iScience

CellPress
OPEN ACCESS

Perspective

Integration and Co-design of Memristive Devices and Algorithms for Artificial Intelligence

Wei Wang,[1] Wenhao Song,[2] Peng Yao,[2] Yang Li,[3] Joseph Van Nostrand,[4,*] Qinru Qiu,[5,*] Daniele Ielmini,[1,*] and J. Joshua Yang[2,6,*]

SUMMARY

Memristive devices share remarkable similarities to biological synapses, dendrites, and neurons at both the physical mechanism level and unit functionality level, making the memristive approach to neuromorphic computing a promising technology for future artificial intelligence. However, these similarities do not directly transfer to the success of efficient computation without device and algorithm co-designs and optimizations. Contemporary deep learning algorithms demand the memristive artificial synapses to ideally possess analog weighting and linear weight-update behavior, requiring substantial device-level and circuit-level optimization. Such co-design and optimization have been the main focus of memristive neuromorphic engineering, which often abandons the "non-ideal" behaviors of memristive devices, although many of them resemble what have been observed in biological components. Novel brain-inspired algorithms are being proposed to utilize such behaviors as unique features to further enhance the efficiency and intelligence of neuromorphic computing, which calls for collaborations among electrical engineers, computing scientists, and neuroscientists.

INTRODUCTION

Artificial intelligence (AI) has made great progress in recent years with the help of the advances of deep learning (DL) technologies (LeCun et al. 2015). However, owing to the high volume of data needed to be frequently transferred between processing units and memories, the performance of deep learning algorithms is limited by the von Neumann bottleneck in conventional computers. The existing von Neumann bottleneck can be overcome by in-memory computing with memristive devices, where the computation takes place in the analog domain in the memory, right at the data location. Memristive devices, existing in several forms, such as resistive switching random access memory (RRAM), phase-change memory (PCM), magnetic random access memory (MRAM), and ferroelectric random access memory (FeRAM) (Figure 1C), have tunable conductance states, similar to the plasticity of biological synapses (Ielmini and Wong, 2018; Wang et al., 2020c), and thus can enable in-memory computing, in analogy to the biological neural system. Thanks to their scalability, stacking-ability, simple device structure, and other intriguing properties, the memristive devices have been considered as leading candidates for synaptic devices for hardware implementation of neural networks and machine learning, providing an energy-efficient and low-latency solution for future AI (Yang et al., 2013; Ielmini and Wong, 2018).

With their ionic transport mechanisms similar to the molecular activities in biological intelligent systems, the memristive devices also exhibit rich dynamics, resembling synaptic and neuronal dynamics found in biological systems. For instance, stochastic switching (Mizrahi et al., 2018), pulse-pair facilitation (Wu et al., 2018), and short-term plasticity (Ohno et al., 2011) have been observed in memristors and are suitable to reproduce the dynamics of biological synapses. This may eventually lead to computational systems capable of faithfully emulating the information representation and processing in the brain with much-improved energy efficiency and fidelity over the conventional systems.

Previous efforts to implement these algorithms that take advantage of memristive devices have met with limited success for two reasons: (1) Current state-of-the-art DL technologies use digitalized values (floating-point in software solutions or integers in CMOS-based hardware accelerators) for connection weights. Representing these

[1]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano and IU.NET, Piazza L. da Vinci 32, Milano 20133, Italy

[2]Electrical and Computer Engineering Department, University of Southern California, Los Angeles, CA, USA

[3]The Andrew and Erna Viterbi Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel

[4]Air Force Research Laboratory, Information Directorate, Rome, NY, USA

[5]Electrical Engineering and Computer Science Department, Syracuse University, NY, USA

[6]Lead Contact

*Correspondence:
joseph.vannostrand.1@us.af.mil (J.V.N.),
qiqiu@syr.edu (Q.Q.),
daniele.ielmini@polimi.it (D.I.),
jjoshuay@usc.edu (J.J.Y.)

https://doi.org/10.1016/j.isci.2020.101809

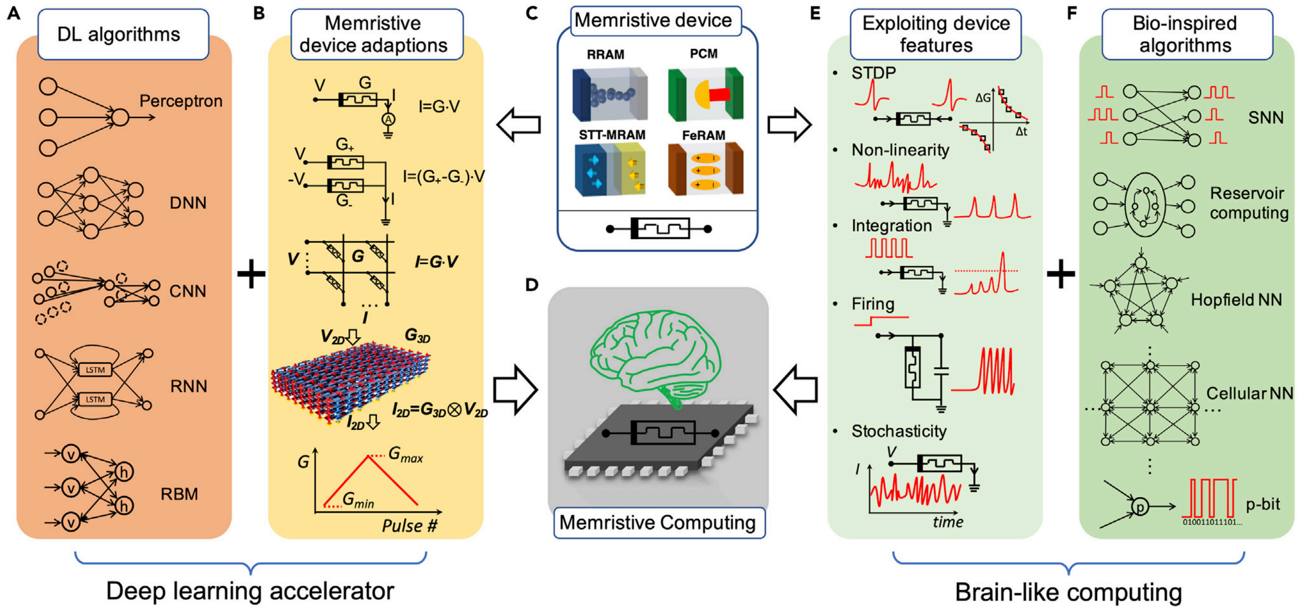


Figure 1. Integration of Learning Algorithms and Memristive Devices for Memristive Neuromorphic Computing

(A) Various algorithms of deep learning (DL) neural networks, including simple perceptron, deep (multiplayer) neural network (DNN), convolutional neural network (CNN), recurrent neural network (RNN), and restricted Boltzmann machine (RBM).

(B) Adaptions and performance enhancements of memristive devices for their synaptic application in DL algorithms: linear read of the memristive device, realization of both positive and negative weight via differential pair, mapping vector matrix multiplication and accumulation (VMMA) to 2D and 3D memristive array, and linear conductance update for identical pulses. Reproduced from (Lin et al., 2020), copyright © 2020, Springer Nature.

(C) Various memristive devices, including resistive random-access memory (RRAM), phase-change memory (PCM), spin-torque transfer random access memory (STT-RAM), and ferroelectric random access memory (FeRAM).

(D) Illustration of a memristive neuromorphic computing system integrated within a monolithic chip.

(E) Unique memristive device features for emerging algorithms, such as the capability of spike-timing dependent plasticity (STDP), nonlinearity for filtering dendrites, integration and firing functions for memristive neurons, and stochasticity.

(F) Emerging algorithms and architectures can be implemented by memristive devices, including spiking neural network (SNN), reservoir computing, Hopfield neural network (Hopfield NN), cellular neural network (Cellular NN), and a probabilistic bit (p-bit).

weights using memristive conductance has suffered from non-ideal behaviors. (2) There are currently no complete algorithms to exploit the bio-plausible behaviors of the memristive devices. As a result, there exist substantial mismatches between algorithms and device properties, and therefore the integration of learning algorithms and memristive devices via co-design is imperative. Specifically, mismatches between memristive devices and learning algorithms can be addressed not only from the device side to engineer the materials for "expected" properties, which has been pursued intensively so far, but also from the algorithm side to either compensate the nonidealities of the devices or exploit some of the "unexpected" properties as valuable features for new types of computing, which has been relatively less explored.

In this perspective, recent achievements in the co-design of memristive devices and learning algorithms are reviewed, to provide a comprehensive overview of the status and remaining challenges for future explorations. We first give an overview of the co-design efforts of learning algorithms and memristive devices (section "The Integration of Learning Algorithms and Memristive Devices"). We then review such co-design efforts in depth from three standpoints: (1) compensating the device nonidealities at the single synapse level and memristive array level in vector-matrix multiplication and accumulation (VMMA) (section "DL Accelerators by Memristive Hardware "); (2) exploiting unique memristive device features in various bio-inspired learning algorithms (section "Exploiting Memristive Properties for Brain-Inspired Algorithms"); and (3) constructing brain-like computing systems with the bio-inspired algorithms enabled by unique memristive features (section "Toward a Brain-like Computing System").

The Integration of Learning Algorithms and Memristive Devices

The artificial neural network (ANN) has a lot of variants, such as the simple perceptron, deep neural network (DNN), deep convolutional neural network (CNN), recurrent neural network (RNN), and deep belief neural network based on restricted Boltzmann machines (RBMs) (Figure 1A), all belonging to the large family of DL techniques (LeCun et al., 2015). The operation of the ANN can be divided into two stages, inference and learning. During the inference phase, each layer of the neural network transforms the input signals by multiplication with the synaptic weights, summation at each output neuron, and activation according to a non-linear function. During the learning phase, the network is trained with data to adjust the synaptic weights for correct inference. Most of the learning tasks can be technically divided into three categories, supervised learning, unsupervised learning, and reinforcement learning, depending on the type and availability of feedback. In most approaches, an objective function or loss function is defined for weight training, that is, how good (or bad) the current weight configuration is, to fulfill the application (e.g., classification or decision) requirement, and the goal of training is to minimize the loss function or maximize the reward.

The most successful learning rule so far is error backpropagation (BP), where the loss function (or error) in the last layer is back-propagated to the preceding layers via the synaptic networks (Rumelhart et al., 1986). BP solves the credit (or blame) assignment problem, i.e., the weight updates to decrease the error, by calculating the gradients of the objective function with respect to the network parameters.

In the DL networks of Figure 1A, the massive synaptic connections and matrix-vector multiplications can be implemented physically within the memristive devices as illustrated in Figure 1B. Memristive devices perform the in-memory computation of the synaptic weighting functions in both forward inference and backward error propagation. Thanks to their ability to nanoscale miniaturization (Figure 1C), hardware acceleration with high efficiency for DL techniques can be achieved (Figure 1D). However, several technical issues still need to be addressed to commercialize such memristive DL accelerators at large scales.

Spiking neural networks (SNNs) take inspirations from neuroscience (Maass 1997) (Figure 1F), by preserving a more biological behavior of the spiking neuron, which can emit spikes in response to a spiking input stimulation. Spike-timing dependent plasticity (STDP) (Bi and Poo 1998) is a widely discussed synaptic property that has been realized by memristive devices and extensively reported (Figure 1E). A major limitation of the local learning rules such as STDP is that they can only be applied to shallow networks with only one layer. Recent research works aim at pushing the network deeper by introducing backpropagation approximation in spike domain (Shrestha et al., 2019), using synthetic local gradient (Kaiser et al., 2020) and adopting backward residual connections and stochastic SoftMax functions (Panda et al., 2020).

Artificial spiking neurons can mimic closely the biological neurons' behaviors, for example, reproducing the exact waveform of the membrane potential in different phases of a fire. However, these bio-realistic models are rarely used in software-simulated SNNs owing to a heavier computational burden. More typically, spiking neurons in SNNs are designed to have a binary response, i.e., "1" when the neuron receives or generates a spike or "0" when it does not. Memristive devices with non-linearity, state-variable accumulation (integration), volatility, as well as stochasticity (Figure 1E) can be used to construct spiking neurons and dendrites, thus combining bio-realistic responses with low computational burdens via the physical properties of the memristor.

There are various methods to encode the information into the spikes, such as using the frequency of a train of spikes (rate coding) or the precise arriving time of each spike (temporal coding) (Masquelier et al., 2009). Different encoding methods running on different neural networks require different training rules, which, however, are usually more computationally expensive than ANNs. Temporal-encoded SNNs are more efficient than rate-encoded SNNs, as the information can be contained in just one spike. Owing to the time-related nature, temporal-coded SNNs are more suitable to process time-related data, such as speech, sound, and vision. Memristive devices with state-variable accumulation, or short-term volatility, (Figure 1E) can be used to process these temporal-encoded spiking patterns in reservoir computing networks (Moon et al., 2019) (Figure 1F). Other advanced methods have been proposed to employ a pseudo-gradient to overcome the non-differentiability of a spiking neuron in feedforward networks. For instance, the e-prop training method in recurrent SNN (Bellec et al., 2020) has been proposed to approximate the backpropagation through time (BPTT) algorithm for traditional RNNs (Werbos 1988), which removes biologically unrealistic computational requirements and makes it possible to build on-chip hardware learning units.

