

# Chalcogenide phase-change devices for neuromorphic photonic computing

Cite as: J. Appl. Phys. **129**, 151103 (2021); <https://doi.org/10.1063/5.0042549>

Submitted: 31 December 2020 • Accepted: 02 April 2021 • Published Online: 21 April 2021

Frank Brücknerhoff-Plückelmann, Johannes Feldmann,  C. David Wright, et al.

## COLLECTIONS

Paper published as part of the special topic on [Plasmonics: Enabling Functionalities with Novel Materials](#)



View Online



Export Citation



CrossMark

## ARTICLES YOU MAY BE INTERESTED IN

[A plasmonically enhanced route to faster and more energy-efficient phase-change integrated photonic memory and computing devices](#)

Journal of Applied Physics **129**, 110902 (2021); <https://doi.org/10.1063/5.0042962>

[Phase change materials in photonic devices](#)

Journal of Applied Physics **129**, 030902 (2021); <https://doi.org/10.1063/5.0027868>

[Photonic tensor cores for machine learning](#)

Applied Physics Reviews **7**, 031404 (2020); <https://doi.org/10.1063/5.0001942>

## Lock-in Amplifiers up to 600 MHz



Zurich  
Instruments



# Chalcogenide phase-change devices for neuromorphic photonic computing

Cite as: J. Appl. Phys. **129**, 151103 (2021); doi: [10.1063/5.0042549](https://doi.org/10.1063/5.0042549)

Submitted: 31 December 2020 · Accepted: 2 April 2021 ·

Published Online: 21 April 2021



Frank Brücknerhoff-Plückelmann,<sup>1</sup> Johannes Feldmann,<sup>2</sup> C. David Wright,<sup>3</sup>  Harish Bhaskaran,<sup>2</sup> and Wolfram H. P. Pernice<sup>1,a)</sup> 

## AFFILIATIONS

<sup>1</sup>Institute of Physics, University of Muenster, Heisenbergstr. 11, 48149 Muenster, Germany

<sup>2</sup>Department of Materials, University of Oxford, Parks Road, OX1 3PH Oxford, United Kingdom

<sup>3</sup>Department of Engineering, University of Exeter, Exeter EX4 4QF, United Kingdom

**Note:** This paper is part of the Special Topic on Plasmonics: Enabling Functionalities with Novel Materials.

**a) Author to whom correspondence should be addressed:** [wolfram.pernice@uni-muenster.de](mailto:wolfram.pernice@uni-muenster.de)

## ABSTRACT

The integration of artificial intelligence systems into daily applications like speech recognition and autonomous driving rapidly increases the amount of data generated and processed. However, satisfying the hardware requirements with the conventional von Neumann architecture remains challenging due to the von Neumann bottleneck. Therefore, new architectures inspired by the working principles of the human brain are developed, and they are called neuromorphic computing. The key principles of neuromorphic computing are in-memory computing to reduce data shuffling and parallelization to decrease computation time. One promising framework for neuromorphic computing is phase-change photonics. By switching to the optical domain, parallelization is inherently possible by wavelength division multiplexing, and high modulation speeds can be deployed. Non-volatile phase-change materials are used to perform multiplications and non-linear operations in an energetically efficient manner. Here, we present two prototypes of neuromorphic photonic computation units based on chalcogenide phase-change materials. First is a neuromorphic hardware accelerator designed to carry out matrix vector multiplication in convolutional neural networks. Due to the neuromorphic architecture, this prototype can already operate at tera-multiply-accumulate per second speeds. Second is an all-optical spiking neuron, which can serve as a building block for large-scale artificial neural networks. Here, the whole computation is carried out in the optical domain, and the device only needs an electrical interface for data input and readout.

Published under an exclusive license by AIP Publishing. <https://doi.org/10.1063/5.0042549>

## INTRODUCTION

In 2016, the computer program “Alpha Go” developed by the British company Google DeepMind beat one of the world’s top players (Lee Sedol) 4:1 at the strategy game GO. Unlike chess, GO cannot be solved deterministically by today’s computers due to the complexity of the game and the fact that a suitable heuristic method to evaluate specific situations does not exist. Therefore, Alpha Go relies on artificial neural networks (ANNs) to play the game GO.<sup>1</sup> Its successor, AlphaZero, just needs the ruleset to learn the game without any additional human input,<sup>2</sup> indicating that computer programs can find solutions and strategies for non-trivial problems on their own. Naturally, artificial neural networks are not only capable of solving well-defined problems in strategic board games but are also heavily deployed in daily life in a wide range of

different applications such as image- and speech recognition, autonomous driving, or medical diagnostics, among others.<sup>3–5</sup>

Even though artificial neural networks (ANNs) lie at the heart of many problem-solving algorithms, providing sufficiently powerful hardware to run them remains challenging due to the large amount of data being processed.<sup>6</sup> In the conventional von Neumann architecture that most processors are based on today, the processing unit is separated from the memory. Consequently, the data need to be shuffled back and forth between both, which leads to a speed barrier known as the von Neumann bottleneck. Moreover, this computing architecture is designed for serial computing, such that the commands are carried out consecutively.<sup>7</sup> Thus, the von Neumann architecture is not optimal to solve data-heavy tasks.

Therefore, new hardware and architectures tailored to ANNs need to be developed. One option is to design application-specific

integrated circuits (ASICs), as, for example, Google's tensor processing unit optimized for matrix vector multiplication (MVM). MVMs are the computationally expensive tasks in many AI applications. Another approach is to build integrated circuits inspired by the working principle of the brain, called neuromorphic computing architectures,<sup>8–12</sup> as biological brains outperform conventional processors in cognitive tasks as speech- and pattern recognition by many orders of magnitude. For example, the simulation of a mouse-scale cortex with  $2.5 \times 10^6$  neurons on a personal computer is 9000 times slower and requires 40 000 times more power than its biological counterpart.<sup>13</sup> Neuromorphic processors aim to work in a highly parallel way and process data directly in memory. Besides several implementations in CMOS electronics, another promising route to building neuromorphic computing systems is to switch to the optical domain. This does not only allow a high degree of parallelization by wavelength division multiplexing but also enables operation speeds up to 100 GHz.<sup>14</sup> This article gives an introduction to photonic approaches to neuromorphic computing.

