wavelengths used to selectively address each one). Bottom chart inset: simplified diagram of AMLE. (b) Corresponding real-time measurement of output probe transmission r of a single cycle learning and forgetting processes in (a). (c) Repeatability of the processes on AMLE over 80 cycles. The levels are denoted by filled circles of different colors that correspond to (b).

80 cycles [Fig. 2(c)], the individual learning weights were clearly identifiable with a standard deviation of ±0.69% in readout transmission.

4. ASSOCIATIVE NETWORK FOR SUPERVISED LEARNING

Up to this point, we have observed associative learning and characterized the workings of our AMLE via single device photonic measurements. The general concepts of Pavlovian associative learning [Fig. 3(a)] and supervised learning [a class of machine learning; Fig. 3(b)] are in essence comparable, both involving the pairing of input (*IN*) with the correct output (*Teacher*) to supervise the learning process. However, the mapping of the input to the desired output in conventional supervised learning architectures is relatively complicated as well as time and energy consuming—the *Teacher* signal propagates backward layer by layer to collectively adjust the network weights such that the actual output R better resembles the desired output (*Teacher*) after each learning iteration. The learning process for our AMLEs is by comparison far more straightforward (and faster and more energy efficient), and to elucidate this, we consider a scaled-up network of AMLE devices, which we illustrate schematically in Fig. 3(c).

To implement the system of Fig. 3(c), we use cascaded Mach–Zehnder modulators (MZMs) to provide a reliable means to split both the *Input* (*IN*) and *Teacher* signals equally with stable optical phases (obtained via the use of integrated NiCr thermo-optic heaters) and feed them to the inputs of the AMLEs. The MZMs also allow for the use of wavelength multiplexing to feed multiple signals to the inputs of multiple AMLEs, before these signals are paired with *Teacher* signals at the associative layer to cumulatively produce and output a transmission response. A more detailed

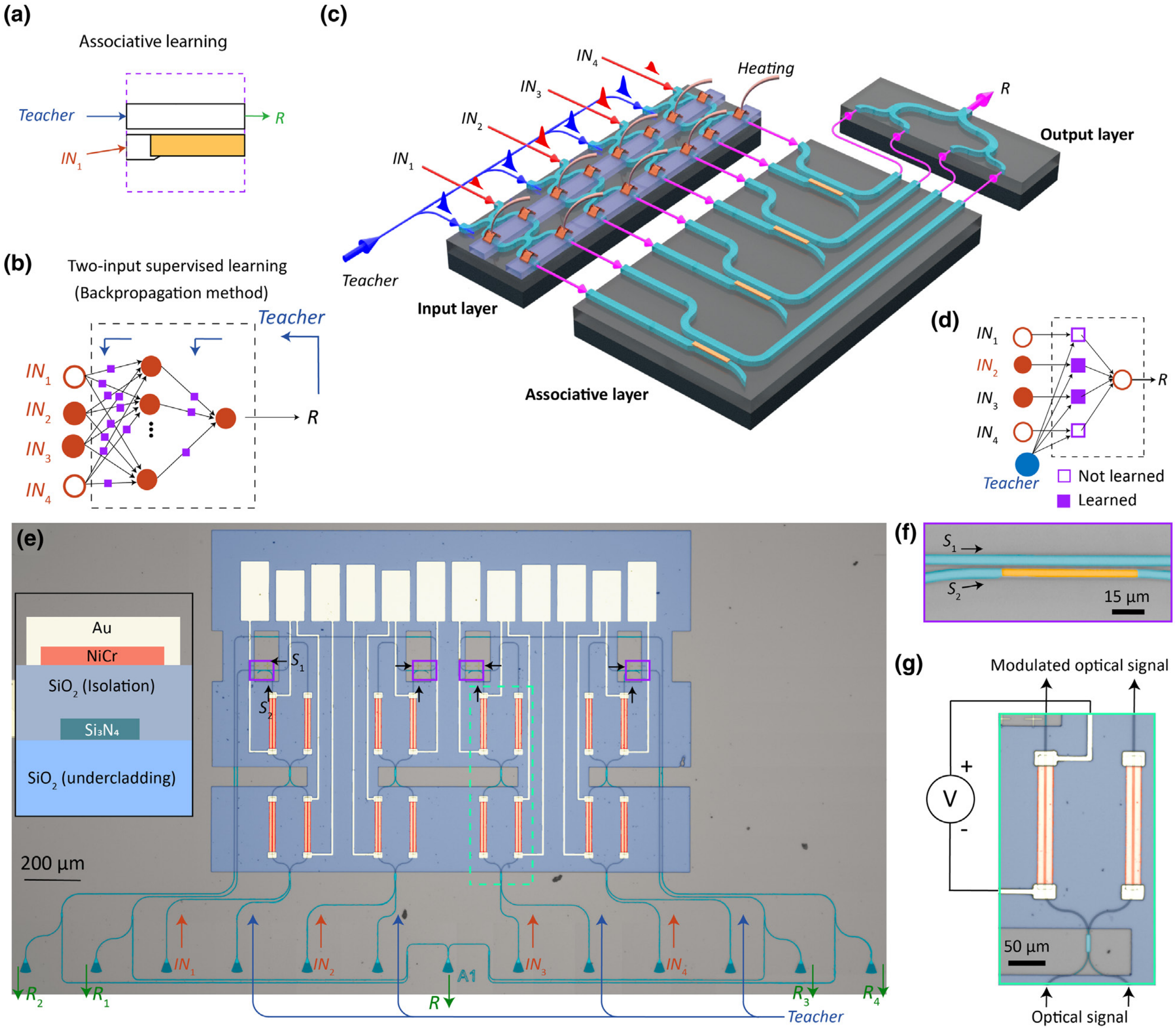

Fig. 3. Supervised Pavlovian associative learning. (a) Pavlovian learning involves pairing the inputs (*IN*) with the correct outputs (*Teacher*) to supervise the learning process. (b) Current conventional supervised learning networks use backpropagation. The network diagram depicts the network for supervised learning. (c) Optical on-chip hardware diagram of supervised learning with two inputs using AMLE. Input signals IN_1 to IN_4 are fed into the system during the learning process to be supervised by the *Teacher* input signal. IN_1 to IN_4 and *Teacher* inputs lead to the AMLE being controlled by thermo-optic (with NiCr heaters) MZMs. (d) Network representation of the hardware diagram in (c). (e) Optical micrograph of supervised learning network in (c), which consists of four AMLEs (boxed, purple). The arrows show the optical input–output connections coupled to the on-chip network via grating couplers. (f), (g) Optical micrographs of a single AMLE and heaters respectively correspond to those in (e).