Besides the SNNs as a signal morphological approach, another approach for brain-like computation is the collective-state computation that emulates brain-like computation at a high level (Csaba and Porod 2020).

This computational model originated from the Hopfield neural network (Hopfield 1984), then extended to cellular neural networks (Chua and Yang 1988), coupled oscillators (Ignatov et al., 2017), and adiabatic annealing machines with probabilistic bits (Borders et al., 2019) (Figure 1F). In these networks, nonlinear synaptic and neural behaviors as well as probabilistic behaviors are needed, which can be provided by the memristive devices (Figure 1E). Similarities between memristive devices and biological components, such as synapses, dendrites, and neurons, have been extensively reported. Some fundamental functionalities or brain-inspired algorithms have also been demonstrated. However, practical computing systems based on these algorithms have rarely been reported.

DL Accelerators by Memristive Hardware

Memristive synapses can be optimized to accelerate DL algorithms. Assembled into a crossbar array configuration, memristive devices are inherently suitable for efficient VMMA operations, which account for a majority percentage of the computation in artificial neural networks, by directly using Ohm's law for multiplications and Kirchhoff's current law for accumulations. Both passive (Prezioso et al., 2015) and active crossbar arrays (Li et al., 2018a) have been reported for VMMA. Three-dimensional (3D) stacking of crossbars provides an additional dimension of parallelism, connectivity, and efficiency for complex neural networks (Lin et al., 2020). The DL neural network has many variants and components, for instance, a fully connected layer, convolutional layer, and recurrent neural network. The mapping between the memristive array and computing layers (Yao et al., 2020) is necessary to achieve the designed topology and is crucial for realistic applications.

DL accelerators require both linear current-voltage (I-V) relation and linear weight update characteristics of the memristive synapses. Optimization of material stacks and electrical operation protocols can ensure very stable and linear I-V characteristics for voltage-conductance multiplication (Hu et al., 2018a). The differential pair (G+/G-) method (Suri et al., 2011) has been widely used in both RRAM and PCM synapses, to allow for both positive and negative synaptic weights using electrical conductance, and more importantly to compensate for any asymmetry between the two weight update directions. Other methods are also used to improve the linearity of synaptic weight updates, for instance, multi-parallel devices for mapping the least significant and most significant weight components in a single synapse (Ambrogio et al., 2018).

So far, memristive DL accelerators have experimentally achieved reasonable accuracy in relatively small-scale tasks such as recognition of the image in the Modified National Institute of Standards and Technology (MNIST) and Canadian Institute For Advanced Research (CIFAR) datasets (Yao et al., 2020; Ambrogio et al., 2018). Such demonstrations remain to be seen in larger-scale DL networks like ResNet and other network structures for ImageNet (He et al., 2016).

Optimizing Linear Response of the Memristive Device

In the inference stage of the neural network, the synapse is essentially carrying out the weight function, that is, scaling the input signal into an output signal. This can be realized by the memristive device through implementing the Ohm's law: $I = G \cdot V$, where G is the conductance of the memristive device, V is the input signal represented by a voltage, and I is the output signal represented by the current through the memristive device. However, memristive devices do not always follow Ohm's law with a linear I-V relation. Since the electron transport in a low conductance state often involves mechanisms like tunneling or hopping transport (Ielmini and Zhang 2007), it is common to observe nonlinear (e.g., exponential) relation between current and voltage (Figure 2A), which compromises the accuracy of directly using Ohm's law for multiplication. This issue is more severe in a passive memristive array where a nonlinear I-V characteristic is deliberately adopted to mitigate the difficulty in selectively programming a target device in the passive array (Alibart et al. 2013).

The linearity of the I-V characteristic can be improved by optimizing the stacks of the memristive device thus the switching dynamics rely on the composition modulation of a localized conduction channel rather than a tunnel barrier (Jiang et al., 2016). A smaller read voltage is preferred for linear read operation since even an exponential I-V relation can be approximated by a linear behavior at small voltages, which, however, may decrease the signal/noise ratio of the circuits and potentially degrade the accuracy of the network.

An alternative way to bypass the nonlinearity of the read operation (for inference) is to use a fixed read voltage while the input signal strength is represented by the duration of the read. The output signal will

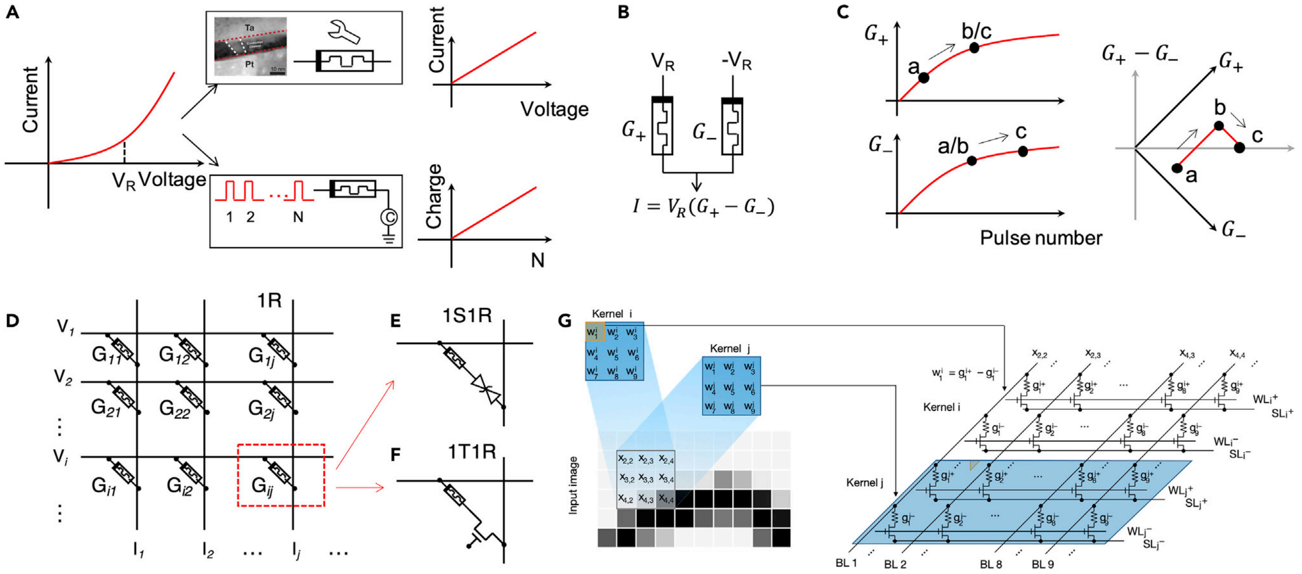


Figure 2. Material- and Device-Level Optimizations to Compensate for the Nonidealities of Memristive Synapses

(A) Dealing with non-linear current-voltage (I-V) relation in memristive devices. Left: a typical I-V relation showing exponential dependence of the reading current on applied voltages. Upper right: material modification for linear I-V relation by using composition modulation of localized conduction channels. Reproduced from (Jiang et al., 2016), CC BY. Lower right: pulse number modulation with the number of pulses representing the strength of the input signal and accumulated charge representing the weighted output, resulting in linear behavior between the charge and the pulse number.

(B) Differential pair for implementing both positive and negative synaptic weights.

(C) Learning errors as the result of the nonlinearity of the weight update. For gradual conductance increases of both G_+ and G_-, potentiation is conducted by applying pulses to the G_+ part (a → b) and depression is conducted by applying pulses to the G_- part (b → c). The same numbers of potentiation and depression should cancel each other; however, it results in different overall weight (G_+-G_-) between (a) and (c).

(D) An array of memristive devices arranged in crossbar structure for VMMA. The voltage vector was applied to the row bars of the memristive array and the current output in each column bar is sensed as the result of Ohm's law and Kirchhoff's current law, $I_j = \sum_i G_{ij} V_i$.

(E) and (F) Illustration of each memristive device in series connected with a passive selective device or a transistor, respectively, for access without affected by or affecting other devices.

(G) Mapping 2D convolutional kernels to memristive array for convolutional layers. Reproduced from (Yao et al., 2020), copyright © 2020, Springer Nature.

be the accumulated charge of the output current. The duration can be the width of a single read pulse (Cai et al., 2019) or digitalized into the number of identical pulses, namely, pulse number modulation (Yao et al., 2017; Cai et al., 2019). This technique may sacrifice the inference speed. In addition, charge accumulation of the output current requires capacitor-based circuits or additional processes in the digital domain. On the other hand, using the number of identical pulses to represent input signals may simplify or even eliminate the need for analog/digital conversion circuits.

As previously mentioned, the mainstream algorithms for artificial neural networks are based on gradient descent of the synaptic weight to minimize the difference between the actual and the target outputs. This results in both negative and positive values of the synaptic weights. The negative synaptic weight also has its biological basis of inhibitory connections in the biological neural system. A practical way to realize negative synaptic weight is to use a differential conductive pair as shown in $I=(G_+-G_-)\cdot V$, where G_+ and G_- are the conductances of the positive and negative branches of the differential pair, respectively. The minus sign can be easily implemented by Kirchhoff's current law, as shown in Figure 2B. Although this reduces the synaptic density by two times, it has additional benefits such as a better redundancy for precise representation of the synaptic weight. For example, if one of the memristive device in the pair is stuck to a certain conductance and becomes non-responsive, the pair can still be used to represent any synaptic weight by just programming the other devices in the pair to an appropriate value. Negative synaptic weights can also be realized by subtracting one shared "negative" column (reference) from all "positive" columns (Milo et al., 2019).

Compensating Device Non-ideality by Online Learning

The synaptic weight can be learned offline by simulating the artificial neural network in traditional computing systems and then transferring the resulting weights to the memristive differential pairs in memristor arrays. This type of weight learning, although consuming a large amount of computational resources on the offline learning platform, is only a one-time cost. To minimize the write errors, it is necessary to utilize iterated program-and-verify steps until the synaptic weight falls into the target range. However, owing to the limited conductance levels and variations of the devices, discrepancies between the original weights and the transferred ones are inevitable. Device failures would also largely degrade network performance (Li et al., 2018).

One solution to compensate for the inaccurate program and device failures would be on-line learning or *in situ* learning. To minimize the output error, the synaptic weight needs to be adjusted according to the gradient descent rule, which can be denoted by $\Delta w = \eta \cdot x \cdot \delta$, where Δw is the weight changes, η is the learning rate, x is the input value, and δ is the output error caused by this synaptic connection (defined as the partial derivative of the total output error to this node's output, and can be backpropagated from the last layer to all nodes in the network). The output error can be obtained by comparison between the actual output of the memristive network and target output. Then periodical weight update can be performed to the memristive synapses via repeated presentations of input patterns from the dataset. With the output error calculation performed *in situ* in the memristive network, the weight learning partially adapts to the nonidealities of memristive devices (write noise and device failure), thus mitigating the impact of the memristive nonidealities (Li et al., 2018). A similar approach can also be applied to offline training where the weight update in network simulation takes the nonidealities of the memristive devices into consideration, namely, hardware aware training (Gokmen et al. 2019; Joshi et al., 2020). Furthermore, a recently proposed hybrid training scheme takes the advantages of both *in situ* and *ex situ* training methods to efficiently realize a memristor-based neuromorphic system (Yao et al., 2020). The weights are initially learned and transferred to memristor conductance as in conventional *ex situ* training case regardless of device non-ideal characteristics, and only a small part of the weights is trained *in situ* (e.g., the weights of the last fully connected layer) to adapt to present imperfections and recover the system performance. All these realizations of online learning are clear examples of the strength of the device/algorithm co-design, where device nonidealities are mitigated via system-level algorithm optimization.

Realizing Linear Weight-Update

Neural network training is a computationally intensive task. Although it can be a one-time cost for some applications, other applications that involve continuous learning or adaption require more frequent retraining of the memristive neural networks. The iterated backpropagation steps in the online training could be a major barrier for training acceleration with memristive networks. It would be ideal if only one write pulse was needed to adjust the memristive synaptic weight in each learning epoch without the need of knowing the current state of the memristive synapse or verifying the updated state. However, this cannot be achieved in most of the memristive devices because their set (increasing the conductance) and reset (decreasing the conductance) switching processes are nonlinear and asymmetric.

A gradual (analog) conductance change upon write pulses during switching can only be achieved in either set or reset operation. RRAM devices can only be gradually reset while the set is abrupt. In contrast, PCM devices can only be gradually set while the reset process is abrupt. As a result, the devices for differential pair composed of RRAM devices are prepared in relatively high conductance states to start with. If potentiation is needed, then the negative conductance branch G_{-} will receive reset operating pulses, whereas for depression, reset pulses will be applied to the positive branch G_{+}. The opposite configuration can be used for a PCM differential pair. However, in both cases, the amount of the conductance change per write pulse is not constant, rather it depends on the current conductance state of the devices, thus leading to substantially non-linear weight update.

Saturation of conductance change in the low-conductance range of RRAM device during the reset operation and in the high-conductance range of PCM during the set operation is generally observed (Figure 2C). During training, numerous weight updates in potentiation and depression should mostly cancel each other and result in only a small amount of net conductance change. The nonlinear changing of conductance in G_{-} and G_{+}, however, prevents such cancellation in the weight updating process and acts as the largest source of accuracy loss in online training (Burr et al., 2015).