In the following, we will first provide an overview of artificial neural networks and explain the working principle of convolutional neural networks (CNNs), which are crucial for image classification. Then, we explain how to implement energy-efficient in-memory computing with phase-change devices in photonic integrated circuits. Based on the principles of phase-change photonics, we present a neuromorphic hardware accelerator that is designed to perform the time demanding task of matrix vector multiplication. Finally, we show an all-optical neuron, which can serve as a building block for large-scale neuromorphic artificial neural networks.

## ARTIFICIAL NEURAL NETWORKS

From a mathematical point of view, an ANN is a function  $h_P: \mathbb{R}^m \rightarrow \mathbb{R}^n$ , which is defined via a set of free parameters  $P$ . Depending on how  $P$  is chosen, the neural network can solve a specific problem. In this context, solving means that it assigns the “correct” output activation  $A_{\text{out}}$  to an input activation  $A_{\text{in}}$ , i.e.,  $h_P(A_{\text{in}}) = A_{\text{out}}$ .

Figure 1(a) shows how a (fully connected) ANN is constructed. It consists of several layers, where each layer contains at least one neuron with an associated neuron activation. Each neuron of the  $n$ th layer is connected with all neurons of the  $(n + 1)$ th layer. The input layer is the interface to the real world, and the output layer presents the computational result of the neural network. The elementary building blocks of ANNs are the neurons [see Fig. 1(b)]. First, all the inputs (i.e., the output signals from the neurons in the previous layer) of the  $j$ th neuron are individually weighted by weights  $w_{ij}$  to  $w_{nj}$  and then added together to obtain the activation energy. Afterward, a non-linear function, for example, a rectified linear unit (ReLU) or Sigmoid, is applied to the weighted sum.<sup>15,16</sup> It is important that the activation function of the neurons is non-linear; otherwise, all layers could be condensed to a single layer. The weights are the free parameters  $P$  of the neural network and need to be chosen such that the neural network fulfills the intended function. The process of appropriately choosing the weights is called training. There are several types of trainings, which can be categorized on a basic level into supervised and unsupervised learning. In supervised learning, a training

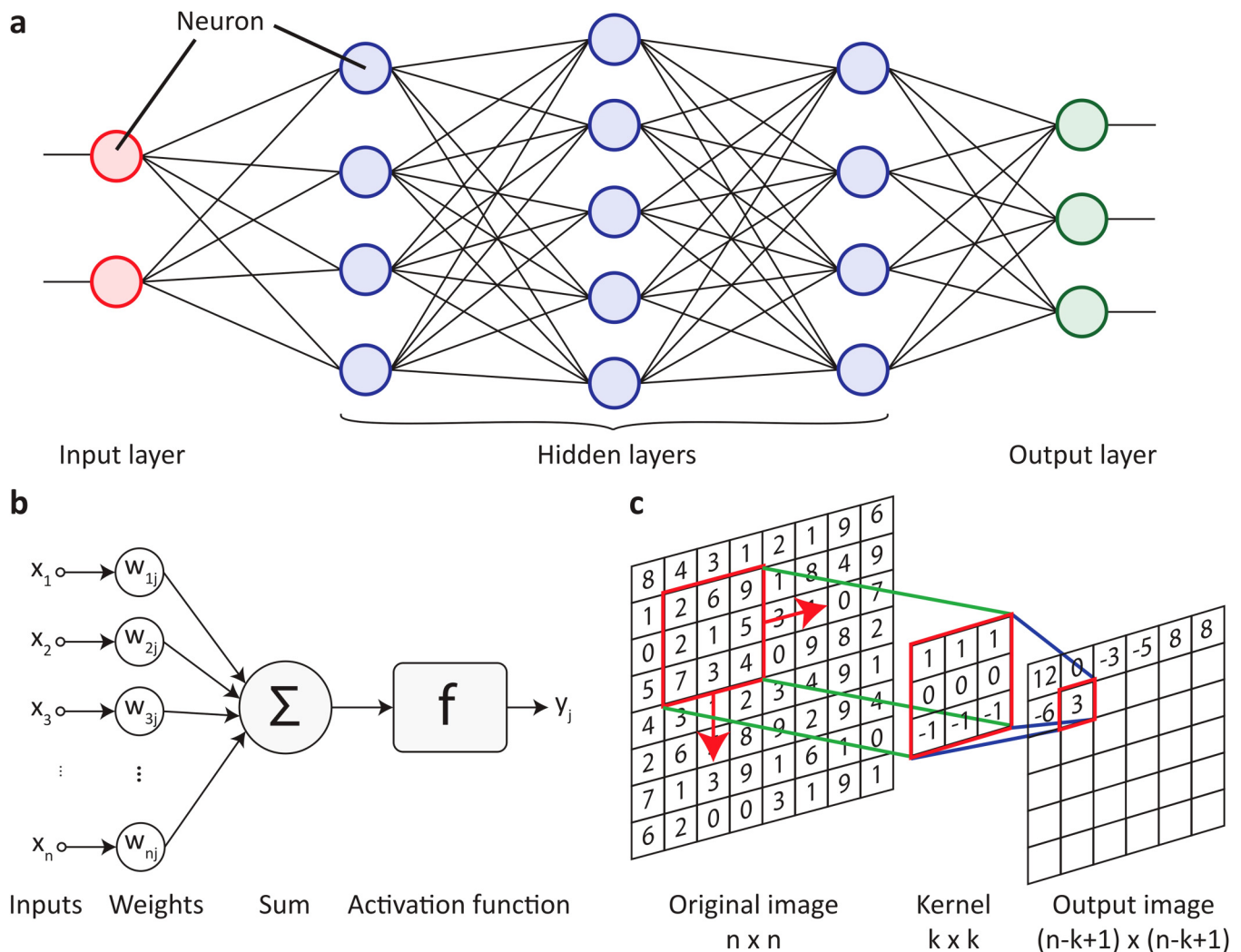
dataset with several pairs  $\{A_{\text{in}}:A_{\text{out}}\}$  must exist and the ANN is fitted to the training dataset. This is typically a very time-demanding task, often implemented with a backpropagation algorithm.<sup>17</sup> Unsupervised learning is applied, when no training set exists and a pattern from an unknown data stream needs to be extracted. This is achieved by implementing a learning rule: for example, inspired by the biological neurons, the Hebbian learning rule “What fires together wires together” can be used.<sup>18</sup>