illustration of the resulting integrated AMLE network chip is thus as shown in Fig. 3(e). The system consists of the associative layer (purple bordered boxes) between an input layer (*Input* data and *Teacher* denoted using red and blue arrows, respectively) and an output layer (response R; green arrow). Optical micrographs with an enlarged view of the AMLE and thermo-optic heaters are shown in Fig. 3(f) and Fig. 3(g), respectively (see Supplement 1 and Figs. S3.1 and S3.2).

We use the AMLE network chip of Fig. 3(e) to carry out a rapid pattern recognition task to verify that once the associations are formed by the AMLEs, simultaneous parallel recognition can be achieved. The GST cells are initially set to their crystalline states (details in Appendix A). We then program the bit pattern "0110" to the AMLEs by respectively sending the pump *Teacher* signal $T = 1.47$ nJ ($\tau = 100$ ns) and pump *Input* signals $IN_1 = 0$, $IN_2 = 1.47$ nJ ($\tau = 100$ ns), $IN_3 = 1.47$ nJ ($\tau = 100$ ns) and $IN_4 = 0$ simultaneously, essentially associating these patterns within the AMLE network [measurement setup is detailed in Supplement 1 and shown in Figs. S4(a) and (b)]. Pattern recognition is then carried out by sending a series of randomized optical binary signals modulated at 1 GHz speed to the AMLEs (with 0 mW for "0" and 0.7 mW for "1"), using different wavelengths (in the range of 1548.51 nm to 1550.92 nm, C33 to C36 in the telecom band) for each of the individual AMLE inputs [see Fig. 4(a)] (further details in Supplement 1). The output from each AMLE device in the network passes to four photodiodes to generate the system outputs (or responses), R. The output transmission response of a single AMLE r_{ij} measured at wavelength λ_j from the photodetectors (PDs) is then summed to obtain the cumulative transmission response $R_j = \Sigma_i\ R_{ij}$ for the specific wavelength, where R_{ij} is the measured output transmission of the ith AMLE and jth wavelength after normalizing (between zero and one range) using $R_{ij} = (r_{ij} - r_{\mathrm{base}})/r_{\mathrm{base}}$, i.e., the average value of two consecutive output transmission values using a trigger pulse (with r_{base} being the baseline of the optical transmission. The average value is taken to minimize the effects of electrical or optical noise. After recording the output pulses from the experiments at four optical wavelengths, we systematically retrieve the cycle(s) at which the cumulative response $R_c = \Sigma_{ij}\ R_{ij}$ is above a pre-determined learning threshold. Figure 4(c) shows the inputs sent at the four different wavelengths C33 to C36, while the corresponding output response of the AMLEs is shown in Fig. 4(d). Figure 4(e) shows the cumulative R_c for all four wavelengths, which exceeds the learning threshold only at cycle 30; this enables us to pinpoint the output bit combinations at which input signals from all four wavelengths match the bit pattern "0110" to be at cycle 30. Figure 4(f) shows the performance of the AMLEs for detection predictions. We measured the prediction error $Error = R - R_{\mathrm{calculated}}$ from the difference between output transmission $R = \Sigma_i\ R_{ij}$ and the calculated transmission $R_{\mathrm{calculated}}$. The error is within the ±0.5 threshold bands.

It is important to point out that the detection speed in the experiments shown in Fig. 4 is limited only by the number of pulses to average (in our case, two) and the signal modulation and detection speed of the modulators and PDs. In principle, this detection speed can be significantly improved by increasing the signal to noise ratio (by using on-chip PDs instead of external PDs connected via grating couplers in our case) and increasing the input signal modulation speed while maintaining detected signal integrity. Although we demonstrated detection specifically for the pattern "0110," this hardware system can be used to detect all other possible (and more) patterns (as shown in Supplement 1 Fig. S6) by additionally representing on other AMLEs another separate set of bits toggled to the bit pattern to be detected, and then additionally sending toggled sets of input bits to these additional AMLEs.

Now, we demonstrate how AMLE-based hardware can achieve generalization on an image recognition task using associative learning, based on the network architecture shown in Fig. 5(a). The network is similar to those in Fig. 3(d), which consists of three main network layers (i.e., input layer, associative layer, and output layer) [36]. During the training process, images to be trained are first pixelated to 15 × 13 input pixel data (195 pixels altogether). These data $IN_1 - IN_{195}$ are then sent to the associative layer, while simultaneously, the *Teacher* signals are likewise preprocessed and sent to the associative layer. Notably, images fed into $IN_1 - IN_{195}$ and *Teacher* signals are exchangeable in the training process. The change in the learning states of AMLEs is determined from their transmission readouts, when preprocessed *Input* data of maximum amplitude (blank white image) and the same dimensions as the training pixel data are fed through the layers. The transmission of these individual AMLEs is summed up at the output layer and rearranged to form a 15 × 13 pixel model representation. The complete computation is obtained by cumulatively adding the model representation from each training pair. A pixel-by-pixel comparison between the model (which generalizes the training images) and measurement is then made to determine whether the testing images are of the image to be detected.

The photonic implementation of the associative learning network architecture for image recognition is shown in Fig. 5(b). Here, the GST elements on the AMLEs are first initialized to the crystalline state before the signals are sent to the on-chip photonic structure. During training, depending on the optical input signal and input pump power sent to the AMLEs, the state of the GST cells on AMLEs either remains in crystalline or structurally switches to an amorphous state. This results in a change in the optical probe output response of the AMLEs. In our experiment, because we have only four AMLEs, we raster the 195 pairs of input pixels across them four.