To linearize the conductance change, material- and device-level optimization has been performed. For instance, by inserting an AlO_x barrier layer, a HfO_2-based RRAM device shows an improved linear potentiation (or depression) behavior of conductance under identical pulses (Woo et al., 2016). Choi et al. limited the conductance changes in a smaller window in Ag^+-based memristive devices by modifying the potentiation pulses to only allow a mild amount of Ag participant in the switching dynamics (Choi et al., 2018). A three-terminal memristive device with an electrochemical gate layer can also make the change of source-drain conductance linear with the pulses applied in the gate terminal (Tang et al., 2018; van de Burgt et al., 2017). These electro-chemical memristive devices, however, suffer from long pulse duration hence a low operational speed. The gradual linear conductance tuning in both directions also allows the potentiation and depression in both sides of the differential pair, avoiding freeze-out of the conductance tuning in unidirectional update (Ielmini and Pedretti 2020).

Capacitor-based synapse has an ideal linear weight update ability but is volatile. A synaptic configuration composed by a major pair of non-volatile PCM devices (G+, G-) and a minor pair of volatile capacitor-based synapses (g+, g-) can thus integrate the advantages of these two parts and lead to linear weight update and non-volatile synaptic weight storage. The minor pair updates its weight frequently in each training cycle and transfers the weight to the major pair periodically. Thanks to the linear weight update ability, a training accuracy comparable with that of a software-based solution can be obtained (Ambrogio et al., 2018). However, this solution results in a bulky individual synapse (2PCM + 3T1C), which occupies a relatively large area on the silicon chip.

Utilizing Memristive Arrays for VMMA

Mapping the fully connected layers of artificial neural networks to the two-dimensional (2D) memristive arrays is straightforward since the computation in the fully connected layers are essentially VMMA operations (Ielmini and Pedretti 2020), $y_j = \sum_i w_{ij} x_i$, where i and j denote the indices of the input vector and output vector, respectively. It can be directly implemented by applying voltages to the rows of a memristive array and sensing the current output at the end of each column as the result of Ohm's law and Kirchhoff's current law, as shown in Figure 2D.

Note that the case shown in Figure 2D illustrates the ideal case of using a passive memristive array (Prezioso et al., 2015). However, there is a dilemma in a passive array between the desire for a linear I-V relation to directly utilize Ohm's law during inference and the desire for a nonlinear I-V relation to suppress the sneak path currents and program individual devices during learning. Adopting a selector in each cell to form the so-called one-selector/one-resistor (1S1R) structure (Figure 2E) increases the nonlinearity for array programming, which, however, makes the pulse duration/number modulation scheme discussed before an only suitable choice for the inference.

A more practical implementation of VMMA is to use an active array where each cell of the array is composed of a memristive device and a transistor connected in series (Figure 2F). With such transistors in the array, an individual memristor can be selected by activating only the corresponding transistor(s) for reading or programming operations without affecting other devices in the array. With a high on/off ratio of the mature transistor technology, a large memristive crossbar array can be achieved, with the only limitation essentially being the wire resistance. Fully parallel reading during inference can be enabled by turning on all the transistors in the array, and semi-parallel programming (e.g., row by row) is also possible. The cell size is limited by the transistor footprint, which results in a lower synaptic density than the passive crossbar array. The overall area efficiency has been limited by the peripheral circuits for neuronal functions (section "Co-design with CMOS Peripheral Circuits") in most hardware implementations of memristive deep networks so far. When innovative designs are devised to significantly reduce the area of such peripheral circuits, the 1T1R synaptic arrays limited by the size of the transistor will then become the primary concern of the area efficiency.

The parasitic effects existing in the array also limit the size of the array. For instance, when the wire resistance is comparable with the device resistance, the voltage drops on the wire connections lower the real voltage applied to the memristive device and might result in program or read errors, i.e., IR drop issues. The actual limit of the size of an array is affected by the wire resistance, the on resistances of the devices, and the on/off ratios. Assuming the memristive device has resistances of 100 kΩ and 10 MΩ in the low-resistance state and the high-resistance state, respectively, and taking the Cu wire interconnection resistance parameter from the ITRS roadmap, Zuloaga et al. predicted that array size up to 256 × 256 can work well after being downscaled to 10 nm technology node (Zuloaga et al., 2015), which is sufficient for most of the demonstrations of deep networks reported so far. If a larger array size for VMMA operation is needed, material-level optimizations to decrease the resistance of the connection wires or to increase the resistance of the memristive devices are needed (Zhang et al., 2019). Alternatively, system reduction schemes that can substantially reduce the weight matrix size can be utilized (Liu et al., 2014). It is also possible to mitigate the IR drop by a compensation scheme retraining the dominant neurons (Jeong et al., 2018). Online adaptive learning (Li et al., 2018), where the automatic weight update will adapt to the mismatch of the actual cell resistance/conductance with the desired synaptic weight, can also mitigate the parasitic effects in array configurations.

Convolutional neural networks are more suitable for recognizing static 2D inputs, such as images, thanks to the biologically inspired convolutional kernels similar to the receptive field of complex cells in the visual cortex. Mapping the convolutional layers to memristive arrays needs first converting the convolutional operation of matrix-matrix multiplication of each kernel and its receptive field to a vector-vector multiplication (Gao et al., 2016). The kernel vectors are aligned in columns or rows of the memristive array, and image input is partitioned in receptive fields (patches) and sequentially fed to the kernel vectors, as illustrated in Figure 2G (Gokmen et al., 2017; Yao et al., 2020). This weight sharing methods by the convolutional kernels is also named as spatial weight sharing (Wang et al., 2020c). Spatially shared memristive weight strongly reduces the number of memristive devices needed for a deep CNN. For comparable accuracy of recognizing handwritten digits from MNIST database, the three-layer fully connected neural network consumes 329,770 PCM devices (accuracy 97.94% [Ambrogio et al., 2018]), whereas the five-layer convolutional neural network uses 5,629 RRAM devices (accuracy 96.19% [Yao et al., 2020]).

Spatially shared kernel vector weight, however, requires repeated vector-vector multiplication between input image patches and the convolutional kernel. The number of patches increases quadratically with the image size (Gokmen et al., 2017), which reduces the throughput and becomes the bottleneck of the system performance. It is also possible to parallelly implement the convolutional operation in fully connected topology (Lin et al., 2020) or partially reduce the spatial share factor. Replicating and transferring identical convolutional kernels to multiple 2D memristor arrays provides a possible solution to boost the parallelism of convolution operations and enhance the throughput accordingly (Yao et al., 2020). However, a careful trade-off between the inference speed and memristive area cost is necessary.

Memristive devices are suitable for 3D integration, a powerful solution for high-density storage, and most importantly, for enhancing the neuronal connectivity required in complex neural networks. The 3D integration enables parallel and faster convolutional calculation for neuromorphic application (Huo et al., 2020; Lin et al., 2020), bypassing the trade-off between the inference speed and memristive area. As the cross-section of the 3D array is a 2D interface to directly accept 2D image inputs, the image patches and convolutional kernel matrices no longer need to be unrolled to vectors (Lin et al., 2020). Moreover, the additional dimension could enable massive connections and increase the flexibility of memristor topologies. By defining the sliding kernels at each patch as zigzag staircases in 3D space and shaping both the top electrode and bottom electrode as vertical pillars, input signal could be fed into different patches through the same pillar simultaneously. In this manner, the whole image could be presented at the same time and the 3D device structure enables all convolutional operations during the sliding procedure to be processed in parallel, saving the sequential shifting time and improving the system performance. Additionally, unrolling 3D convolutional kernels to 2D matrices, a 3D convolutional neural network for stereoscopic object recognition can be realized (Huo et al., 2020).

Other variants of the DL neural network, like long short-term memory (LSTM) (Li et al., 2019; Wang et al., 2019c) and deep belief neural network composed by RBMs (Eryilmaz et al., 2016), can also exploit the benefits of VMMA capability of memristive arrays, which is very attractive as VMMA consists the major part of the computations in these networks. However, other essential functionalities, like the gate unit controlling the memory time for LSTM cells and probability generations in RBM neurons, are carried out in software. These functionalities can also be implemented by exploiting the unique features of memristive devices more morphologically, which will be covered in the next section.

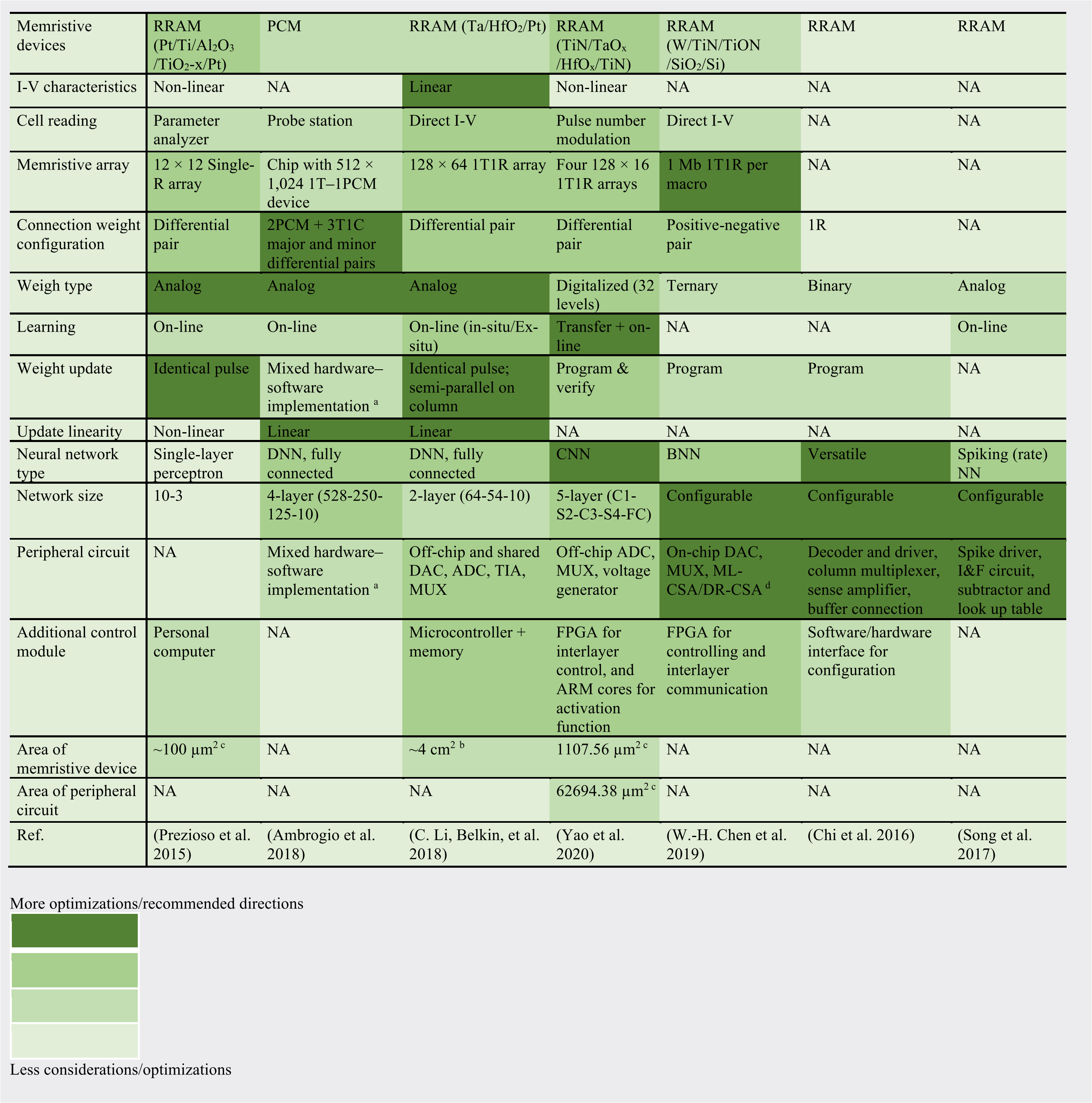
Co-design with CMOS Peripheral Circuits

The program and read operations on the memristive array and the weight update calculation need to be carried out by complementary metal–oxide–semiconductor (CMOS) circuits mimicking the behavior of biological neurons, which should be integrated closely with the memristor array on the same chip to further enhance the efficiency. Tailored according to the targeted application and specific hardware architecture, the generally utilized circuitry blocks could include the sample-and-hold (S&H) module and analog-to-digital converter (ADC) to temporarily hold the summed analog currents and transform them to the digital domain, respectively. The digital-to-analog converter (DAC) that converts digital inputs into appropriate voltage amplitudes should be counted if a voltage amplitude-encoding scheme is adopted. Digital control, processing, and routing blocks are also necessary to realize activation functions and monolithically integrate a complete neuromorphic system (Shafiee et al., 2016; Hu et al., 2018). Furthermore, extra peripheral circuits need to be considered to realize various kinds of on-chip learning rules. It is worth mentioning that in practical system implementation, the bottom-level device characteristics and the top-level algorithm optimizations would jointly determine the circuit and architecture design to meet the necessary hardware performance. The array size, the precision and speed of ADCs, and other circuit aspects need to be carefully considered with trade-offs between hardware efficiency and cost.