A main issue of fully connected ANNs is that the number of free parameters tends to be huge. For example, an ANN designed for image classification that takes input images with  $1 \times 10^6$  pixels and has 1000 neurons in the first hidden layer would already have  $1 \times 10^9$  free parameters. Moreover, many hidden layers are deployed in deep neural networks to implement complex functionalities that further increase the number of free parameters.<sup>19</sup> To overcome this challenge and reduce computational complexity, special classes of ANNs have been developed as the aforementioned convolutional neural networks (CNNs). CNNs reduce the number of parameters by introducing a preprocessing step to detect local features between neighboring pixels in the input images. In this step, the image is convolved with several filters, as shown in Fig. 1(c). Those filters are the free parameters of a convolutional layer and are determined during the training process. In the following, we will elucidate how these elementary concepts can be realized with integrated optical or nanophotonic devices in which non-linearity and the capability for learning are implemented with phase-change materials.

## PHASE-CHANGE PHOTONICS

Phase-change photonics is the conjunction between phase-change materials and nanophotonics, which enables integrated photonic circuits (PICs) with novel functionalities. Phase-change materials (PCMs) are materials that can be rapidly switched between an (unordered) amorphous and (ordered) crystalline state and thereby exhibit stark contrast in the optical properties between both phases of matter. The transition between the states is reversible and can be induced via optical or electrical heating. Figure 2(a) schematically shows the switching dynamics of a PCM. If the material is heated (and kept) above the glass transition temperature but below the melting point, the atoms have enough energy to arrange themselves in the energetically preferred crystalline order. If the material is instead further heated above the melting temperature and subsequently rapidly cooled down below the glass transition temperature without giving it time to crystallize, the unordered amorphous state is obtained. Typically, the PCM needs to be cooled down with a rate of 1–100 K/ns to be switched to the amorphous state.<sup>20,21</sup> In the amorphous state, no long-range order is present and covalent bonds between the atoms are dominant.<sup>22</sup> Therefore, the electrons are strongly localized leading to low conductivity. In contrast, resonant bonds between several atoms are formed in the crystalline state leading to highly delocalized electrons and enabling high conductivity.<sup>23,24</sup> Similarly, depending on the stoichiometry, the refractive index also varies greatly. Therefore, PCMs are already used in the field of rewritable optical data storage for decades.<sup>25</sup>

Integrated with photonic circuits, phase-change materials enable active control over phase and amplitude of light propagating



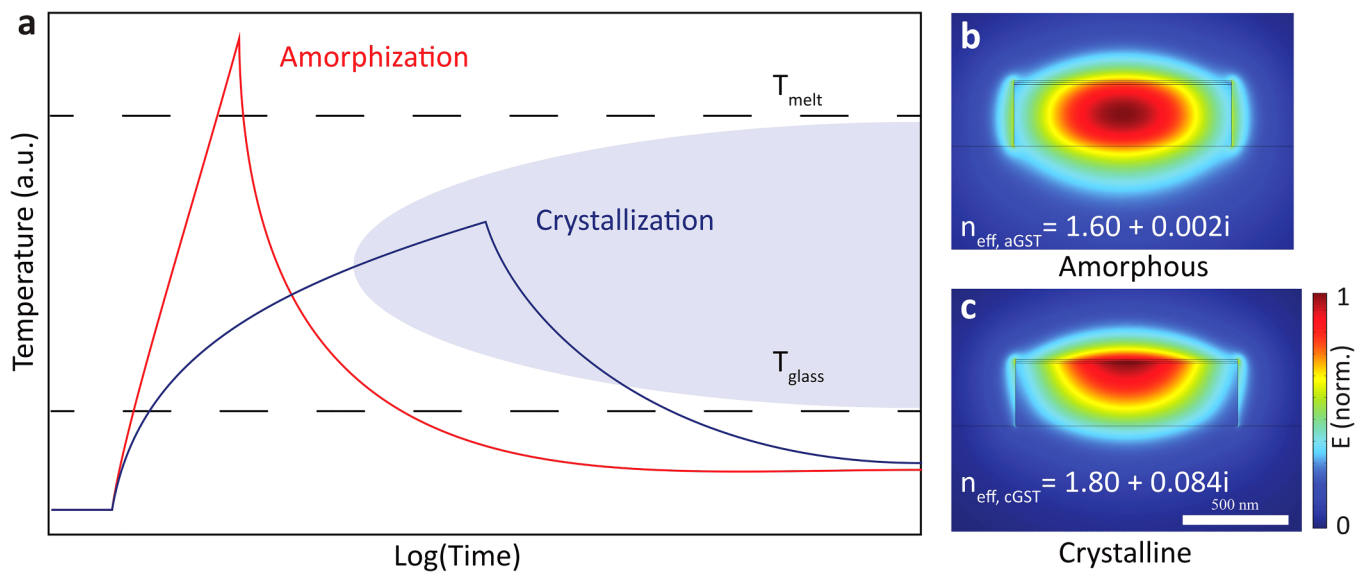
**FIG. 1.** (a) Structure of a fully connected artificial neural network. The neural network consists of several layers, where each layer contains at least one neuron. The input layer is the interface to real world and the output layer presents the computational result of the neural network. (b) Structure of a single neuron. Each neuron has an activation assigned to it, which is calculated from all neurons in the previous layer. First, the activations from the previous layer are individual weighted and then added together. From the sum, a non-linear function  $f$  determines the neuron's activation output. (c) For image classification tasks, convolutional layers are usually applied to the input image first. Those layers convolve the input image with a filter kernel to highlight specific features like edges.

through optical waveguides when evanescently coupled to the PCM. The following devices are based on the well-studied PCM  $\text{Ge}_2\text{Sb}_2\text{Te}_5$  (GST),<sup>26</sup> which belongs to the popular group of chalcogenide solids based on germanium (Ge)–antimony (Sb)–tellurium (Te) alloys. Due to its non-volatility, no static energy supply is required to maintain the PCMs state. In combination with switching energies below 20 pJ, GST enables energy-efficient computation.<sup>27</sup> However, it should be noted that a wide range of PCMs can be found especially in the ternary phase-diagram of Ge:Sb:Te with different properties in terms of switching energies and stability. Also monatomic phase-change materials are developed recently,<sup>28,29</sup> potentially leading to very high switching speeds.