We examine the cat image classification capabilities of our associative learning network using the "Dogs versus Cats" dataset from fast.ai [37] (single dataset collected from CIFAR10, CIFAR 100 [38], Caltech 101 [39], Oxford-IIIT Pet [40], and Imagewoof [41]). The cat images that we used for the training process are shown in Fig. 5(c). After the training process using these images, we obtain the model representation of a cat shown in Fig. 5(d) by feeding in a 15 × 13 pixel blank white test input of maximum probe input power magnitude (1.3 mW; further details of the experiment in Supplement 1).

Thus far, we have demonstrated searching for patterns (in the form of pixel amplitudes) from the 15 × 13 pixel image sent to the network of monadic AMLEs in a single step. After each training iteration with a model image, the network's ability to distinguish the appearance of a cat improves. The feature subtleties can then be captured from the model representation, giving us a valuable means to distinguish a cat from other objects. To test the model representation, we use the testing images shown in Fig. 5(e) to compare them (im) with the representation of the cat rep and measure the error function with respect to pixel j, given by $\tanh\,[(rep_j - im_j)^2\ e]$ for every pixel j. We set the threshold $min(Error) + (max(Error) - min(Error))/2$, which is 22.625

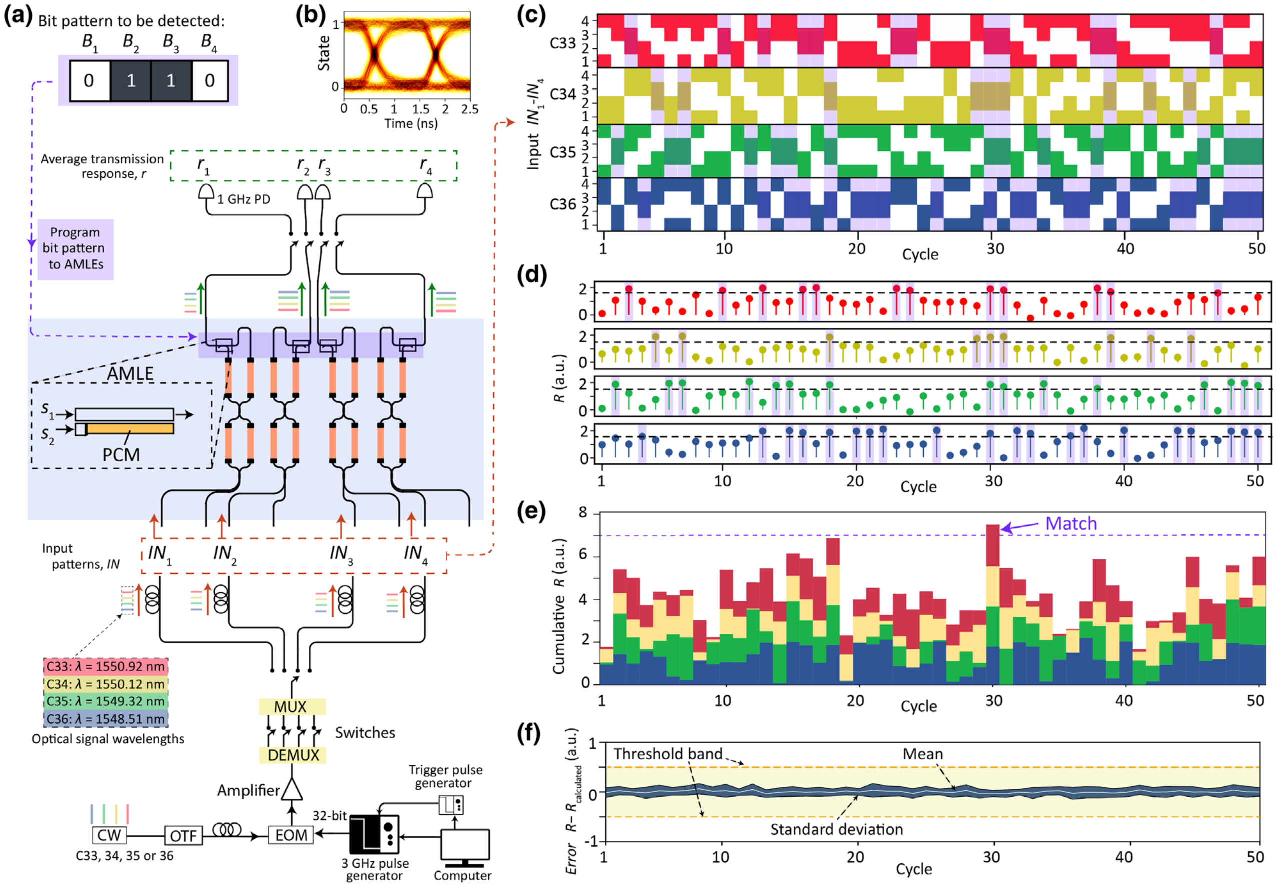

Fig. 4. Pattern recognition with AMLE. (a) Measurement setup for pattern recognition. Bit patterns to be detected "0110" are associated into the AMLE via a learning step. 1 GHz optical detection pulses defined by computer-generated pseudo-random bits at wavelengths from C33 to C36 (1548.51 nm to 1550.92 nm) are sent through AMLE devices and detected by photodiodes as the signals are temporally multiplexed. Measured average output transmission R cumulatively summed to determine if pattern is recognized. (b) Eye-diagram obtained from non-return-to-zero (NRZ) pseudo-random binary sequence (PRBS) pulses modulated at 1 GHz shows clear distinction between states "0" and "1." (c) Inputs IN_1 to IN_4 sent at four different wavelengths i.e., C33 to C36 (input = "1" represented in red, yellow, green, and blue respectively). Purple shades represent "0110" pattern sent to the devices. (d) R for the different wavelengths (C33 to C36). Purple shades represent "0110" pattern detected. (e) Cumulative R obtained for the respective wavelengths and detection cycles. Values above cumulative transmission threshold $R = 7$ (dashed black line) denote all-matching response (i.e., in cycle 30). (f) Calculated error $R - R_{\mathrm{calculated}}$ (difference between measured average transmission response and expected response) shows mean error is negligibly low with standard deviation well within the ±0.5 threshold band.