CMOS implementation of the peripheral circuits can result in a large area and power consumption. For instance, the DAC and ADC circuits can occupy a much larger area (e.g., 21.57 mm^2) than the dense synaptic array (e.g., 0.14 mm^2) (Cai et al., 2019). A commonly used strategy to improve the system area efficiency is to temporally share the DAC and ADC elements. State-of-the-art high-precision ADCs consume large area; however, they can have high sampling rates. Thus multiple output nodes of the memristive array can share a single ADC sequentially (Gokmen and Vlasov 2016). An additional multiplexer (MUX) circuit is needed for the sequential selection of output nodes. ADCs of 6–8 bit precision are needed for acceptable accuracy loss of neural networks with a relatively small size while further lowering the precision may induce high accuracy deterioration (Li et al., 2016). Nevertheless, the required precision for large-scale neural networks depends on the exact network structure or dataset. Further reducing the precision can be achieved by separately handling the outlier values and normal values with 4-bit ADCs, showing a relatively small loss of performance (Park et al. 2018). A binary neural network is a possible option to address the area and power inefficiency in ADC- and DAC-based neuronal function realizations, which will be further discussed in section ''Exploiting Bistable Behavior of Memristive Devices.'' It is also possible to use neurons working directly in the analog domain (Krestinskaya et al., 2018), thus eliminating the need for the conversion between analog and digital signals. This requires the analog neuronal circuit to perform (in the analog domain) activation functions, like sigmoid or ReLU, which, in most of the hardware deep neural network demonstrations so far, have been implemented by controlling computers or microcontrollers in the digital domain after the conversion of the signal from the analog domain. Using integrate&fire neurons to convert the output of VMMA to spiking trains, where the spike count or frequency denotes the analog value of the neuron output, can also eliminate the need of analog to digital conversion (Yan et al., 2019).

For a large-scale integration of the synaptic and neuronal components with the learning algorithms in a practical system, utilizing the above strategies, in-memory computing macros or neuromorphic computing macros have been proposed, such as ISAAC (Shafiee et al., 2016), PRIME (Chi et al., 2016), and Pipelayer (Song et al., 2017). These macro circuits can be tiled together according to the structure of deep neural networks to be constructed. For a more comprehensive review of peripherical circuits and large-scale integration, readers can refer to Yan et al., 2019. These memristive DL accelerators are projected to be superior to CMOS-based or other solutions in several aspects, such as in performance (operation per second, OPS), area, and power efficiency (Zhang et al., 2020a; Sebastian et al., 2020). Unfortunately, ideal switching characteristics and linear I-V characteristic are assumed in these designs. In terms of the chip-level demonstration, a fully integrated memristive in-memory computing macro, named as non-volatile computing-in-memory (nvCIM), has been demonstrated (Chen et al., 2019). However, the nvCIM macro works on the binary-input ternary-weight model and will not fully exploit the analog in-memory computing ability of a memristive array. The realization of activation and intra-layer communication is carried out by off-chip field-programmed gate array (FPGA). Recently, this field has been rapidly developing toward monolithically integrated memristive neuromorphic systems, even though the memristive analog behavior has not been fully exploited (Liu et al., 2020; Wan et al., 2020).

In Table 1, we summarize the optimization and design efforts in various levels of memristive neuromorphic computing with the purpose of DL accelerators. The color code refers to the degree of optimization/co-design of each implementation. Despite the numerous efforts in Table 1, a general-purpose memristor accelerator for general neural networks is still missing, partially because device reliability and uniformity



Memristive devices	RRAM ($Pt/Ti/Al_2O_3$ /TiO_{2-x}/Pt)	PCM	RRAM (Ta/HfO_2/Pt)	RRAM (TiN/TaO_x /HfO_x/TiN)	RRAM (W/TiN/TiON /SiO_2/Si)	RRAM	RRAM
I-V characteristics	Non-linear	NA	Linear	Non-linear	NA	NA	NA
Cell reading	Parameter analyzer	Probe station	Direct I-V	Pulse number modulation	Direct I-V	NA	NA
Memristive array	12 × 12 Single-R array	Chip with 512 × 1,024 1T–1PCM device	128 × 64 1T1R array	Four 128 × 16 1T1R arrays	1 Mb 1T1R per macro	NA	NA
Connection weight configuration	Differential pair	2PCM + 3T1C major and minor differential pairs	Differential pair	Differential pair	Positive-negative pair	1R	NA
Weigh type	Analog	Analog	Analog	Digitalized (32 levels)	Ternary	Binary	Analog
Learning	On-line	On-line	On-line (in-situ/Ex-situ)	Transfer + on-line	NA	NA	On-line
Weight update	Identical pulse	Mixed hardware–software implementation [a]	Identical pulse; semi-parallel on column	Program & verify	Program	Program	NA
Update linearity	Non-linear	Linear	Linear	NA	NA	NA	NA
Neural network type	Single-layer perceptron	DNN, fully connected	DNN, fully connected	CNN	BNN	Versatile	Spiking (rate) NN
Network size	10-3	4-layer (528-250-125-10)	2-layer (64-54-10)	5-layer (C1-S2-C3-S4-FC)	Configurable	Configurable	Configurable
Peripheral circuit	NA	Mixed hardware–software implementation [a]	Off-chip and shared DAC, ADC, TIA, MUX	Off-chip ADC, MUX, voltage generator	On-chip DAC, MUX, ML-CSA/DR-CSA [d]	Decoder and driver, column multiplexer, sense amplifier, buffer connection	Spike driver, I&F circuit, subtractor and look up table
Additional control module	Personal computer	NA	Microcontroller + memory	FPGA for interlayer control, and ARM cores for activation function	FPGA for controlling and interlayer communication	Software/hardware interface for configuration	NA
Area of memristive device	~100 μm^2 [c]	NA	~4 cm^2 [b]	1107.56 μm^2 [c]	NA	NA	NA
Area of peripheral circuit	NA	NA	NA	62694.38 μm^2 [c]	NA	NA	NA
Ref.	(Prezioso et al. 2015)	(Ambrogio et al. 2018)	(C. Li, Belkin, et al. 2018)	(Yao et al. 2020)	(W.-H. Chen et al. 2019)	(Chi et al. 2016)	(Song et al. 2017)

More optimizations/recommended directions

🟩
🟩
🟩
🟩

Less considerations/optimizations

Table 1. A Survey of Optimization and Design Considerations in Various Levels of Memristive Neuromorphic Computing

NA: not applicable (or not discussed).

[a] Training within a transfer interval was performed in software with device models and read PCM devices are operated when transfer needed.

[b] Data retrieved from (Li et al., 2018).

[c] Projected for all peripheral circuits integrated on chips with 128 × 128 array and four columns of the memristive array sharing one ADC converter.

[d] ML-CSA: multi-level current-mode sense amplifier; DR-CSA: distance racing current-mode sense amplifier.

issues across multiple arrays are yet to be solved. On the other hand, conventional DAC and ADC solutions consume a large area and energy, becoming the bottleneck of system performance (Cai et al., 2019). Novel routing schemes with the least requirement for the on-chip memory are also needed to make the most of the memristor neuromorphic system.

Exploiting Memristive Properties for Brain-Inspired Algorithms

In addition to serving as a static memory of synaptic states for in-memory-computing in deep learning algorithms, memristive devices also have a variety of dynamical properties that share close similarities with biological components, which can potentially lead to computing with augmented efficiency and intelligence. Novel brain-inspired learning algorithms are needed to utilize these intrinsic properties of memristive devices, e.g., the stochasticity of the state (Yu et al., 2013), the dynamics of state transition, and second-order effects (Du et al., 2017).

Exploiting the Stochasticity

Various random physical phenomena exist in memristive devices, resulting in stochastic variations of conductance levels and switching parameters. For instance, for RRAM and PCM devices, owing to the nature of the ionic-electronic coupled transport mechanism, intrinsic stochasticity exists as random telegraph noise in the reading phase and variation of switching parameters in the weight update phase (Carboni and Ielmini 2019). Stochasticity is a critical problem for memory and storage applications and their usage as synaptic weights in DL accelerators. However, stochasticity as a physical entropy source can be exploited for generating true random numbers or physical unclonable functions for information security. Additionally, they can provide a low-cost solution for implementing some specific neural network algorithms where stochasticity is essential for computation.

Under a weak programming condition, the set transition of metal-oxide memristive devices becomes probabilistic. A winner-take-all network can be realized by the competition among post-neurons utilizing the probabilistic switching in synapses (Yu et al., 2013), as illustrated in Figure 3A. In RBM, the sampling and reconstruction stages heavily rely on the probabilistic of hidden or visible units being activated. A dot product circuit incorporating the stochasticity coming from the intrinsic noise of the memristor array for the RBM has also been demonstrated (Mahmoodi et al., 2019). Using a passive memristor crossbar, a single-layer RBM with ten visible and eight hidden neurons is demonstrated with the energy function minimization (Figure 3B). In another work of implementing the Hopfield neural network, the intrinsic noise of a memristive crossbar was used for a combinatorial optimization problem (Cai et al., 2020). A moderate noise level was found useful for the network to escape from local minimum points in the energy landscape better than both the noise-free and the high noise level situations (Figure 3C).

The stochasticity of magnetic tunnel junctions (or MRAM) was reported to implement three-terminal probabilistic bits (p-bits) (Borders et al., 2019). These p-bits can be viewed as probabilistic neurons that are electrically connected to form an asynchronous network for factorizing integers up to 945 (63 × 15) adapting diabatic quantum computing algorithm. Tunable probability of the random switching of the superparamagnetic tunnel junction also allows the population coding where each neuron embodied by a superparamagnetic tunnel junction is associated with a specific range of inputs, which is then computed as a weighted sum of the rates of each neuron (Mizrahi et al., 2018). Similar tunable probability can be achieved by utilizing the inherent random noises of analog RRAM devices and was used to construct a Bayesian inference neural network that shows high resilience to adversarial testing samples (Lin et al., 2019).

Exploiting the Current-Voltage Non-linearity

I-V non-linearity of memristive devices could be a major issue when the devices are used as synapses for DL accelerators. However, such non-linearity is an essential synaptic or neuronal behavior in cellular neural networks (Duan et al., 2015; Caravelli et al. 2017), where it enriches the dynamics of the system. It has also been shown that the non-linearity can be used to mimic the non-linear integration of biological dendrites (Lavzin et al., 2012), constructing a memristive dendrite (Li et al., 2020).

In conventional algorithms of DL neural networks, neurons act as simple elements summing all inputs from synapses. It is found in neuroscience that the non-linear integration of synaptic signals by the dendrites provides primitive processing before the signals reach the neuron body (Agmon-Snir et al., 1998). Recently, a memristive dendrite component has been demonstrated using a Pt/TaO_x/AlO_δ/Al-based dynamic memristor (Li et al., 2020), exploiting the non-linearity provided by the Schottky-like barrier in the Pt/TaO_x interface. By adding non-linearity before the neural summation to realize the non-linear dendritic function can further enhance the performance of the neural network. With memristive dendrites filtering the background signals, the spiking output of the neuron shows more distinction between false patterns and

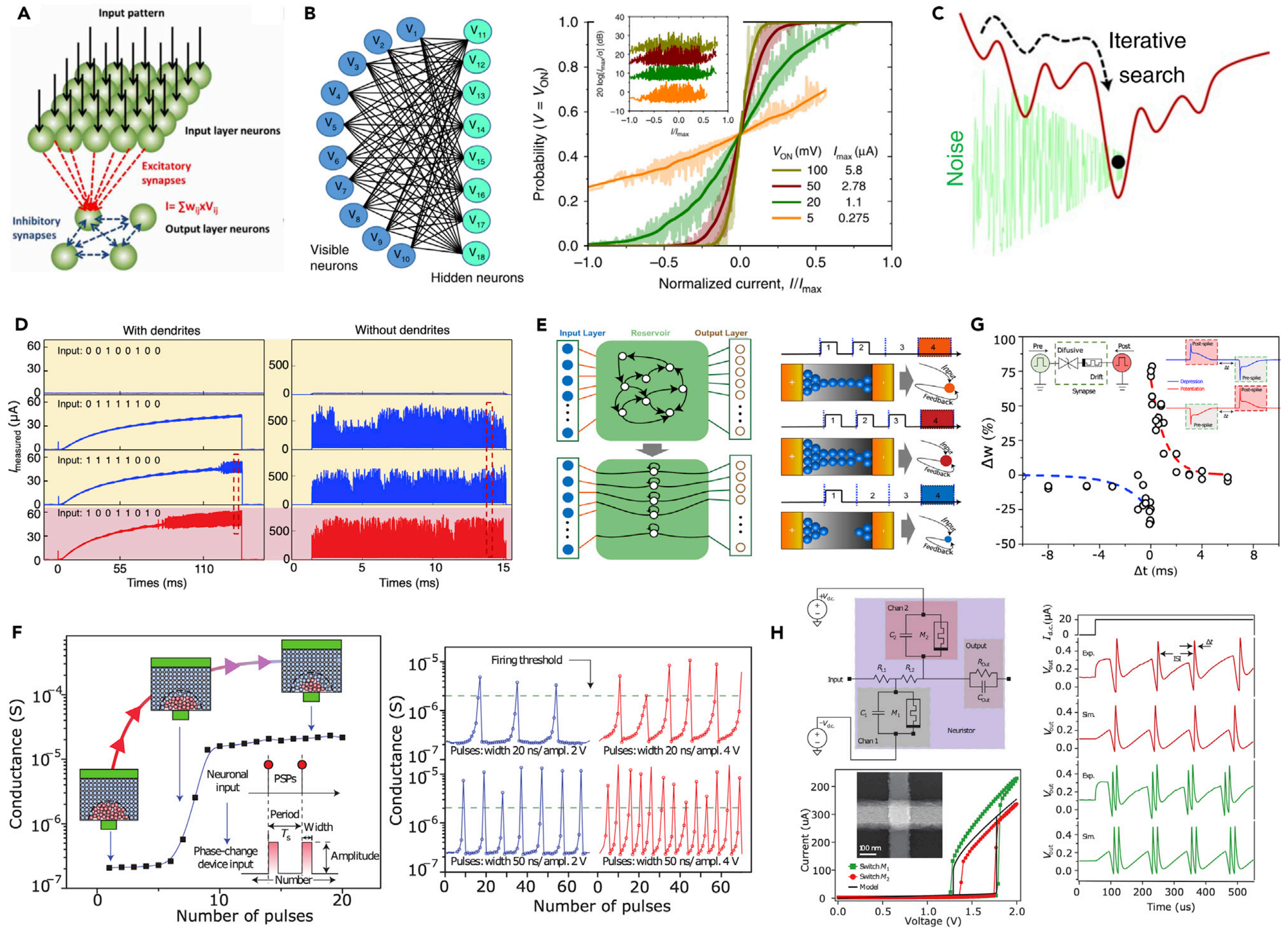


Figure 3. Various Bioinspired Algorithms Exploiting the Unique Features of Memristive Devices

(A) A winner-take-all neural network exploiting the stochasticity of the memristive devices. Reproduced from (Yu et al., 2013), CC BY.