In the telecom wavelength regime at 1550 nm, the refractive indices of the two different GST states are as follows:<sup>30</sup>

$$n_{\text{aGST}} = 3.94 + i0.045, \quad n_{\text{cGST}} = 6.11 + i0.83. \quad (1)$$

Therefore, a thin layer of a PCM is deposited on the top of the waveguide and covered with indium tin oxide (ITO) to prevent oxidation of the PCM. Figures 2(b) and 2(c) show the  $\text{TE}_{00}$  mode profile of 330 nm thick silicon nitride (SiN) on silicon oxide ( $\text{SiO}_2$ ) waveguide with 10 nm GST covered by 10 nm ITO. In the amorphous state, the imaginary part of the effective index is significantly lower than in the crystalline state, leading to a large absorption



**FIG. 2.** (a) Switching dynamics of phase change materials. If the material is heated above the glass transition temperature but not melted, the atoms have sufficient energy to rearrange in a crystal lattice. The PCM becomes amorphous instead, if it is first melted and then cooled down rapidly, to freeze the atoms in the unordered (amorphous) state. (b) and (c) Optical mode profiles of a silicon nitride waveguide (330 nm height) with a 10 nm layer of GST and a 10 nm ITO capping layer on top. Due to the refractive index difference between the amorphous and crystalline state, also the effective index of the guided mode can be varied.

contrast between both modes,

$$\alpha_{\text{SiN-aGST}} = -0.07 \frac{\text{dB}}{\mu\text{m}}, \quad \alpha_{\text{SiN-cGST}} = -2.96 \frac{\text{dB}}{\mu\text{m}}. \quad (2)$$

Using standard lithography processes, the PCM can be selectively deposited in specific areas of the waveguide, enabling two functionalities: first, the transmission of the waveguide can be locally varied by partially switching the GST patch between the amorphous and highly absorptive crystalline state. This can either be done electrically using external heater structures<sup>31</sup> or optically since the light absorbed in the GST will heat it up.<sup>32,33</sup> For optical switching, up to 64 intermediate transmission states in a GST patch have been demonstrated.<sup>34</sup> Second, the GST can be used as a non-linear element inside a photonic circuit, due to the threshold behavior of the switching process.

Overall, phase-change photonics enable the realization of non-volatile memory cells and non-linear power-dependent elements inside an integrated photonic circuit. The low energy consumption and passive interaction behavior between a light pulse and the PCM make phase-change photonics an ideal building block for high-speed neuromorphic computing.

### NEUROMORPHIC HARDWARE ACCELERATOR

A first step to true neuromorphic computing is to build neuromorphic photonic integrated circuits for mathematical operations that are time demanding in the conventional von Neumann architecture. In a fully connected layer of an ANN, the activations from the previous layer need to be weighted with various weights and

accumulated. In a convolutional layer, the activation from the previous layer is convolved with several filters. Both operations can be written in the form of matrix vector multiplications, which are therefore often a bottleneck for computing ANNs.

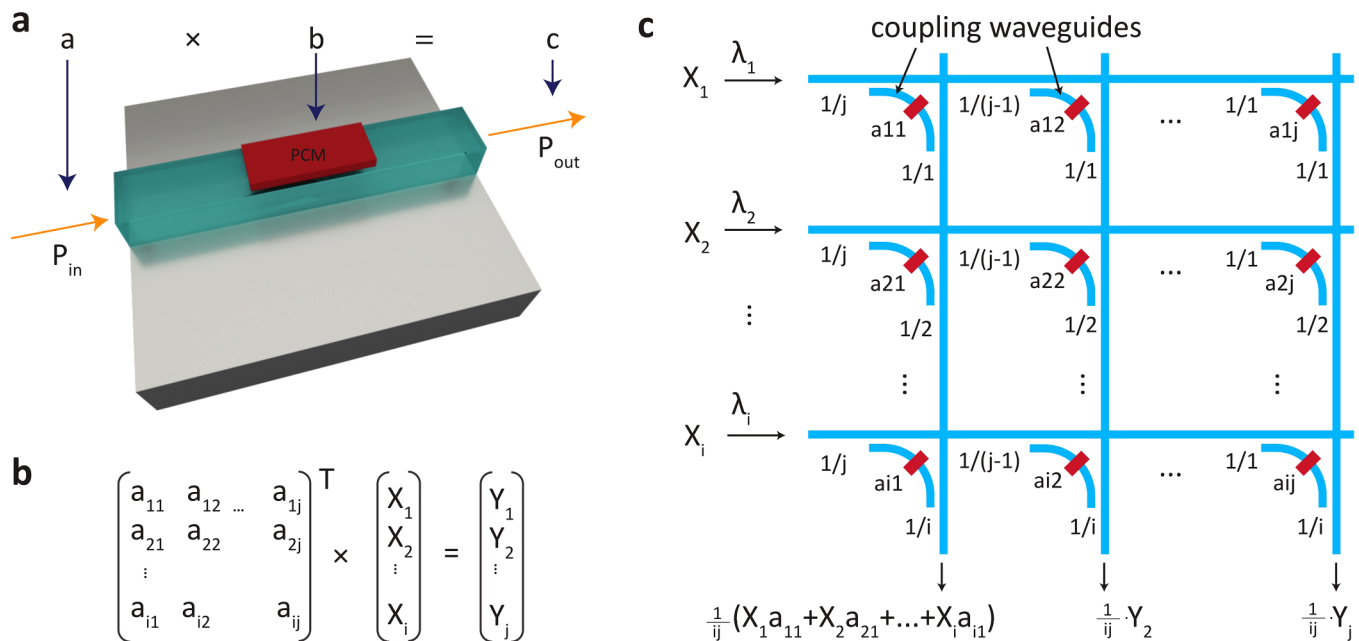
In order to perform MVMs with a PIC, several multiply and accumulate (MAC) operations need to be carried out. Figure 3(a) shows how the multiplication of an (fast modulated) input pulse with power  $P_{\text{in}}$  and a (tunable) matrix element is carried out with phase-change material cells. Depending on the phase state, transmission  $T$  through the PCM cell can be set; consequently, the power of the transmitted pulse is  $P_{\text{out}} = T \times P_{\text{in}}$ . Since the PCM is only evanescently coupled to the waveguide leading to absorption, the multiplication time is just the time the pulse needs to propagate through the PCM cell.