for our case, to determine the *Testing Images* that resemble the model representation in Fig. 5(d). Our results, summarized in Fig. 5(f), reveal that *Testing Images* 1, 3, and 5 resemble the cat representation. With the error [*Error* = squared Euclidean distance $\Sigma\ (rep_j - im_j)^2$] of *Images* 2, 4, and 6 above the threshold, the network predicts that these images are not of a cat. The associative learning network thus accurately classifies images of cats from the model representation obtained from the training iterations. In contrast to conventional numeric-based artificial intelligence (AI) approaches, our symbolic associative learning system interprets from the "cat representation" results to logically conjure up the proposition that an image must have the shown clear, distinctive, and comprehensible physical attributes to constitute a cat image. Our image recognition example adopts symbolic AI, while conventional image recognition uses connectionist AI. Our approach is typically simpler and faster, but comes at the expense of not having the ability to acquire deep features—an ability that may not be needed for many less complex machine learning tasks.

5. EVALUATION METRICS FOR ASSOCIATIVE LEARNING DEVICES

Our application-specific system offers convergence in one training step due to its straightforward approach of detecting similarities without having to randomize network weights and perform backpropagation as in conventional neural networks.

We identify relevant device-level evaluation metrics by contextualizing the AMLE with a typical machine learning data load, which requires data to be transferred back and forth from the data source to be run using cloud computing and/or supercomputers. For a more energy-efficient locally run neural network, it is important to shrink the network for greater portability and reduce the energy consumption of the learning process. Table 1 summarizes the minimum active device volume (in our case, volume of GST and the waveguide below) and learning energy of other associative learning devices [14–27]. The electronic and optoelectronic associative learning devices range from ~0.1–10^{10} μm^3 in active volume and consume ~2.63–10^5 nJ of energy per learning event [14–27]. In comparison, the all-optical AMLE in our work fares favorably relative to these devices in terms of