(B) A restricted Boltzmann machine exploiting the intrinsic reading noise of the memristive array. Reproduced from (Mahmoodi et al., 2019), CC BY.

(C) Exploiting the intrinsic noise of memristive array to avoid trapping in local minimum sites in a Hopfield memristive neural network. Reproduced from (Cai et al., 2020), copyright © 2020, Springer Nature.

(D) Memristive dendrites exploiting the non-linearity of memristive devices for filtering and integration functions. Reproduced from (Li et al., 2020), copyright © 2020, Springer Nature.

(E) Reservoir computing realized by the short-memory effect of diffusive memristors. Reproduced from (Midya et al., 2019), CC BY.

(F) A PCM device-based neuron with the gradual set switching of the PCM device was used to mimic the integration function of neurons. Reproduced from (Tuma et al., 2016), copyright © 2016, Springer Nature.

(G) Spike-timing-dependent plasticity (STDP) enabled by the short-term memory (volatility) of the diffusive memristive device. Reproduced from (Wang et al., 2017), copyright © 2016, Springer Nature.

(H) A Mott transition memristor device-based Hodgkin-Huxley neuron faithfully reproducing biological spike shape and tunable spiking trains. Reproduced from (Pickett et al., 2013), copyright © 2013, Springer Nature.

true pattern (Figure 3D), and, at the same time, reduces the power consumption. Thus, performance enhancement in both energy efficiency and accuracy is obtained with the addition of non-linear memristive dendrites.

Exploiting the State-Variable Accumulation

Upon stimulation of weak electrical pulses, the memristive devices sometimes do not show explicit conductance changes. However, explicit conductance may be induced by subsequent pulses only if there are prior pulses. This behavior happens because of the internal state-variable, for instance, the temperature being an internal memory for historical stimuli (Kim et al., 2015). The accumulation of the state-variable may offer an internal timing mechanism and enables an activity-history-dependent modulation of the first-order state, namely, conductance; thus, is also called second-order effect. It has been used to mimic the effect of the Ca^{2+} dynamics of biological synapses and enables the temporal learning of timing-encoded information (Zidan et al., 2017).

In Ag-based diffusive memristors, the configuration of the transporting ions in the dielectric layer before the threshold switching shares a similar behavior. The accumulation of ions before their final formation of a continuous filament bridging the electrodes can act as internal memory for historical stimuli. The drift of Ag ions under electrical stimuli and the diffusion of Ag ions under zero electrical bias faithfully emulate the ion dynamics that play a critical role in the neuromorphic functions of biological intelligent systems (Wang et al., 2017). For instance, pair-pulsed facilitation and depression were found in the diffusive memristor-based synapse for a high frequency of pulses and low frequency of pulses, respectively (Wang et al., 2017), which also enables the temporal learning naturally.

Reservoir computing (RC) can offer efficient temporal processing of recurrent neural networks with a low training cost. RCs based on the second-order effect (Du et al., 2017; Moon et al., 2019) and the accumulation of ions in diffusive memristors (Midya et al., 2019) have been explored. The RC exploiting the accumulation or integration of the state-variable (secondary internal variable or ion configurations) acts as a framework extracting features from temporal inputs. Taking advantage of the rich short-term dynamics of the diffusive memristive device, an RC system is constructed with one reservoir layer of a diffusive memristors and one readout layer of a nonvolatile memristor-based trainable perceptron neural network, with which classification of temporally rearranged handwritten digits from the MNIST database is achieved with a much-reduced training workload, as shown in Figure 3E (Midya et al., 2019). Without the short-term dynamics in the state-variable accumulation, neural networks for processing temporal information should have additional memory gate to control the learning and forgetting of the historical information, resulting in extra costs in circuitry and energy.

The state-variable accumulation upon electrical stimuli can be used for the integration function of an artificial memristive neuron (Tuma et al., 2016). The gradual set of PCM device has been utilized to demonstrate an artificial neuron capable of integrating post-synaptic potential at the nanoscale, where the phase configuration (thus the conductance) of the nanoscale PCM device represents the membrane potential, as shown in Figure 3F (Tuma et al., 2016). With the gradual internal state change upon pulses mimicking the integrating function and its consequential abrupt switching representing the fire behavior, a silicon oxide RRAM cell is reported to emulate a biological neuron (Mehonic and Kenyon 2016). A similar function can also be realized by the accumulation of ion transport in a diffusive memristor (Hao et al., 2020). This approach results in a capacitor-free version of a solid-state neuron; however, it requires a reset of the memristive device back to its original state after each fire.

Exploiting the Volatile Memristive Switching

Industrial memory storage application of memristive devices requires that the device can retain its conductance state for at least 10 years; thus, it is also called non-volatile memory. In neuromorphic computing for DL accelerators, a similar requirement should be fulfilled, that is, the memristive conductance for synaptic weight should remain stable for a long time to preserve the learned knowledge. However, some memristive devices with Ag as one of its electrode shows short retention time for the high conductance state (Bricalli et al., 2018; Wang et al., 2019b). The retention time is usually reported to be in the range of sub-microseconds to tens of milliseconds. This can be viewed as short-term synaptic plasticity and is reported to demonstrate some time-related computing functionalities (Wang et al., 2018a). The volatilities can be modulated by the strength of the stimuli. An increase in the frequency of applied pulses (Ohno et al., 2011) or using a higher compliance current (Wang et al., 2019b) can cause a transition from volatile to non-volatile memory, corresponding to the short-term plasticity and long-term plasticity, respectively.

The short-term plasticity of volatile memristive device allows the STDP learning with non-overlapping spikes to be demonstrated in a combined synapse of one volatile memristive device and one non-volatile memristive device (Wang et al., 2017), with the finite delay time of the volatile memristive device bridging the time gap of the non-overlapping spikes (Figure 3G).

Another approach for memristive neurons is to utilize the abrupt and volatile switching of memristive devices for the fire functionality (Wang et al., 2018d; Zhang et al., 2018), whereas the integration function is completed by charge accumulation in an external or parasitic parallel capacitor or internal state accumulation before the abrupt switching (Zhang et al., 2018b). The volatility of the memristive device enables the artificial neuron to recover its resting state spontaneously after the abrupt switching on, which is obtained by a device reset operation after each firing event in the nonvolatile memristive neurons discussed in section "Exploiting the State-Variable Accumulation." Owing to the simple structure and nanoscale-level scalability, these memristive neurons can be much more compact than the bulky CMOS neurons. Moreover, in case long time constants, such as tens of milliseconds, are needed to match the normal time constants of the biological systems, huge capacitors would be required in the CMOS neurons (Qiao et al., 2015). In contrast, a nanoscale diffusive memristor would readily provide such time constant. The implementation of memristive neurons has also enabled fully memristive neuromorphic computing (Wang et al., 2018), further enhancing the integration level of the hardware neuromorphic computing.

Volatile memristive switching sometimes accompanies negative differential resistance arising from an insulating-to-conducting phase transition or Mott transition, namely, Mott memory device (del Valle et al., 2019; Zhang et al., 2020b). Using two Mott memristors with transient memory as ionic channel and two capacitors as charge storage, a neural circuit named as neuristor was built as a hardware Hodgkin-Huxley model (Hodgkin and Huxley 1952) that faithfully mimicked the action potential generation in biological axons, as shown in Figure 3H (Pickett et al. 2013). In another work, more biologically plausible and intrinsically stochastic neurons were built with vanadium dioxide Mott memristors, which exhibited twenty-three types of biological neuronal behaviors (Yi et al., 2018). The controllable frequency of spikes in these artificial neurons also finds applications in coupled oscillator networks (Csaba and Porod 2020).

Exploiting Bistable Behavior of Memristive Devices

Without fine material-level and device-level optimization, the memristive device usually shows limited conductance levels other than the capability of analog conductance tuning. With limited conductance states, the conventional artificial neural network needs to be adapted. This can be done by quantizing the analog weight value from offline learning (Milo et al., 2019), which generally results in some loss of recognition accuracy. Many memristive devices only show binary stable states, i.e., high conductance state and low conductance state. For memristive devices embodied as STT-RAM and FeRAM, analog switching is generally more challenging.

To exploit the bistable behavior of memristive devices for synaptic applications, a binary neural network was proposed relying on binary synapses (only with two states) and binary node value (Hirtzlin et al., 2020). In the binary neural network, since the weights and inputs from the preceding layer are both binary valued, the weighted outputs are also binary, thus the vector-matrix multiplication becomes an XOR operation (Luo et al., 2019). The accumulation/summation function in the neural nodes degenerates to POPCOUNT operations, i.e., counting the number of "1"s in a series of bits, eliminating the needs of a high-precision current sensor. The activation function afterward is only a sign function, further reducing the computational needs in the neuron nodes. The binary neural network also shows high tolerance to weight bit error (Hirtzlin et al., 2020).

Ternary content-addressable memory (TCAM) is another algorithm that intrinsically exploits the bistable behavior of memristive devices (Yang et al., 2019; Ni et al., 2019). TCAM can perform in-memory search and pattern matching between the query feature vector and stored vectors of binary bits. In the study by Yan et al., 2019a, Yan et al., 2019b, 2-transistor/2-RRAM TCAM cells were used to store the TCAM vectors. For each TCAM cell, the stored TCAM datum was defined as the bit "1" for RRAM1 in HRS and RRAM2 in LRS, the bit "0" for RRAM1 in LRS and RRAM2 in HRS, the bit "X" (do not care bit) for both RRAMs in HRS. Thus, only two states of the memristive device were required. Within a similar scenario, ferroelectric TCAM with each cell only consisting of two ferroelectric field-effect transistors (FeFETs, three-terminal forms of FeRAM) has also been proposed (Ni et al., 2019). Recently, an analog memristive TCAM was introduced by taking advantage of the analog programming in RRAM devices (Li et al., 2020a).