In order to add the power of several pulses (multiplication results) together, they are overlapped in a single waveguide. However, two coherent laser beams with frequency  $w_1$  and  $w_2$  propagating in the same direction with same polarization will interfere with each other,

$$E \sin(w_1 t) + E \sin(w_2 t) = 2E \sin\left(\frac{w_1 + w_2}{2} t\right) \cos\left(\frac{w_1 - w_2}{2} t\right). \quad (3)$$

The beat term consists of two parts, the fast oscillating one with frequency  $(w_1 + w_2)/2$  and a slow oscillating one with frequency  $(w_1 - w_2)/2$ . The fast oscillating one will be averaged out by a photodetector. However, the slow oscillating one can be visible, depending on the detuning and detector bandwidth. Therefore, in order to avoid oscillations, the accumulated laser





**FIG. 3.** (a) Scalar multiplication carried out using a PCM cell. Here, the first factor is encoded in the power of the light pulse and the second factor in the transmission level of the PCM. The product of both factors can be obtained from the amplitude of the output signal. (b) Mathematical operation carried out by the PIC shown in (c). Each component of the matrix is encoded in a different PCM cell, whereas the input vector is modulated on the power of the incoming light. (c) Sketch of the PIC used for matrix-vector multiplication. The components of the input vector are sent into different rows. With directional couplers, the signal is equally split between the columns and individually weighted by the corresponding PCM cell. In the output waveguide, the signals from the different rows are added incoherently.

pulses need different wavelengths with sufficient detuning. In the following, this approach to add two coherent lasers is called “incoherent” accumulation. The opposing way would be to use two lasers at the exact same wavelength and add them together “coherently” by fixing the phase relation between both.

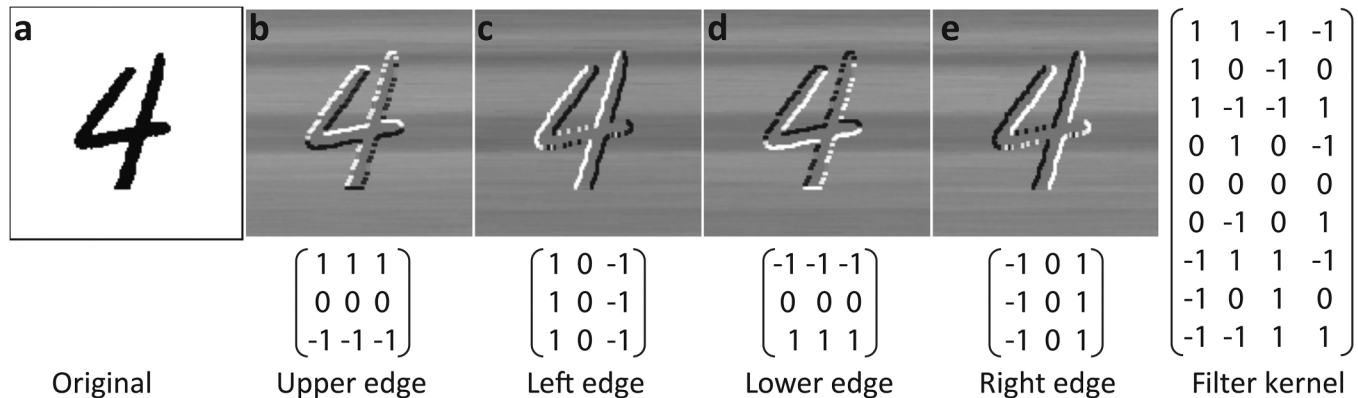
There are several approaches to add two signals incoherently. One way is using multiplexing techniques, such as wavelength-division multiplexing (WDM) or mode multiplexing. The advantage of this approach is that it is theoretically lossless. However, multiplexing requires very precise fabrication and potentially a way to actively tune the devices afterwards, due to the sensitivity of, e.g., ring resonators and Bragg filters used for WDM.<sup>35</sup> Moreover, especially narrow band Bragg filters are large, increasing the footprint of the PIC. Another option is to combine two different signals with directional couplers. While this method relaxes the requirements for the fabrication, it unavoidably leads to optical losses.

Figure 3(c) shows a PIC that is designed to carry out MVMs as described in Fig. 3(b). The multiplication is carried out by choosing the pulse height and the PCM’s transmission state, and the pulses are added together onto a common waveguide with directional couplers. The different inputs  $X_1$  to  $X_i$  have different wavelengths  $\lambda_1$  to  $\lambda_i$  leading to  $j$  outputs. A fixed fraction of the input light in the input rows is transferred to each coupling waveguide that connects the horizontal input waveguides to the vertical output waveguides. After a fraction of the incoming pulse is transferred to the coupling waveguide, it is partially absorbed by the PCM cell to carry out the

multiplication. Finally, light is coupled from the coupling waveguide into the vertical output waveguide. In order to ensure that all matrix cells contribute equally to the final power in the output waveguide, the coupling fraction for the different cells has to be chosen properly. Therefore, only  $1/ij$  of the input power of the first waveguide will reach the output waveguide. Here,  $1/i$  is attributed to losses caused by the directional couplers and  $1/j$  to the number of columns. We term this architecture photonic tensor core (PTC).