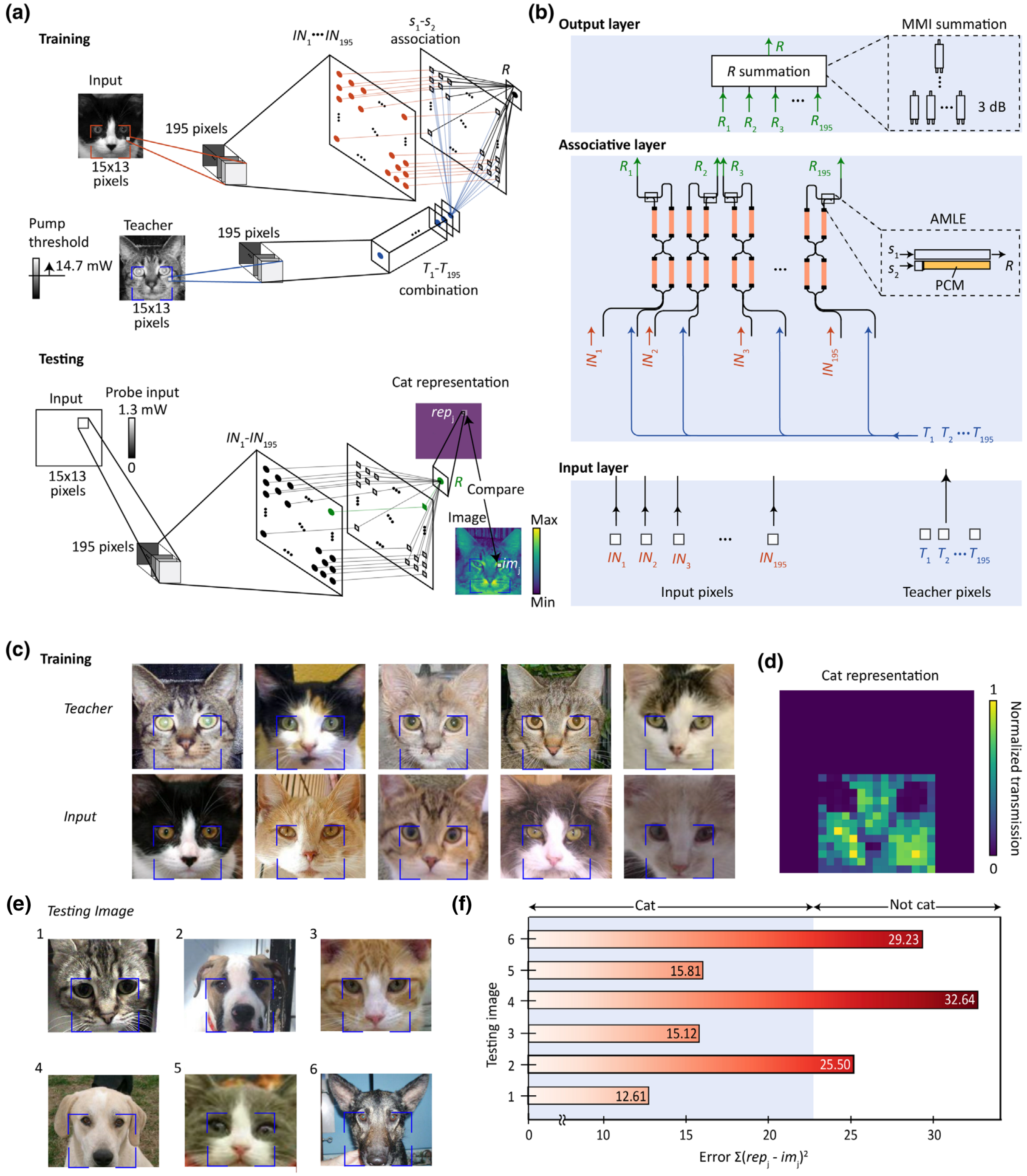

Fig. 5. Scaling architecture for image recognition using associative learning network. (a) General associative neural network composed of an input layer ($IN_1 - IN_{195}$, $T_1 - T_{195}$), associative layer (stimuli $s_1 - s_2$ association), and output layer (transmission R). The input signal is the pattern (pixels from image) to be classified, and the external teacher provides the desired set of output values chosen to represent the class of patterns being learned. During training, the input layer ($IN_1 - IN_{195}$, $T_1 - T_{195}$) is fed into the associative layer. Associative layer consists of AMLEs with states that are modified when both the input and external teacher signals are paired together. A model representation of the trained images is then obtained by sending a preprocessed input signal of a blank white image to propagate through the layers. (b) Optical on-chip implementation of the input layer, associative layer, and output layer. The associative layer, which consists of thermo-optic NiCr heaters, distributes the combination of the input and external teacher signals (from the input layer) as s_1 and s_2 inputs to the respective AMLEs. The output layer consists of a summation unit to sum up the output response from the AMLEs to form the net output. Conventional 2 × 1 combiner has 3 dB loss each. (c) Training the associative learning network to identify cat images, with bounding box corners to indicate region of interest. (d) After five training iterations, the network learns the model representation of the cat from the output response R. (e) Images used to test whether the network can correctly classify pictures as cat and non-cat. (f) The network successfully recognizes cat and non-cat images based on the squared Euclidean distance $\Sigma\,(rep_j - im_j)^2$ measured over the pixelated testing images for each of the 15 × 13 pixels.

Table 1. Comparison of Active Volume and Learning Energy in Associative Learning Devices

Type	Active Volume (μm^3)	Min. Learning Energy (nJ)	Refs.
Electronic			
• Memresistive			
i. Chalcogenide	0.12–10.5	4.7×10^4	[14]
	8	2.63	[15]
ii. Manganite	~0.1	1.35×10^3	[16]
	1.25×10^{10}	1.02×10^5	[17]
iv. Nickelate	4.7×10^3	7.20×10^5	[18]
	4.8×10^4	2.04×10^5	[19]
v. Metal oxide	~10−900	$\sim 10^3 - 10^5$	[20,21]
vi. Organic	~ 0.1−0.5	$\sim 10^3 - 10^4$	[22,23]
• Electrochemical	6×10^3	6×10^4	[24]
	9.6×10^5	125	[25]
• Memcapacitive	26.9	~ 30	[26]
Optoelectronic	1.62×10^3	2.1×10^3	[27]
Optical AMLE	0.12	1.8	**This work**

dimensions and very favorably in terms of energy usage, with a low active volume at 0.12 μm^3 and minimum learning energy at only 1.8 nJ. The overall size of the single-element device is of dimensions 2 μm × 17 μm × 0.33 μm. Despite the relatively low active volume (of GST) in our case, compared to electronic hardware, our devices are still larger; however, the speed and multiplexing of photonic devices can lead to higher computational density. Experimentally, we have demonstrated 118 TOPS/mm^2, which was limited by the available setup in the laboratory (see Supplement 1). Scaling the device to more wavelength channels and modulating at a higher speed potentially leads to significant gains in the compute density. Considering, for example, modulation at 50 GHz [42] and using 16 wavelengths, the device would deliver approximately 2.5×10^4 TOPS/mm^2. The overall net energy usage of the learning system is directly related to the number of iterations required to train a learning system (details in Supplement 1). The energy–speed trade-off—particularly evident in CMOS devices [43]—is another important evaluation metric that requires further investigation and research. Our device learns in ~100 ns, compared to ~ms in a previous associative learning device [15].

6. CONCLUSION

Our results show the first demonstration of an AMLE implemented on a photonic platform. We provide a supervised learning framework that facilitates the transition from a monadic Pavlovian single *Input–Teacher* association on an AMLE to any arbitrary *n Input–Teacher* associations, thus enabling backpropagation-free, single-layer weight artificial neural network architectures.

We have elucidated the inner workings of the network building block, which can spatiotemporally correlate two initially distinct inputs (s_1 and s_2) to a same output when both inputs are simultaneously applied at a predetermined $\Delta\varphi$ optical delay. Given that light signals inherently do not interfere at different wavelengths in linear media (including the AMLE), such input–input association can handle associations of multiple data streams consisting of different wavelengths over a single element, as we have experimentally demonstrated. Our photonic platform allows for wavelength multiplexing, which is inherently suited to the highly parallel nature of machine learning. We anticipate further improvements in other relevant metrics (e.g., overall device volume and learning energy) on different material platforms and with other optimization methods.

More generally and interestingly, our work can extend the one-way learning (s_2 becomes associated to s_1) to a customizable form of learning, for example, mutual/two-way learning (s_2 becomes associated to s_1 while s_1 is associated to s_2; details in Supplement 1. This customizable feature when combined with demonstrations of deterministic weights using identical, fixed-energy, and fixed-duration pulses [35] will provide unprecedented design flexibility for a wide range of machine learning applications. As to whether the nonlinear scaling with the number of inputs for nonlinear classification problems is an inherent attribute in associative learning is still an open question (discussed in Supplement 1). However, as shown in Figs. 4 and 5, practical applications such as pattern recognition and image recognition can be readily demonstrated with an associative learning network that scales linearly with the number of inputs. The compact single-element implementation in our work will allow the use of the AMLE as a building block in machine learning/statistical inference in general, thus potentially opening up new avenues of research in machine learning algorithms and architectures.