Toward a Brain-like Computing System

The first and second panels of Figure 4 summarize the projections of various memristive features and the corresponding brain-inspired functions discussed in previous section. Building upon these components, the next step would be the construction of brain-like algorithms and realization of cognitive computations as alternative solutions to the DL techniques. Mainly, two approaches can be seen in recent developments

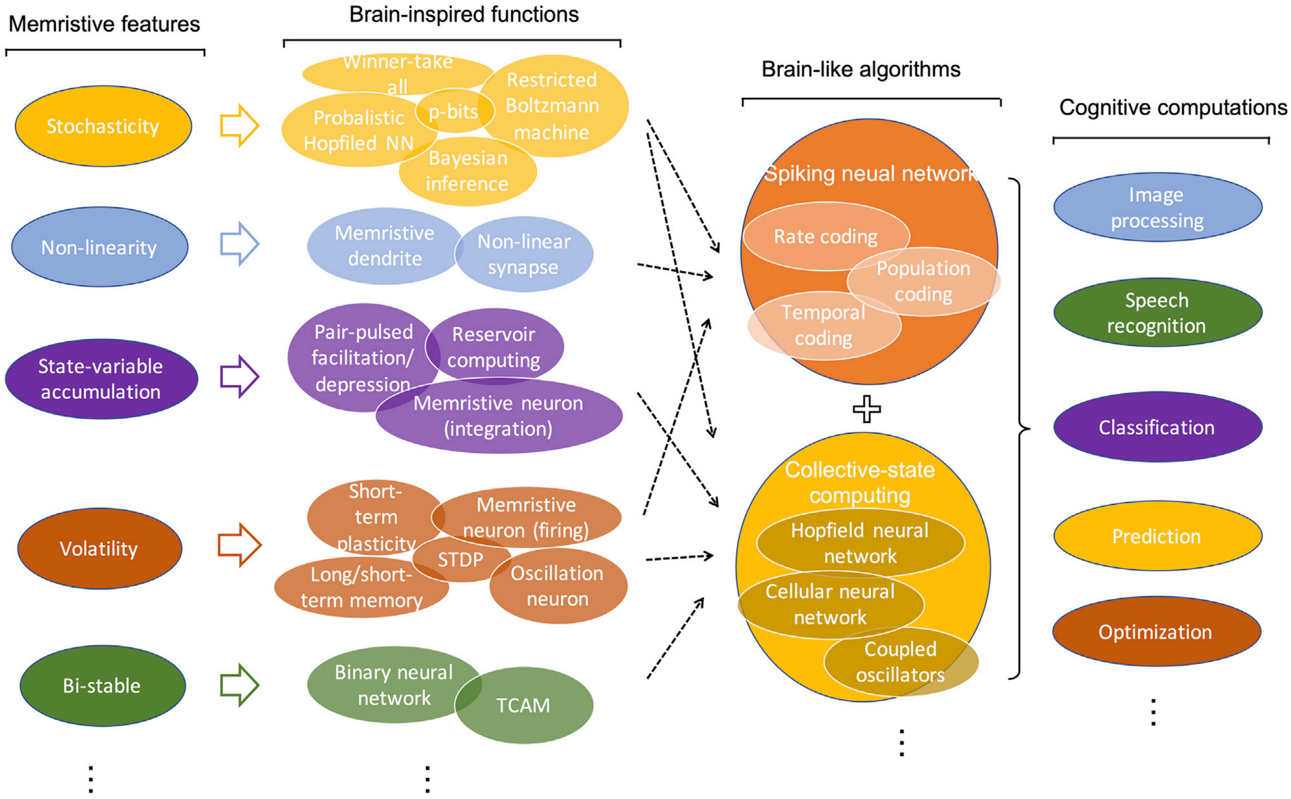


Figure 4. Exploiting Memristive Features in Various Brain-Inspired Algorithms and Projections of Using These Brain-Inspired Functions for Cognitive Computations with Brain-Like Algorithms

The unique features of memristive devices have been proposed to realize various brain-inspired functions for a neural network (from the first column to the second column); however, how to combine these brain-inspired functionalities to realize brain-like algorithms (from the second to the third) and for practical cognitive functions (from the third to the fourth) has no clear paths.

of memristive neuromorphic systems. One is from the signal morphological aspect, to emulate the spiking behavior of the biological neural network (Roy et al. 2019). The other one is from the connection morphological aspect, to emulate the collective state dynamics and evolution of the biological network. The SNN (Wang et al., 2018b), closely mimicking the information presentation in biological neural systems, is considered as a viable way to achieve brain-like computing with high energy efficiency and error tolerance. Collective-state computation has several forms, such as Hopfield neural network (Hopfield 1982), cellular neural network (Chua and Yang 1988), and coupled oscillators (Csaba and Porod 2020), mimicking brain activity at a high level, solving the problem by the system automatically finding its stable states in its energy landscape. A hybrid solution of these two approaches is also possible. However, there are no clear projections of which and how many brain-inspired functions can be utilized in these two approaches as illustrated in Figure 4.

Spike-Timing Dependent Plasticity and Spiking Neural Network

SNN is considered as the third generation of neural networks (Maass 1997), following the first generation based on McCulloch-Pitts neurons (McCulloch and Pitts 1943) with digital input and output, and the second generation composed of multiple perceptron layers with gradient descent learning algorithm, which applies activation functions after the weighted sum of the inputs and achieves analog-valued input and output (Lecun et al., 1998).

The various DL neural networks with analog-valued input and output in section ''DL Accelerators by Memristive Hardware'' can be in general converted to an SNN, with the spiking rate of each neuron proportional to the analog value. Instead of ADC/DAC conversion introduced in ''Co-design with CMOS Peripheral

Circuits'', the analog current/voltage in the memristor array can be converted by integrate and fire (IF) neurons (Milo et al., 2016; Yang et al., 2019). However, this does not fully exploit the benefits and capabilities of an SNN. In the human brain, by encoding information using spike timing, an extremely sparse and energy-efficient representation can be achieved (VanRullen et al. 2005). Conversion of the analog current/voltage in the memristor array into spatial-temporal spike representation in the digital domain using leaky integrate and fire (LIF) neurons with temporal dynamics (Fang et al., 2019) provides the possibility of constructing spatiotemporal spiking neural network.

Instead of error backpropagation, one commonly utilized mechanism for learning in SNNs is the STDP of synapses (Bi and Poo, 1998). The STDP learning rule has its biological root originated from the Hebbian learning rule, where ''neurons that fire together, wire together'' (Hebb 1949). Memristive synapses capable of STDP and triplet-based learning have been widely reported (Wang et al., 2015, 2020b). The weight updates depend on the timing of the presynaptic and postsynaptic spikes: the synapse weight is potentiated if the presynaptic spike precedes the postsynaptic spike, and depressed otherwise. The general realization of this STDP property in memristive devices is based on the engineered shapes of the presynaptic spike signal and the postsynaptic spike signal and their overlap in time (Linares-Barranco and Serrano-Gotarredona 2009; Stoliar et al., 2019). It can also be realized without spike overlapping by utilizing the internal dynamics of volatile diffusive memristors, which faithfully emulate what happens in biological synapses (Wang et al., 2017) (see also section ''Exploiting the Volatile Memristive Switching'').

Several neuromorphic systems based on the STDP weight update mechanism have been reported for pattern recognition. A PCM-based one-layer neural network for online pattern learning and recognition has been demonstrated by assuming the alternation of pattern and noise spikes from the pre-neurons and competition between post-neurons (Ambrogio et al., 2016). The essential idea is that the simultaneous pattern spikes in the pre-neurons result in a spike in one of the post-neurons, and potentiation will be induced in their connecting synaptic devices via the STDP rule, while noise spikes following the spike of the post-neuron result in depression of the according synaptic devices. The same methodology can be applied to the neuromorphic system based on RRAM memristive synaptic devices (Pedretti et al., 2017). Based on a similar methodology, the detection of the coincidence of simultaneous spikes representing an image among noise was developed (Sebastian et al., 2017; Prezioso et al., 2018) (Figure 5A).

Spatiotemporal spiking patterns can also be learned in a memristive neuromorphic system via a modified STDP learning rule, where the potentiation and depression of memristive synapse can be related to the precise timing of its received spike (Wang et al., 2018b) (Figure 5B). This can potentially enable direct learning and recognition of spatiotemporal signals in the real world, such as speech, motion, and gesture recognition (Wang et al., 2019a).

Unsupervised learning based on STDP like learning rule has been demonstrated in a fully memristive neural network integrated with memristive synapse and diffusive memristor-based neurons. Pattern classification has been realized with such fully memristive neural network after unsupervised learning (Wang et al., 2018).

Collective-State Computing

Hopfield networks realistically describe neurophysiological processes and exhibit associative memory behaviors with the system automatically evolving to attractor states (Hopfield 1982). In the Hopfield network, each neuron receives input from all other neurons, and integrate-and-fire neurons can be employed (Eryilmaz et al., 2014). Thus, when a fixed spiking pattern is presented to the neurons, the synapse can receive overlapping stimuli between self-generated spikes and the input spikes in its two terminals (Figure 5C). The synapse weight can thus be updated with Hebbian-like rules, such as STDP (Milo et al., 2017). The learned configuration of synaptic weights forms an attractor state. After the learning, even if only part of the spiking pattern is presented to the neurons, the full spiking pattern can be recalled (Milo et al., 2018), which is the basic concept of associative memory or content-addressable memory. The number of attractors that can be learned in a single synaptic array largely depends on the size of the network and is also affected by the learning rules. The Oja rule is reported to have a larger memory capacity roughly 10 times better than the Hebbian rule (Wang et al., 2020a).

The cellular neural network only allows local connections between neighboring neural cells (Chua and Yang 1988). In a standard cellular neural network, the neuron cells are arranged in a 2D array and the synapses

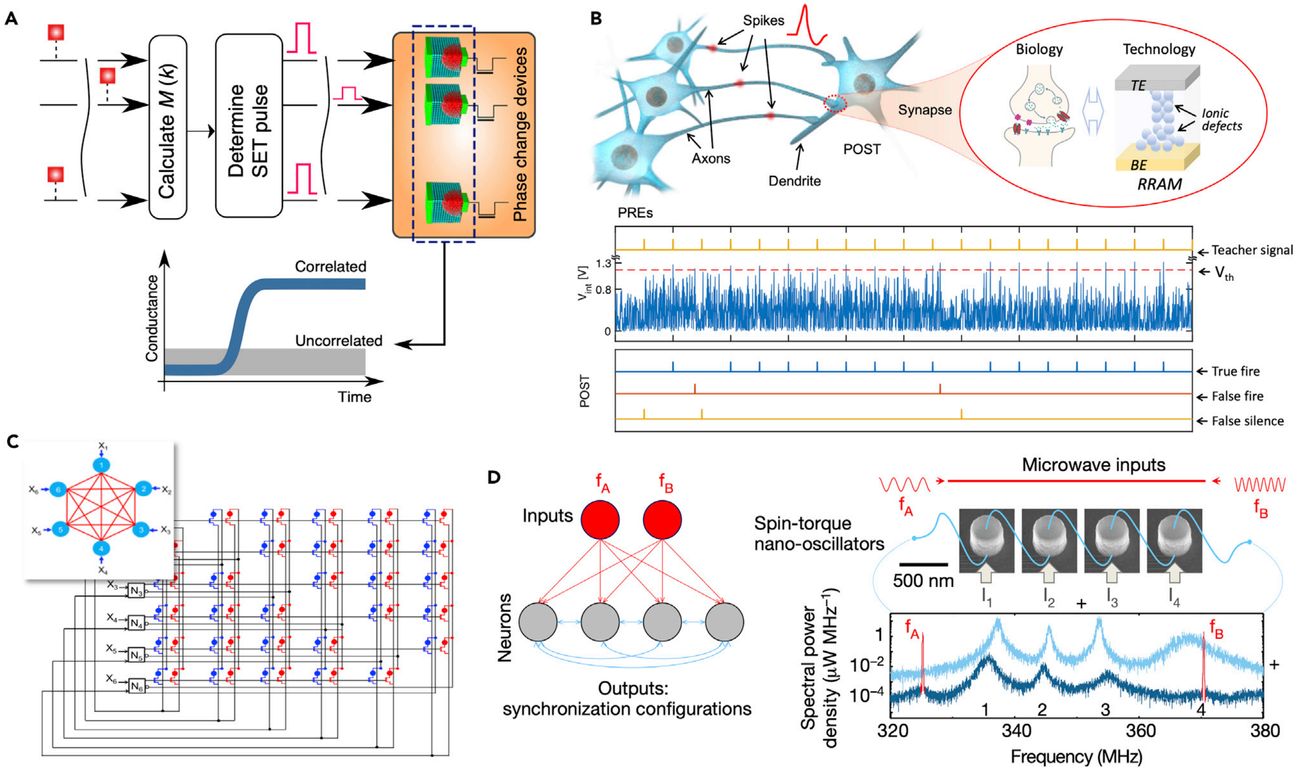


Figure 5. Construction of Memristive Neuromorphic System Utilizing Brain-Inspired Algorithms Enabled by Memristive Devices

(A) Learning and recognition of an image by the detection of the coincidence of simultaneous spikes with the help of PCM synaptic devices capable of the STDP. Reproduced from (Sebastian et al., 2017), copyright © 2016, Springer Nature, CC BY.

(B) Spatiotemporal computation considering the precise times of each spike within a spiking pattern. Reproduced from (Wang et al., 2018b), CC BY.

(C) A Hopfield neural network for associative learning by fully connecting all neurons via bidirectional synapses. Reproduced from (Milo et al., 2017), copyright © IEEE 2017.

(D) Coupled nano-oscillators enabling mimicking of neural synchrony for vowel recognition. Reproduced from (Romera et al., 2018), copyright © 2018, Springer Nature.

bridge each cell with its neighboring cells (Duan et al., 2015). Many two-dimensional tasks such as pattern and image analysis can be solved parallelly with such a 2D arrangement of cellular neural networks. Owing to the localized synaptic connections or communication between cells and the fully parallel operations of each cell, it is more suitable to be implemented with the hardware within the neuromorphic regime. Memristive synapses can further reduce the area cost compared with a CMOS only solution (Dominguez-Castro et al., 1997). Simulation results have shown that memristive cellular neural networks execute functions of image processing such as horizontal line detection, edge extraction, and noise removal (Duan et al., 2015). The non-linear I-V relation in memristive devices is incorporated in the analysis and simulation and has not proven to be an issue. However, the impact of other non-idealities of the memristive devices on the system performance needs further investigation. Disordered graphical network maps can be viewed as a special case of cellular neural networks. Theoretical analysis of these networks based on memristive connections shows much richer dynamic behaviors (Caravelli et al. 2017).

The oscillation network is another example of collective-state computation. Coupled with a memristive circuit, two self-sustained relaxation oscillators show frequency synchronization and phase locking (Ignatov et al., 2016). This is believed to convey two essential principles of biological computing, namely, synchronization and memory. More recently, the memristive-coupled oscillator network is extended for temporal binding of different attributes of the same object (Marina Ignatov et al., 2017). In another work, the oscillators are implemented by spin-torque memristive devices (Figure 5D), while the coupling factors among the oscillators are tuned by the direct current through each oscillator (Romera et al., 2018). Vowel recognition with four coupled spin-torque oscillators was experimentally demonstrated.