The advantage of a PTC is that it is a completely passive device (PCM is non-volatile) and therefore does not require any energy to preserve the matrix state. The calculation is carried out in a transmission measurement. Figure 4 shows the experimental result of four different convolution operations [see Fig. 1(c)] calculated with the neuromorphic hardware accelerator. As designed, it clearly detects the upper/lower and left/right edges of the input picture. This basic hardware accelerator was then used as a part of a convolutional network and tested with the MNIST database of handwritten digits. For the chosen CNN, the optimal prediction efficiency is 96.1%. When using the hardware accelerator instead of a conventional PC to carry out convolution, the efficiency only slightly drops to 95.3%.<sup>36</sup>

By employing a second tier of multiplexing, several matrix vector multiplications can be carried out in parallel without changing the PTC itself. In this case, the first vector uses the wavelengths  $\lambda_1$  to  $\lambda_i$  and the second  $\lambda_{i+1}$  to  $\lambda_{2i}$  and so on. By demultiplexing the signal at the output of the PTC accordingly, the results of the



**FIG. 4.** Convolution operations calculated with a photonic neuromorphic hardware accelerator. The input image (a) is convolved with four different filters (b)–(e) that detect edges to the upper/lower and left/right side. The  $9 \times 4$  filter kernel needed to carry out all convolutions in parallel is then written into the neuromorphic hardware accelerator and the input image subsequently passed to it.<sup>36</sup>

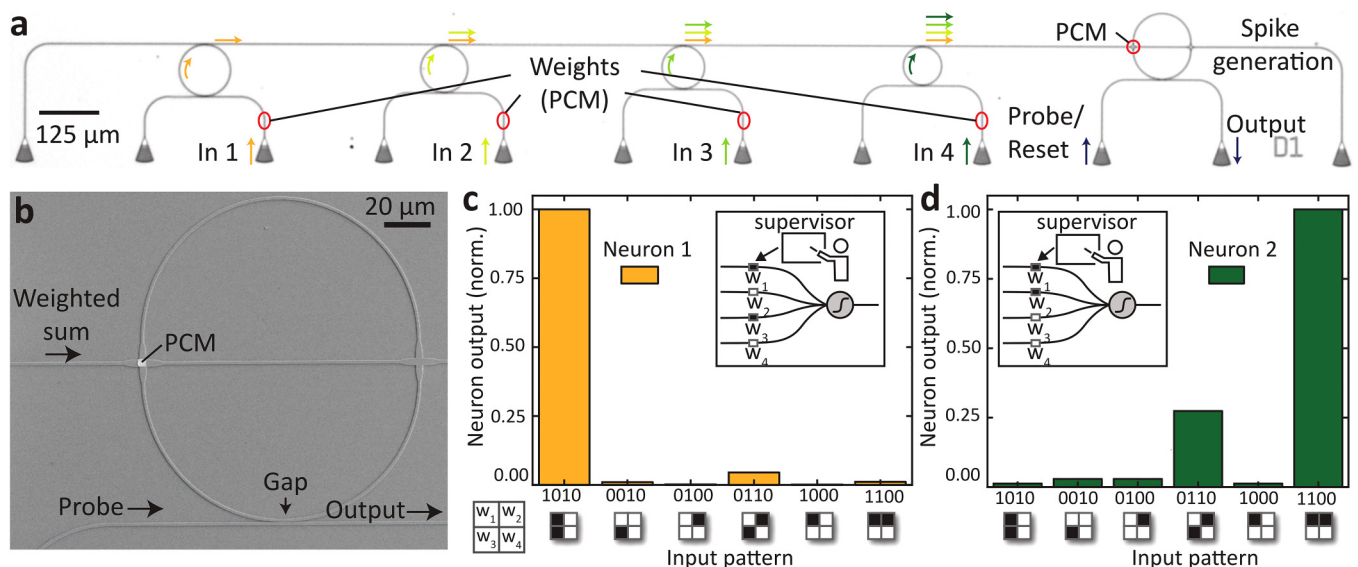
individual MVMs can be obtained in parallel. The total bandwidth is only limited by the wavelength dependency of the directional couplers. In the experiments, the photonic neuromorphic hardware accelerator was able to operate at a rate up to 2 TMAC/s.

### ALL OPTICAL SPIKING NEURONS

A further step toward all-optical neuromorphic computing is to perform the entire data processing in a photonic integrated

circuit. In order to calculate an ANN, several MAC operations must be carried out first to weight and accumulate the input from the previous layer. Afterward, a non-linear activation function determines the neurons' activation.

Figure 5(a) shows a photonic circuit designed to mimic a single neuron with four inputs. The same principles as in the photonic hardware accelerator are used for carrying out the MAC operations: multiplication is achieved with PCM cells and afterward incoherent addition of the weighted inputs is carried out. In this



**FIG. 5.** (a) The photonic implementation of an artificial neural network with four inputs and one output neuron. The inputs In 1–In 4 are weighted with the weights  $w_1$ – $w_4$  via PCM cells. Afterward, the input signals are added together via a ring multiplexer. Finally, the activation unit determines the output neuron activation. (b) SEM image of the deployed activation unit. Initially, the PCM cell is in the crystalline state and the probe signal absorbed in the ring resonator. If the weighted sum of all inputs exceeds a power threshold, which is set by the melting temperature of the GST, the PCM becomes amorphous. Now, the probe pulse is mainly transmitted. (c) and (d) The neuron is trained to detect different pattern sent into the neuron via laser pulses. The ANN can clearly distinguish between the trained pattern and random patterns.<sup>37</sup>

alternative WDM framework, here ring resonators are deployed instead of directional couplers to add the weighted inputs. The obtained weighted input power is sent to the activation unit that is shown in Fig. 5(b). The activation unit consists of a ring resonator with an integrated PCM cell and a probe waveguide (fixed wavelength). In the beginning, the PCM cell is crystalline and the probe pulse on resonance is mainly absorbed in the ring resonator. If the total weighted input power exceeds the threshold set by the melting temperature of the PCM, the PCM amorphizes. This reduces the losses inside the ring resonator and therefore the maximal extinction ratio and shifts the resonance frequency because of the change in both the real and imaginary parts of the refractive index. Now the probe pulse that previously was on resonance with the resonator is mainly transmitted. A switching contrast of up to 10 dB can be achieved in this way.<sup>37</sup> Figures 5(c) and 5(d) show the experimental result of measurements performed with this type of artificial photonic neuron. In both cases, the neuron is trained to detect a specific pattern, and in both cases, it can clearly distinguish between the desired and various different patterns. The shown neuron design can serve as a building block for larger multilayer neural networks. In this case, the output pulse of the neuron in Fig. 5(a) can serve as an input to the neurons in the next layer. Moreover, unsupervised learning according to a Hebbian-like learning rule is possible by overlapping the output pulse with the input pulse in the PCM weights. By doing so, the weights will change depending on whether the neuron fires together with an input pulse or not.<sup>37</sup>