APPENDIX A: METHODS

1. FDTD SIMULATION

Three-dimensional finite-difference time-domain (FDTD) simulations were performed using the FDTD Solutions software from Lumerical Inc. A fundamental quasi-transverse electric (TE; magnetic field component H_z dominant) optical mode source input of 1 V/m at 1.58 μm wavelength is let incident onto the AMLE waveguide input. The simulation plots in Figs. 1(c) and 1(d) show optical field $|E|$ profiles taken at the central cross section in the $x-y$ plane and $y-z$ plane [axis depicted in Fig. 1(c)] of the AMLE structure. Our numerical simulations summarized in Figs. S1(g) and (h) indicate that the output transmission at 1.55 μm wavelength is within the same range as those at 1.58 μm wavelength.

2. DEVICE FABRICATION AND CHARACTERIZATION

The AMLE is fabricated on a Si_3N_4/SiO_2 platform. Electron beam lithography (JEOL 5500FS, JEOL Ltd.) is used at 50 kV to define the Si_3N_4 structure on the Ma-N 2403 negative-tone resist-coated substrate. After the development process, reactive ion etching (PlasmaPro 80, Oxford Instruments) is performed in $CHF_3/O_2/Ar$ to etch down 330 nm of Si_3N_4. Electron beam lithography is then implemented on a poly(methyl methacrylate) (PMMA) positive resist-coated substrate to open a window for the GST cell. This is followed by sputter-deposition (Nordiko RF Sputter Tool) of 10 nm GST/10 nm ITO on the substrate. For the heater-based layout, windows to deposit the SiO_2 isolation layer are opened on photoresist S1813, exposed using mask aligner (SUSS MicroTec SE). An NiCr layer (for heaters) is deposited on the SiO_2 isolation layer using the sputtering tool. A gold layer (for on-chip electrical pads) is deposited on the NiCr layer using thermal evaporation (Edwards 306 Vacuum Coater/Deposition systems). The AMLE characterization process is performed using high-resolution emission gun SEM (Hitachi S-4300 SEM system, Ibaraki, Japan) with low accelerating voltage (1–3 kV) at a working distance of ~13 mm, and using an optical microscope (Eclipse LV100ND, Nikon).

3. OPTICAL MEASUREMENT

The experiment setup to measure AMLE on a ring-based layout (Layout 1 for Fig. 2) builds upon a previously described probe–pump configuration [11], and is described in Supplement 1. To probe AMLE transmission, two low-power continuous-wave (CW) probe diode lasers (N7711A, Keysight Tech.) are used as probe lasers. The signals coupled from the layout are filtered by optical tunable bandpass filters (OTFs; OTF-320, Santec Corp.) and detected by PDs (2011-FC, Newport Spectra-Physics Ltd.). To induce learning on the AMLE, optical pump pulses are sent to the AMLE. A CW diode laser (TSL-550, Santec Corp.) is used for the pump signal. The pulse shape of the optical signal is defined by an electro-optic modulator (EOM; 2623NA, Lucent) based on the electrical pulse shape generated by the arbitrary function generator (AFG; AFG 3102C, Tektronix). The optical pump pulses are then amplified by a low-noise erbium-doped fiber amplifier (EDFA; AEDFA-CL-23, Amonics) and sent to the AMLE.

In the stabilization step carried out prior to the experiment, a set of amorphizing pulses is sent to the AMLE, followed by a set of crystallizing pulses. These sets of pulses are exactly the same as the pulses applied during the "associative learning" and "forgetting" process shown in Fig. 2(a). Here, the set of amorphizing pulses is consecutive 100 ns wide pulses at 0.66 nJ, 0.87 nJ, 1.26 nJ, and 1.45 nJ, while the set of crystallizing pulses is a 100 ns wide pulse at 0.43 nJ for 10 times, followed by five 0.19 nJ 100 ns wide pulses at 1 MHz repetition rate for 10 times. These forgetting pulses introduce sustained heating at temperatures above crystallization temperature and below melting temperature to crystallize the GST.

Our pattern and image recognition experiments in Figs. 4 and 5 are carried out by measuring the associative hardware network shown in Fig. 3(e). These experiments are based on the experiment setup to measure a single AMLE device on a heater-based layout (Layout 2), which is thoroughly described in Section S3 and Fig. S3.1 in Supplement 1.

The pattern recognition experiment in Fig. 4 consists of two steps: pattern programming and pattern detection. In the first step, we program the AMLE with bit patterns by sending pump pulses to the AMLEs in the associative hardware network. In the second step, we send probe pulses at high speed (1 GHz modulation rate) and measure the AMLE output transmission responses to determine if the patterns sent match the patterns programmed into the AMLEs. The experiment setup to program a set of patterns to the AMLEs is shown in Figs. S4(a) and (b) and described in Section S4 in Supplement 1. The experiment setup for the detection step is shown in Fig. 4(a). In the setup, the AFG (AFG 3102C, Tektronix) specifically picks the continuous 32 bit 1 GHz pulses from the 3 GHz pulse generator (HP 8133A, Hewlett Packard). The EDFA (AEDFA-CL-23, Amonics Ltd.) amplifies the pulse from EOM (2623 NA, Lucent Tech.). The AMLE output pulses coupled from the layout are detected by the 1 GHz fiber optic PD (1611FC-AC, Newport). Technical details in this step are provided in Supplement 1.

The image recognition experiment in Fig. 5 consists of two steps: training and testing. As in other face recognition methods, the datasets are preprocessed to filter out irrelevant datasets to ensure that they have reliably sufficient facial landmarks, before the associative learning method engages in the image recognition process. Both steps are implemented using the integrated pump–probe measurement setup shown in Fig. S5(a) in Supplement 1. In the setup, input signals IN_1-IN_{195} are represented as optical signals wavelength-multiplexed from the spectrally filtered (SuperK Split, NKT Photonics) laser (WhiteLase Micro, NKT Photonics). Another set of separate supervisory *Teacher* signals T_1-T_{195} from the CW laser (TSL-550, Santec Corp.) is distributed to each input. These input pairs are rastered across the four hardware AMLEs shown in Fig. 3(e). The combined signals are selectively routed to either a probe line or pump line amplified by EDFA (FA-15, Pritel and FA-33-IO, Pritel), filtered by OTF (OTF-320, Santec Corp.), and channeled to the layout. At the output end of the layout, the readouts are filtered by OTF (OTF-930, Santec Corp.) and detected by PD (2011-FC, Newport Spectra-Physics Ltd.). During pumping, the optical attenuator (V1550PA, Thorlabs Inc.) connected prior to OTF is activated to filter out pump signals from PD. To modulate optical inputs to the AMLEs in the associative hardware network, voltage biasing to the on-chip NiCr waveguide heaters is applied. Details of the setup are provided in Supplement 1. The measured optical transmission to obtain the net cat representation is provided in Fig. S5(b) in Supplement 1. The testing process is carried out by comparing pixel-by-pixel the testing image with the net cat representation.