Outlook

The integration of the memristor-based neuromorphic computing systems requires a detailed co-design at various levels, ranging from material optimization to system engineering. At each level, there are various integration methods depending on the approach and goal of the final system. Joint efforts and collaborations from experts in various research fields are needed. This perspective clarifies the goals of the efforts at various integration levels for two approaches to memristive neuromorphic systems: the DL accelerator and the brain-like computation.

The implementation of state-of-the-art DL techniques enabled by the material- and device-level optimization and by the array level adaption has been a fruitful exploration in memristive neuromorphic computing. It can be viewed as a model for the co-design of memristive devices and algorithms. This methodology mainly relies on the maturity of DL algorithms and the popularity of these techniques in the AI era. Thanks to this popularity, materials scientists and electrical engineers working on memristive devices have sufficient prior knowledge to explore current machine learning infrastructures and slightly modify the algorithms as needed for real situations encountered in memristive synapses. Prototype systems realizing benchmark cognitive functions, for instance, the image classification for MNIST dataset, CIFAR, have been demonstrated or simulated (Yao et al., 2020; Ambrogio et al., 2018). However, demonstrations of large-scale fully integrated memristive neuromorphic solutions for DL acceleration beyond the relatively small tasks (e.g., MNIST, CIFAR), toward more practical applications (e.g., in the scale of ImageNet [He et al., 2016]), are still lacking.

Exploiting unique features, including those traditionally viewed as non-idealities, of memristive devices enables a more direct and efficient implementation of brain-inspired algorithms, resulting in artificial synapses, dendrites, and neurons closely resembling their biological counterparts, as well as some basic functionalities in biological systems. However, these brain-inspired algorithms do not directly result in practical computational capabilities. Compared with memristive DL accelerators, memristive brain-like computations are limited to a smaller scale or toy applications so far. Besides the technical issues that need to be addressed, memristive SNNs are relatively underdeveloped mainly because of the lack of a clear understanding of biological information representations and processes that occur in the brain.

SNN and collective-state computation are two possible frameworks, both resembling essential features of biological computations, to utilize the brain-inspired algorithms for brain-like computation. The artificial synapses, neurons, and dendrites that can faithfully emulate their biological counterparts may eventually provide building blocks for bio-realistic artificial neural networks. Such neural networks not only serve as computation tools that can generate natural intelligence but also act as faithful biological emulators to verify neuroscience principles. Compared with the biological tissues that essentially compose a "Blackbox" for neuroscience experiments, such an electronic testbed could be considered a "Whitebox" where every node in the neural network can be monitored, measured, and understood. In this way, memristor-based brain-like neural networks will not only benefit from, but also be beneficial for, the understanding of how biological neural networks naturally process information. Co-design between memristive hardware and neural network algorithms is critical for developing such brain-like neural networks.

ACKNOWLEDGMENT

This work was partially supported by Air Force Office of Scientific Research (AFOSR) for the support through the MURI program under Contract No. FA9550-19-1-0213. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of AFRL. This work has received funding from the European Union's Horizon 2020 research and innovation program (grant agreement No. 824164).

AUTHOR CONTRIBUTIONS

J.J.Y., W.W., W.S., and P.Y. developed the original conception and structure of this perceptive paper. W.W. prepared the original draft. Y.L., J.V.N., Q.Q., and D.I. contributed to the solidation of the conception and revision of the manuscript. J.J.Y. supervised this work.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Agmon-Snir, H., Carr, C.E., and Rinzel, J. (1998). The role of dendrites in auditory coincidence detection. Nature *393*, 268–270, https://doi.org/10.1038/246170a0.

Alibart, F., Zamanidoost, E., and Strukov, D.B. (2013). Pattern classification by memristive crossbar circuits using ex situ and in situ training. Nat. Commun. *4*, 2072, https://doi.org/10.1038/ncomms3072.

Ambrogio, S., Ciocchini, N., Laudato, M., Milo, V., Pirovano, A., Fantini, P., and Ielmini, D. (2016). Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses. Front. Neurosci. *10*, 56, https://doi.org/10.3389/fnins.2016.00056.

Ambrogio, S., Narayanan, P., Tsai, H., Shelby, R.M., Boybat, I., Di Nolfo, C., Sidler, S., Giordano, M., Bodini, M., Farinha, N.C.P., et al. (2018). Equivalent-accuracy accelerated neural-network training using Analogue memory. Nature *558*, 60–67, https://doi.org/10.1038/s41586-018-0180-5.

Bellec, G., Scherr, F., Anand, S., Elias, H., Salaj, D., Legenstein, R., and Maass, W. (2020). A solution to the learning dilemma for recurrent networks of spiking neurons. Nat. Commun. *11*, 3625, https://doi.org/10.1038/s41467-020-17236-y.

Bi, G.Q., and Poo, M.M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J. Neurosci. *18*, 10464–10472, https://doi.org/10.1523/JNEUROSCI.18-24-10464.1998.

Borders, W.A., Pervaiz, A.Z., Fukami, S., Camsari, K.Y., Ohno, H., and Datta, S. (2019). Integer factorization using stochastic magnetic tunnel junctions. Nature *573*, 390–393, https://doi.org/10.1038/s41586-019-1557-9.

Bricalli, A., Ambrosi, E., Laudato, M., Maestro, M., Rodriguez, R., and Ielmini, D. (2018). Resistive switching device technology based on silicon oxide for improved ON–OFF ratio—Part II: select devices. IEEE Trans. Electron. Devices *65*, 122–128, https://doi.org/10.1109/TED.2017.2776085.

van de Burgt, Y., Lubberman, E., Fuller, E.J., Keene, S.T., Faria, G.C., Agarwal, S., Marinella, M.J., Talin, A.A., and Salleo, A. (2017). A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing. Nat. Mater. *16*, 414–418, https://doi.org/10.1038/nmat4856.

Burr, G.W., Shelby, R.M., Sidler, S., Di Nolfo, C., Jang, J., Boybat, I., Shenoy, R.S., Narayanan, P., Virwani, K., Giacometti, E.U., et al. (2015). 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. Devices *62*, 3498–3507, https://doi.org/10.1109/TED.2015.2439635.

Cai, F., Correll, J.M., Lee, S.H., Lim, Y., Bothra, V., Zhang, Z., Flynn, M.P., and Lu, W.D. (2019). "A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations. Nat. Electron. *2*, 290–299, https://doi.org/10.1038/s41928-019-0270-x.

Cai, F., Kumar, S., Van Vaerenbergh, T., Xia, S., Liu, R., Li, C., Liu, Z., Foltin, M., Yu, S., Xia, Q., et al. (2020). Power-efficient combinatorial optimization using intrinsic noise in memristor Hopfield neural networks. Nat. Electron. *3*, 409–418, https://doi.org/10.1038/s41928-020-0436-6.

Caravelli, F., Traversa, F.L., and Di Ventra, M. (2017). Complex dynamics of memristive circuits: analytical results and universal slow relaxation. Phys. Rev. E *95*, 022140, https://doi.org/10.1103/PhysRevE.95.022140.

Carboni, R., and Ielmini, D. (2019). Stochastic memory devices for security and computing. Adv. Electron. Mater. *5*, 1900198, https://doi.org/10.1002/aelm.201900198.

Chen, W.H., Dou, C., Li, K.X., Lin, W.Y., Li, P.Y., Huang, J.H., Wang, J.H., Wei, W.C., Xue, C.X., Chu, Y.C., et al. (2019). CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors. Nat. Electron. *2*, 420–428, https://doi.org/10.1038/s41928-019-0288-0.

Chi, P., Li, S., Xu, C., Zhang, T., Zhao, J., Liu, Y., Wang, Y., and Yuan, X. (2016). PRIME: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory. In Annual International Symposium on Computer Architecture (ISCA) (IEEE), pp. 27–39, https://doi.org/10.1109/ISCA.2016.13.

Choi, S., Tan, S.H., Li, Z., Kim, Y., Choi, C., Chen, P.Y., Yeon, H., Yu, S., and Kim, J. (2018). SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations. Nat. Mater. *17*, 335–340, https://doi.org/10.1038/s41563-017-0001-5.

Chua, L.O., and Lin, Y. (1988). Cellular neural networks: theory. IEEE Trans. Circ. Syst. *35*, 1257–1272, https://doi.org/10.1109/31.7600.

Csaba, G., and Porod, W. (2020). Coupled oscillators for computing: a review and perspective. Appl. Phys. Rev. *7*, 011302, https://doi.org/10.1063/1.5120412.

Dominguez-Castro, R., Servando Espejo, A., Rodriguez-Vazquez, R.A., Carmona, P.F., Zarandy, A., Szolgay, P., Sziranyi, T., and Roska, T. (1997). A 0.8-Mm CMOS two-dimensional programmable mixed-signal focal-plane array processor with on-chip binary imaging and instructions storage. IEEE J. Solid State Circ. *32*, 1013–1026, https://doi.org/10.1109/4.597292.

Du, C., Cai, F., Zidan, M.A., Wen, M., Lee, S.H., and Wei, D.L. (2017). Reservoir computing using dynamic memristors for temporal information processing. Nat. Commun. *8*, 2204, https://doi.org/10.1038/s41467-017-02337-y.

Duan, S., Hu, X., Dong, Z., Wang, L., and Mazumder, P. (2015). Memristor-based cellular nonlinear/neural network: design, analysis, and applications. IEEE Trans. Neural Networks Learn. Syst. *26*, 1202–1213, https://doi.org/10.1109/TNNLS.2014.2334701.

Eryilmaz, S.B., Kuzum, D., Jeyasingh, R., Kim, S.B., BrightSky, M., Lam, C., and Philip Wong, H.-S. (2014). Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array. Front. Neurosci. *8*, 205, https://doi.org/10.3389/fnins.2014.00205.

Eryilmaz, S.B., Neftci, E., Joshi, S., Kim, S., Brightsky, M., Lan Lung, H., Lam, C., Cauwenberghs, G., and Philip Wong, H.S. (2016). Training a probabilistic graphical model with resistive switching electronic synapses. IEEE Trans. Electron Devices *63*, 5004–5011, https://doi.org/10.1109/TED.2016.2616483.

Fang, H., Shrestha, A., Zhao, Z., Li, Y., and Qiu, Q. (2019). An event-driven neuromorphic system with biologically plausible temporal dynamics. In 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD) (IEEE), pp. 1–8, https://doi.org/10.1109/ICCAD45719.2019.8942083.

Gao, L., Chen, P.-Y., and Yu, S. (2016). Demonstration of convolution kernel operation on resistive cross-point Array. IEEE Electron Device Lett. *37*, 870–873, https://doi.org/10.1109/LED.2016.2573140.

Gokmen, T., Onen, M., and Haensch, W. (2017). Training deep convolutional neural networks with resistive cross-point devices. Front. Neurosci. *11*, 538, https://doi.org/10.3389/fnins.2017.00538.

Gokmen, T., Rasch, M.J., and Haensch, W. (2019). The marriage of training and inference for scaled deep learning analog hardware. In 2019 IEEE International Electron Devices Meeting (IEDM) (IEEE), pp. 22.3.1–22.3.4, https://doi.org/10.1109/IEDM19573.2019.8993573.

Gokmen, T., and Vlasov, Y. (2016). Acceleration of deep neural network training with resistive cross-point devices: design considerations. Front. Neurosci. *10*, 333, https://doi.org/10.3389/fnins.2016.00333.

Hao, S., Ji, X., Zhong, S., Pang, K.Y., Lim, K.G., Chong, T.C., and Zhao, R. (2020). "A monolayer leaky integrate-and-fire neuron for 2D memristive neuromorphic networks. Adv. Electron. Mater. *6*, 1901335, https://doi.org/10.1002/aelm.201901335.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), *45*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

Hebb, D.O. (1949). The Organization of Behavior: A Neuropsychological Theory (John Wiley & Sons, Inc.; Champan & Hall, Limited). https://doi.org/10.2307/1418888.

Hirtzlin, T., Bocquet, M., Penkovsky, B., Klein, J.-O., Nowak, E., Vianello, E., Portal, J.-M., and Querlioz, D. (2020). Digital biologically plausible implementation of binarized neural networks with differential hafnium oxide resistive memory arrays. Front. Neurosci. *13*, 1383, https://doi.org/10.3389/fnins.2019.01383.

Hodgkin, A.L., and Huxley, A.F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. *117*, 500–544, https://doi.org/10.1113/jphysiol.1952.sp004764.

Hopfield, J.J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. Proc. Natl. Acad. Sci. USA *81*, 3088–3092, https://doi.org/10.1073/pnas.81.10.3088.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. USA *79*, 2554–2558, https://doi.org/10.1073/pnas.79.8.2554.

Hu, M., Graves, C.E., Li, C., Li, Y., Ge, N., Montgomery, E., Davila, N., Jiang, H., Williams, R.S., Yang, J.J., et al. (2018). Memristor-based analog computation and neural network classification with a dot product engine. Adv. Mater. *30*, 1705914, https://doi.org/10.1002/adma.201705914.