Since the PCM in the activation unit is switched continuously, cycle-to-cycle variations are present like in the electrical counterpart.<sup>38</sup> However, for the operation of the optical neuron, a certain degree of noise can even be beneficial to avoid local minima in the training process and are also present in biochemical neural networks.<sup>39</sup> Additionally, neuromorphic circuits are comparably tolerant to small variations of the weights and inputs, allowing for the precision of the calculations to be reduced significantly.<sup>40</sup> After training the neuron, the PCM weights are stable over months<sup>27</sup> since the transmission of the PCM cell does not depend on conductive filaments in the PCM but results from the evanescent coupling between the waveguide and the PCM. Therefore, spatial variations in the PCM state, which are small in comparison to the optical wavelength, are averaged and therefore have a small impact on the overall transmission.

## CHALLENGES AND OUTLOOK

Recent work on neuromorphic computing demonstrates prospects of building brain inspired photonic integrated circuits. Nevertheless, there are several challenges to overcome before they can commercially challenge conventional architectures in the field of artificial intelligence.

Even though a photonic hardware accelerator can theoretically reach unprecedented performance in the PMAC/s range for a single matrix,<sup>36</sup> the device footprint is substantially larger than electronic hardware. The silicon nitride waveguides deployed in both presented PICs have a width of 1.2  $\mu\text{m}$ , and thus, the total device size is on the order of square millimeters to square centimeters. In comparison, nowadays electric circuits can be fabricated in a 5 nm process. However, this is the result of decades of commercial optimization,

starting from 10  $\mu\text{m}$  MOSFET processes in the 1970s. There are several approaches of how to reduce the footprint of PICs. First, one can use another platform with a higher refractive index contrast than SiN on SiO<sub>2</sub> leading to smaller waveguides and smaller possible bend radii as, for example, silicon on insulator (SOI).<sup>41</sup> Moreover, one could build the PIC not only in a plane but use multilayer processes to move toward 3D architectures.<sup>42</sup> The larger footprint of photonic circuits compared to electronics can also be compensated by the ability to multiplex signals on different wavelengths (a feature that is not available in electronics). This way the same circuit can be used for different computations at the same time increasing the computational density and parallelism.

Furthermore, for a fully functional system, various optical components need to be integrated on chip and a sufficient interface needs to be provided before it can be used commercially. In the examples outlined above, only the computational unit itself is integrated on the chip, whereas the required laser sources, multiplexer, modulators, and detectors remain off chip. This makes it challenging to scale and is unfeasible to use outside laboratories. Switching to a different platform, for example, InP or SOI, is a promising route to integrate all components on the chip. Finally, both designs need an electrical interface, in order to make it compatible with existing technology. Since the PIC works in the analog domain and conventional electronics is digital, digital to analog converters play a crucial role.

Overall, neuromorphic computing implemented in photonic integrated circuits using phase-change devices is a promising way to satisfy the rapidly growing computational demands of artificial intelligence. Due to the high modulation speeds achievable in the optical domain and the inherent capability for parallelization via multiplexing, it is well suited to process the large amount of data in artificial neural networks. In-memory computing with non-volatile phase-change materials also enables an overall energy-efficient process. The next step will be to move the experimental designs from the laboratories to commercial applications.

## DATA AVAILABILITY

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

## ACKNOWLEDGMENTS

This research was supported by EPSRC via Grant Nos. EP/J018694/1, EP/M015173/1, and EP/M015130/1 in the United Kingdom and the Deutsche Forschungsgemeinschaft (DFG) under Grant No. PE 1832/5-1 in Germany. W.P. gratefully acknowledges support by the European Research Council through Grant No. 724707. We further acknowledge funding for this work from the European Union's Horizon 2020 Research and Innovation Program (Fun-COMP project, No. 780848 and PheMtronics project, No. 899598).

## REFERENCES

- <sup>1</sup>D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, *Nature* **529**, 484 (2016).