Funding. European Commission (101017237, 780848); Engineering and Physical Sciences Research Council (EP/J018694/1, EP/M015173/1 and EP/M015130/1); National Key Research and Development Program of China (2020YFA0308800); National Natural Science Foundation of China (62074042); Science and Technology Commission of Shanghai Municipality (20501130100); The Young Scientist Project of MOE Innovation Platform.

Acknowledgment. The authors acknowledge discussions with and the experimental work of Andrew Katumba, Nikolaos Farmakidis, Eugene J.H. Soh, Wen Zhou, and Hazim Hanif. All authors acknowledge funding from the European Union's Horizon 2020 Research and Innovation Programme. This research was also supported via the Engineering and Physical Sciences Research Council. Z.C. acknowledges support from National Key Research and Development Program of China, NNSF China, Science and Technology Commission of Shanghai Municipality, and The Young Scientist Project of MOE Innovation Platform.

Author contributions. H.B. led the project with inputs from C.D.W and W.H.P.; H.B., J.Y.S.T., and Z.C. conceived of and planned the project, and drafted the paper. J.Y.S.T. and Z.C. performed the analytical calculations, simulations, fabrications, and characterizations, with all authors contributing to the experiments or analysis and discussing the results and paper.

Disclosures. The authors acknowledge that patents related to this technology have been filed through the University of Oxford.

Data availability. All data needed to evaluate the conclusions in the paper are present in the paper and/or Supplement 1. Additional data related to this paper may be requested from the authors.

Supplemental document. See Supplement 1 for supporting content.

[†]These authors contributed equally to this paper.