- <sup>2</sup>D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. Van Den Driessche, T. Graepel, and D. Hassabis, *Nature* **550**, 354 (2017).
- <sup>3</sup>N. D. Lane, S. Bhattacharya, A. Mathur, P. Georgiev, C. Forlivesi, and F. Kawsar, *IEEE Pervasive Comput.* **16**, 82 (2017).
- <sup>4</sup>M. A. A. Babiker, M. A. O. Elawad, and A. H. M. Ahmed, in *Proceedings of the International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)* (IEEE, 2019).
- <sup>5</sup>F. Amato, A. López, E. M. Peña-Méndez, P. Vañhara, A. Hampl, and J. Havel, *J. Appl. Biomed.* **11**, 47 (2013).
- <sup>6</sup>See [https://www.mckinsey.com/~/media/McKinsey/Industries/Semiconductors/Our Insights/Artificial intelligence hardware New opportunities for semiconductor companies/Artificial-intelligence-hardware.pdf](https://www.mckinsey.com/~/media/McKinsey/Industries/Semiconductors/Our%20Insights/Artificial%20intelligence%20hardware%20New%20opportunities%20for%20semiconductor%20companies/Artificial-intelligence-hardware.pdf) for “The McKinsey report” on Artificial-intelligence hardware: New opportunities for semiconductor companies which gives an outlook on the development of the AI market size and demands.
- <sup>7</sup>J. Backus, *Commun. ACM* **21**, 613 (1978).
- <sup>8</sup>Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic, *Nat. Photonics* **11**, 441 (2017).
- <sup>9</sup>T. Ferreira de Lima, B. J. Shastri, A. N. Tait, M. A. Nahmias, and P. R. Prucnal, *Nanophotonics* **6**, 577 (2017).
- <sup>10</sup>D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, *Nat. Commun.* **4**, 1364 (2013).
- <sup>11</sup>Q. Vinckier, F. Duport, A. Smerieri, K. Vandoorne, P. Bienstman, M. Haelterman, and S. Massar, *Optica* **2**, 438 (2015).
- <sup>12</sup>S. Schmitt, J. Klähn, G. Bellec, A. Grübl, M. Güttler, A. Hartel, S. Hartmann, D. Husmann, K. Husmann, S. Jeltsch, V. Karasenko, M. Kleider, C. Koke, A. Kononov, C. Mauch, E. Müller, P. Müller, J. Partzsch, M. A. Petrovici, S. Schiefer, S. Scholze, V. Thanasoulis, B. Vogginger, R. Legenstein, W. Maass, C. Mayr, R. Schüffny, J. Schemmel, and K. Meier, *arXiv:1703.01909*.
- <sup>13</sup>B. V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J. M. Bussat, R. Alvarez-Icaza, J. V. Arthur, P. A. Merolla, and K. Boahen, *Proc. IEEE* **102**, 699 (2014).
- <sup>14</sup>P. O. Weigel, J. Zhao, K. Fang, H. Al-Rubaye, D. Trotter, D. Hood, J. Mudrick, C. Dallo, A. T. Pomerene, A. L. Starbuck, C. T. DeRose, A. L. Lentine, G. Rebeiz, and S. Mookherjee, *Opt. Express* **26**, 23728 (2018).
- <sup>15</sup>J. Feng and S. Lu, *J. Phys.: Conf. Ser.* **1237**, 022030 (2019).
- <sup>16</sup>P. Sibi, S. Allwyn Jones, and P. Siddarth, *J. Theor. Appl. Inf. Technol.* **47**, 1344 (2013).
- <sup>17</sup>Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Proc. IEEE* **86**, 2278 (1998).
- <sup>18</sup>G. Q. Bi and M.-m. Poo, *Annu. Rev. Neurosci.* **24**, 139 (2001).
- <sup>19</sup>J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. A. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng, *Adv. Neural Inf. Process. Syst.* **2**, 1223 (2012).
- <sup>20</sup>C. Peng, L. Cheng, and M. Mansuripur, *J. Appl. Phys.* **82**, 4183 (1997).
- <sup>21</sup>A. Redaelli, A. Pirovano, A. Benvenuti, and A. L. Lacaita, *J. Appl. Phys.* **103**, 111101 (2008).
- <sup>22</sup>G. Lucovsky and R. M. White, *Phys. Rev. B* **8**, 660 (1973).
- <sup>23</sup>D. Lencer, M. Salinga, and M. Wuttig, *Adv. Mater.* **23**, 2030 (2011).
- <sup>24</sup>M. Wuttig, H. Bhaskaran, and T. Taubner, *Nat. Photonics* **11**, 465 (2017).
- <sup>25</sup>M. Wuttig and N. Yamada, *Nat. Mater.* **6**, 824 (2007).
- <sup>26</sup>S. Abdollahramezani, O. Hemmatyar, H. Taghinejad, A. Krasnok, Y. Kiarashinejad, M. Zandehshahvar, A. Alù, and A. Adibi, “Tunable nanophotonics enabled by chalcogenide phase change materials,” *arxiv:2001.06335*.
- <sup>27</sup>J. Feldmann, M. Stegmaier, N. Gruhler, C. Rios, H. Bhaskaran, C. D. Wright, and W. H. P. Pernice, “Calculating with light using a chip-scale all-optical abacus,” *Nat. Commun.* **8**, 1256 (2017).
- <sup>28</sup>M. Salinga, B. Kersting, I. Ronneberger, V. P. Jonnalagadda, X. T. Vu, M. Le Gallo, I. Giannopoulos, O. Cojocar-mirédin, R. Mazzarello, and A. Sebastian, *Sebastian, Nat. Mater.* **17**, 681 (2018).
- <sup>29</sup>Z. Cheng, T. Milne, P. Salter, J. S. Kim, S. Humphrey, M. Booth, and H. Bhaskaran, *Sci. Adv.* **7**, 1 (2021).
- <sup>30</sup>M. Stegmaier, C. Ríos, H. Bhaskaran, and W. H. P. Pernice, *ACS Photonics* **3**, 828 (2016).
- <sup>31</sup>J. Zheng, Z. Fang, C. Wu, S. Zhu, P. Xu, J. K. Doylend, S. Deshmukh, E. Pop, S. Dunham, M. Li, and A. Majumdar, *Adv. Mater.* **32**, 1 (2020).
- <sup>32</sup>Z. Cheng, C. Ríos, W. H. P. Pernice, C. David Wright, and H. Bhaskaran, *Sci. Adv.* **3**, 1 (2017).
- <sup>33</sup>X. Li, N. Youngblood, C. Ríos, Z. Cheng, C. D. Wright, W. H. Pernice, and H. Bhaskaran, *Optica* **6**, 1 (2019).
- <sup>34</sup>C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, and M. Li, *arXiv:2004.10651*.
- <sup>35</sup>D. Mu, H. Qiu, J. Jiang, X. Wang, Z. Fu, Y. Wang, X. Jiang, H. Yu, and J. Yang, *IEEE Photonics J.* **11**, 1 (2019).
- <sup>36</sup>J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, 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, *Nature* **589**, 52 (2021).
- <sup>37</sup>J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, *Nature* **569**, 208–214 (2019).
- <sup>38</sup>N. Gong, T. Idé, S. Kim, I. Boybat, A. Sebastian, V. Narayanan, and T. Ando, *Nat. Commun.* **9**, 1 (2018).
- <sup>39</sup>T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian, and E. Eleftheriou, *Nat. Nanotechnol.* **11**, 693 (2016).
- <sup>40</sup>S. Gupta, A. Agrawal, K. Gopalakrishnan, Y. Heights, P. Narayanan, and S. Jose, *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.* **37**, 1737 (2015).
- <sup>41</sup>Y. A. Vlasov and S. J. McNab, *Opt. Express* **12**, 1622 (2004).
- <sup>42</sup>P. Koonath and B. Jalali, *Opt. InfoBase Conf. Pap.* **15**, 12686 (2007).