REFERENCES

1. M. Usama, J. Qadir, A. Raza, H. Arif, K.-L. A. Yau, Y. Elkhatib, A. Hussain, and A. Al-Fuqaha, "Unsupervised machine learning for networking: techniques, applications and research challenges," IEEE Access **7**, 65579–65615 (2019).
2. B. A. M. Velasco, Á. G. Herranz, M. Á. López, F. Pulvirenti, and N. Maravitsas, "A telecom analytics framework for dynamic quality of service management," in *1st International Workshop on Big Data Applications and Principles* (Springer, 2014), pp. 103–134.
3. R. J. Bolton and D. J. Hand, "Statistical fraud detection: a review," Stat. Sci. **17**, 235–255 (2002).
4. A. S. Koyuncugil and N. Ozgulbas, *Surveillance Technologies and Early Warning Systems: Data Mining Applications for Risk Detection* (Information Science Reference, 2011).
5. T. Paine, H. Jin, J. Yang, Z. Lin, and T. Huang, "GPU asynchronous stochastic gradient descent to speed up neural network training," presented at 2nd International Conference on Learning Representations (ICLR 2014), Banff, AB, Canada (April 14-16 2014).
6. N. P. Jouppi, C. Young, N. Patil, *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *44th Annual International Symposium on Computer Architecture (ISCA)* (2017).
7. A. S. Cassidy, J. Georgioub, and A. G. Andreou, "Design of silicon brains in the nano-CMOS era: spiking neurons, learning synapses and neural architecture optimization," Neural Netw. **45**, 4–26 (2013).
8. M. Gschwind, H. P. Hofstee, B. Flachs, M. Hopkins, Y. Watanabe, and T. Yamazaki, "Synergistic processing in cell's multicore architecture," IEEE Micro **26**, 10–24 (2006).
9. J. Misra and I. Saha, "Artificial neural networks in hardware: a survey of two decades of progress," Neurocomputing **74**, 239–255 (2010).
10. J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. L. Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**, 52–58 (2021).
11. C. Ríos, M. Stegmaier, P. Hosseini, D. Wang, T. Scherer, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "Integrated all-photonic non-volatile multi-level memory," Nat. Photonics **9**, 725–733 (2015).
12. B. J. Shastri, A. N. Tait, H. Bhaskaran, C. D. Wright, T. F. de Lima, and P. R. Prucnal, "Photonics for artificial intelligence and neuromorphic computing," Nat. Photonics **15**, 102–114 (2021).
13. Y. Shen, N. C. Harris, M. Hochberg, S. Skirlo, X. Sun, S. Zhao, M. Prabhu, H. Larochelle, T. Baehr-Jones, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**, 441–446 (2017).
14. M. Ziegler, R. Soni, T. Patelczyk, M. Ignatov, T. Bartsch, P. Meuffels, and H. Kohlstedt, "An electronic version of Pavlov's dog," Adv. Funct. Mater **22**, 2744–2749 (2012).
15. Y. Li, L. Xu, Y.-P. Zhong, Y.-X. Zhou, S.-J. Zhong, Y.-Z. Hu, L. O. Chua, and X.-S. Miao, "Associative learning with temporal contiguity in a memristive circuit for large-scale neuromorphic networks," Adv. Electron. Mater. **1**, 1500125 (2015).
16. K. Moon, S. Park, J. Jang, D. Lee, J. Woo, E. Cha, S. Lee, J. Park, J. Song, Y. Koo, and H. Hwang, "Hardware implementation of associative memory characteristics with analogue-type resistive-switching device," Nanotechnology **25**, 495204 (2014).
17. Z.-H. Tan, X.-B. Yin, R. Yang, S.-B. Mi, C.-L. Jia, and X. Guo, "Pavlovian conditioning demonstrated with neuromorphic memristive devices," Sci. Rep. **7**, 713 (2017).
18. S. G. Hu, Y. Liu, Z. Liu, T. P. Chen, Q. Yu, L. J. Deng, Y. Yin, and S. Hosaka, "Synaptic long-term potentiation realized in Pavlov's dog model based on a NiOx-based memristor," J. Appl. Phys. **116**, 214502 (2014).
19. S. D. Ha, J. Shi, Y. Meroz, L. Mahadevan, and S. Ramanathan, "Neuromimetic circuits with synaptic devices based on strongly correlated electron systems," Phys. Rev. Appl. **2**, 064003 (2014).
20. C. Wu, T. W. Kim, T. Guo, F. Li, D. U. Lee, and J. J. Yang, "Mimicking classical conditioning based on a single flexible memristor," Adv. Mater. **29**, 1602890 (2017).
21. M. Kumar, S. Abbas, J.-H. Lee, and J. Kim, "Controllable digital resistive switching for artificial synapses and Pavlovian learning algorithm," Nanoscale **11**, 15596–15604 (2019).
22. O. Bichler, W. Zhao, F. Alibart, S. Pleutin, S. Lenfant, D. Vuillaume, and C. Gamrat, "Pavlov's dog associative learning demonstrated on synaptic-like organic transistors," Neural Comput. **25**, 549–566 (2013).
23. C. Eckel, J. Lenz, A. Melianas, A. Salleo, and R. T. Weitz, "Nanoscopic electrolyte-gated vertical organic transistors with low operation voltage and five orders of magnitude switching range for neuromorphic systems," Nano Lett. **22**, 973–978 (2022).
24. C. Wan, J. Zhou, Y. Shi, and Q. Wan, "Classical conditioning mimicked in junctionless IZO electric-double-layer thin-film transistors," IEEE Electron. Dev. Lett. **35**, 414–416 (2014).
25. F. Yu, L. Q. Zhu, H. Xiao, W. T. Gao, and Y. B. Guo, "Restickable oxide neuromorphic transistors with spike-timing-dependent plasticity and Pavlovian associative learning activities," Adv. Funct. Mater. **28**, 1804025 (2018).
26. Z. Wang, M. Rao, J.-W. Han, *et al.*, "Capacitive neural network with neuro-transistors," Nat. Commun. **9**, 3208 (2018).
27. R. A. John, F. Liu, N. A. Chien, M. R. Kulkarni, C. Zhu, Q. Fu, A. Basu, Z. Liu, and N. Mathews, "Synergistic gating of electro-iono-photoactive 2D chalcogenide neuristors: coexistence of Hebbian and homeostatic synaptic metaplasticity," Adv. Mater. **30**, 1800220 (2018).
28. H. Zhang, M. Lin, H. Shi, W. Ji, L. Huang, X. Zhang, S. Shen, R. Gao, S. Wu, C. Tian, Z. Yang, G. Zhang, S. He, H. Wang, T. Saw, Y. Chen, and Q. Ouyang, "Programming a Pavlovian-like conditioning circuit in Escherichia coli," Nat. Commun. **5**, 3102 (2014).
29. I. I. Lederhendler, S. Gart, and D. Alkon, "Classical conditioning of Hermissenda: origin of a new response," J. Neurosci. **6**, 1325–1331 (1986).
30. S. Commins, *Behavioral Neuroscience* (Cambridge University, 2018), Chap. 9, pp. 100–117.
31. I. P. Pavlov, "Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex," Ann. Neurosci. **17**, 136–141 (1927).
32. J. Siegel, A. Schropp, J. Solis, C. N. Afonso, and M. Wuttig, "Rewritable phase-change optical recording in $Ge_2Sb_2Te_5$ films induced by picosecond laser pulses," Appl. Phys. Lett. **84**, 2250 (2004).
33. Y. Xie, W. Kim, Y. Kim, S. Kim, J. Gonsalves, M. BrightSky, C. Lam, Y. Zhu, and J. J. Cha, "Self-healing of a confined phase change memory device with a metallic surfactant layer," Adv. Mater. **30**, 1705587 (2018).
34. A. Pirovano, A. L. Lacaita, A. Benvenuti, F. Pellizzer, S. Hudgens, and R. Bez, "Scaling analysis of phase-change memory technology," in *IEEE International Electron Devices Meeting* (Institute of Electrical and Electronics Engineers (IEEE), 2003), pp. 699–702.
35. Z. Cheng, C. Ríos, W. H. P. Pernice, C. D. Wright, and H. Bhaskaran, "On-chip photonic synapse," Sci. Adv. **3**, e1700160 (2017).
36. D. L. Alkon, K. T. Blackwell, G. S. Barbour, A. K. Rigler, and T. P. Vogl, "Pattern-recognition by an artificial network derived from biologic neuronal systems," Biol. Cybern. **62**, 363–376 (1990).
37. https://course.fast.ai/datasets, Accessed on 27 August 2021.
38. A. Krizhevsky, "Learning multiple layers of features from tiny images," (2009).
39. R. F. L. Fei-Fei and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," in *Conference on Computer Vision and Pattern Recognition Workshop* (IEEE, 2004).
40. A. V. Omkar, M. Parkhi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2012).

41. W. D. J. Deng, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: a large-scale hierarchical image database,” in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009).
42. X. Wang, P. O. Weigei, J. Zhao, M. Ruesing, and S. Mookherjea, “Achieving beyond-100-GHz large-signal modulation bandwidth in hybrid silicon photonics Mach Zehnder modulators using thin film lithium niobate,” APL Photon. **4**, 096101 (2019).
43. V. Stojanovic, D. Markovic, B. Nikolic, M. A. Horowitz, and R. W. Brodersen, “Energy-delay tradeoffs in combinational logic using gate sizing and supply voltage optimization,” in *European Solid-state Devices and Circuits Conference (ESSCIRC): 28th European Solid-State Circuits Conference* (IEEE, 2002).