Huo, Q., Liu, M., Song, R., Lei, D., Luo, Q., Wu, Z., Wu, Z., Zhao, X., Zhang, F., and Li, L. (2020). Demonstration of 3D convolution kernel function based on 8-layer 3D vertical resistive random access memory. IEEE Electron. Device Lett. *41*, 497–500, https://doi.org/10.1109/LED.2020.2970536.

Ielmini, D., and Pedretti, G. (2020). Device and circuit architectures for in-memory computing. Adv. Intell. Syst. *2*, 2000040, https://doi.org/10.1002/aisy.202000040.

Ielmini, D., and Wong, H.-S.P. (2018). In-memory computing with resistive switching devices. Nat. Electron. *1*, 333–343, https://doi.org/10.1038/s41928-018-0092-2.

Ielmini, D., and Zhang, Y. (2007). Analytical model for subthreshold conduction and threshold switching in chalcogenide-based memory devices. J. Appl. Phys. *102*, 054517, https://doi.org/10.1063/1.2773688.

Ignatov, M., Hansen, M., Ziegler, M., and Kohlstedt, H. (2016). Synchronization of two memristively coupled van Der pol oscillators. Appl. Phys. Lett. *108*, 084105, https://doi.org/10.1063/1.4942832.

Ignatov, M., Ziegler, M., Hansen, M., and Kohlstedt, H. (2017). Memristive stochastic plasticity enables mimicking of neural synchrony: memristive circuit emulates an optical illusion. Sci. Adv. *3*, e1700849, https://doi.org/10.1126/sciadv.1700849.

Jeong, Y., Zidan, M.A., and Lu, W.D. (2018). Parasitic effect analysis in memristor-array-based neuromorphic systems. IEEE Trans. Nanotechnol. *17*, 184–193, https://doi.org/10.1109/TNANO.2017.2784364.

Jiang, H., Han, L., Lin, P., Wang, Z., Jang, M.H., Wu, Q., Barnell, M., Yang, J.J., Xin, H.L., and Xia, Q. (2016). Sub-10 Nm Ta channel responsible for superior performance of a HfO_2 memristor. Sci. Rep. *6*, 28525, https://doi.org/10.1038/srep28525.

Joshi, V., Le Gallo, M., Haefeli, S., Boybat, I., Nandakumar, S.R., Piveteau, C., Dazzi, M., Rajendran, B., Sebastian, A., and Eleftheriou, E. (2020). Accurate deep neural network inference using computational phase-change memory. Nat. Commun. *11*, 2473, https://doi.org/10.1038/s41467-020-16108-9.

Kaiser, J., Mostafa, H., and Neftci, E. (2020). Synaptic plasticity dynamics for deep continuous local learning (DECOLLE). Front. Neurosci. *14*, 424. https://doi.org/10.3389/fnins.2020.00424.

Kim, S., Du, C., Sheridan, P., Wen, M., Choi, S., and Lu, W.D. (2015). Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity. Nano Lett. *15*, 2203–2211, https://doi.org/10.1021/acs.nanolett.5b00697.

Krestinskaya, O., Khaled Nabil, S., and Alex Pappachen, J. (2018). Analog backpropagation learning circuits for memristive crossbar neural networks. In 2018 IEEE International Symposium on Circuits and Systems (ISCAS) (IEEE), pp. 1–5, https://doi.org/10.1109/ISCAS.2018.8351344.

Lavzin, M., Rapoport, S., Alon, P., Garion, L., and Schiller, J. (2012). Nonlinear dendritic processing determines angular tuning of barrel cortex neurons in vivo. Nature *490*, 397–401, https://doi.org/10.1038/nature11451.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proc. IEEE *86*, 2278–2324, https://doi.org/10.1109/5.726791.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444, https://doi.org/10.1038/nature14539.

Li, B., Gu, P., Wang, Y., and Yang, H. (2016). Exploring the precision limitation for RRAM-based analog approximate computing. IEEE Des. Test *33*, 51–58, https://doi.org/10.1109/MDAT.2015.2487218.

Li, C., Belkin, D., Li, Y., Peng, Y., Hu, M., Ning, G., Jiang, H., Montgomery, E., Lin, P., Wang, Z., et al. (2018a). Efficient and self-adaptive in-situ learning in multilayer memristor neural networks. Nat. Commun. *9*, 2385, https://doi.org/10.1038/s41467-018-04484-2.

Li, C., Graves, C.E., Xia, S., Miller, D., Foltin, M., Pedretti, G., and Strachan, J.P. (2020a). Analog content-addressable memories with memristors. Nat. Commun. *11*, 1638, https://doi.org/10.1038/s41467-020-15254-4.

Li, C., Hu, M., Li, Y., Jiang, H., Ning, G., Montgomery, E., Zhang, J., Song, W., Dávila, N., Graves, C.E., et al. (2018b). Analogue signal and image processing with large memristor crossbars. Nat. Electron. *1*, 52–59, https://doi.org/10.1038/s41928-017-0002-z.

Li, X., Tang, J., Zhang, Q., Gao, B., Yang, J.J., Song, S., Wu, W., Zhang, W., Yao, P., Deng, N., et al. (2020b). Power-efficient neural network with artificial dendrites. Nat. Nanotechnol. *15*, 776–782, https://doi.org/10.1038/s41565-020-0722-5.

Li, C., Wang, Z., Rao, M., Belkin, D., Song, W., Jiang, H., Yan, P., Li, Y., Lin, P., Hu, M., et al. (2019). Long short-term memory networks in memristor crossbar arrays. Nat. Machine Intelligence *1*, 49–57, https://doi.org/10.1038/s42256-018-0001-4.

Lin, P., Li, C., Wang, Z., Li, Y., Jiang, H., Song, W., Rao, M., Zhuo, Y., Upadhyay, N.K., Barnell, M., et al. (2020). Three-dimensional memristor circuits as complex neural networks. Nat. Electron. *3*, 225–232, https://doi.org/10.1038/s41928-020-0397-9.

Lin, Y., Xiaobo, S.H., He, Q., Huaqiang, W., Qingtian, Z., Jianshi, T., Bin, G., Zhu, J., Lu, J., Hu, X.S., et al. (2019). Bayesian neural network realization by exploiting inherent stochastic characteristics of analog RRAM. In 2019 IEEE International Electron Devices Meeting (IEDM) (IEEE), pp. 14.6.1–14.6.4.

Linares-Barranco, B., and Serrano-Gotarredona, T. (2009). Memristance can explain spike-time-dependent-plasticity in neural synapses. Nat. Precedings, 1–4, https://doi.org/10.1038/npre.2009.3010.1.

Liu, Q., Gao, B., Yao, P., Wu, D., Chen, J., Pang, Y., Zhang, W., Liao, Y., Xue, C.X., Chen, W.H., et al. (2020). A fully integrated analog ReRAM based 78.4TOPS/W compute-in-memory chip with fully parallel MAC computing. In 2020 IEEE International Solid- State Circuits Conference - (ISSCC) (IEEE), pp. 500–502.

Liu, B., Li, H., Chen, Y., Li, X., Huang, T., Wu, Q., and Barnell, M. (2014). Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems. In 2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD) (IEEE), pp. 63–70, https://doi.org/10.1109/ICCAD.2014.7001330.

Luo, Q., Gong, T., Xu, X., Peng, Y., Li, X., Lu, T., Qi, L., Lv, H., and Liu, M. (2019). Complementary memory cell based on field-programmable ferroelectric diode for ultra-low power current-SA free BNN applications. In 2019 IEEE International Electron Devices Meeting (IEDM) (IEEE), pp. 38.5.1–38.5.4.

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. Neural Networks *10*, 1659–1671, https://doi.org/10.1016/S0893-6080(97)00011-7.

Mahmoodi, M.R., Prezioso, M., and Strukov, D.B. (2019). Versatile stochastic dot product circuits based on nonvolatile memories for high performance neurocomputing and neurooptimization. Nat. Commun. *10*, 5113, https://doi.org/10.1038/s41467-019-13103-7.

Masquelier, T., Guyonneau, R., and Thorpe, S.J. (2009). Competitive STDP-based spike pattern learning. Neural Comput. *21*, 1259–1276, https://doi.org/10.1162/neco.2008.06-08-804.

McCulloch, W.S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. *5*, 115–133, https://doi.org/10.1007/BF02478259.

Mehonic, A., and Kenyon, A.J. (2016). Emulating the electrical activity of the neuron using a silicon oxide RRAM cell. Front. Neurosci. *10*, 57, https://doi.org/10.3389/fnins.2016.00057.

Milo, V., Pedretti, G., Carboni, R., Calderoni, A., Ramaswamy, N., Ambrogio, S., and Ielmini, D. (2016). Demonstration of hybrid CMOS/RRAM neural networks with spike time/rate-dependent plasticity. In 2016 IEEE International Electron Devices Meeting (IEDM) (IEEE), pp. 16.8.1–16.8.4, https://doi.org/10.1109/IEDM.2016.7838435.

Milo, V., Ielmini, D., and Chicca, E. (2017). Attractor networks and associative memories with STDP learning in RRAM synapses. In 2017

IEEE International Electron Devices Meeting (IEDM) (IEEE), pp. 11.2.1–11.2.4, https://doi.org/10.1109/IEDM.2017.8268369.

Midya, R., Wang, Z., Asapu, S., Zhang, X., Rao, M., Song, W., Ye, Z., Upadhyay, N., Xia, Q., and Yang, J.J. (2019). Reservoir computing using diffusive memristors. Adv. Intell. Syst. *1*, 1900084, https://doi.org/10.1002/aisy.201900084.

Milo, V., Chicca, E., and Ielmini, D. (2018). Brain-inspired recurrent neural network with plastic RRAM synapses. In IEEE International Symposium on Circuits and Systems (ISCAS) (IEEE), pp. 1–5, https://doi.org/10.1109/ISCAS.2018.8351523.

Milo, V., Zambelli, C., Olivo, P., Pérez, E., Mahadevaiah, M.K., Ossorio, O.G., Wenger, Ch., and Ielmini, D. (2019). Multilevel HfO_2-based RRAM devices for low-power neuromorphic networks. APL Mater. *7*, 081120, https://doi.org/10.1063/1.5108650.

Mizrahi, A., Hirtzlin, T., Fukushima, A., Kubota, H., Yuasa, S., Grollier, J., and Querlioz, D. (2018). Neural-like computing with populations of superparamagnetic basis functions. Nat. Commun. *9*, 1533, https://doi.org/10.1038/s41467-018-03963-w.

Moon, J., Wen, M., Jong Hoon, S., Cai, F., Du, C., Lee, S.H., and Lu, W.D. (2019). Temporal data classification and forecasting using a memristor-based reservoir computing system. Nat. Electron. *2*, 480–487, https://doi.org/10.1038/s41928-019-0313-3.

Ni, K., Yin, X., Franchesca Laguna, A., Joshi, S., Dünkel, S., Martin, T., Müeller, J., Beyer, S., Niemier, M., Hu, X.S., and Datta, S. (2019). Ferroelectric ternary content-addressable memory for one-shot learning. Nat. Electron. *2*, 521–529, https://doi.org/10.1038/s41928-019-0321-3.

Ohno, T., Hasegawa, T., Tsuruoka, T., Terabe, K., Gimzewski, J.K., and Aono, M. (2011). Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. Nat. Mater. *10*, 591–595, https://doi.org/10.1038/nmat3054.

Panda, P., Aketi, S.A., and Roy, K. (2020). Toward scalable, efficient, and accurate deep spiking neural networks With backward residual connections, stochastic softmax, and hybridization. Front. Neurosci. *14*, 653. https://doi.org/10.3389/fnins.2020.00653.

Park, E., Kim, D., and Yoo, S. (2018). Energy-efficient neural network accelerator based on outlier-aware low-precision computation. In 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA) (IEEE), pp. 688–698, https://doi.org/10.1109/ISCA.2018.00063.

Pedretti, G., Milo, V., Ambrogio, S., Carboni, R., Bianchi, S., Calderoni, A., Ramaswamy, N., Spinelli, A.S., and Ielmini, D. (2017). Memristive neural network for on-line learning and tracking with brain-inspired spike timing dependent plasticity. Sci. Rep. *7*, 5288, https://doi.org/10.1038/s41598-017-05480-0.

Pickett, M.D., Medeiros-Ribeiro, G., and Williams, R.S. (2013). A scalable neuristor built with Mott memristors. Nat. Mater. *12*, 114–117, https://doi.org/10.1038/nmat3510.

Prezioso, M., Mahmoodi, M.R., Merrikh Bayat, F., Nili, H., Kim, H., Vincent, A., and Strukov, D.B. (2018). Spike-timing-dependent plasticity learning of coincidence detection with passively integrated memristive circuits. Nat. Commun. *9*, 5311, https://doi.org/10.1038/s41467-018-07757-y.

Prezioso, M., Merrikh-Bayat, F., Hoskins, B.D., Adam, G.C., Likharev, K.K., and Strukov, D.B. (2015). Training and operation of an integrated neuromorphic network based on metal-oxide memristors. Nature *521*, 61–64, https://doi.org/10.1038/nature14441.

Qiao, N., Mostafa, H., Corradi, F., Osswald, M., Stefanini, F., Sumislawska, D., and Indiveri, G. (2015). A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. Front. Neurosci. *9*, 141, https://doi.org/10.3389/fnins.2015.00141.

Romera, M., Talatchian, P., Tsunegi, S., Abreu Araujo, F., Cros, V., Bortolotti, P., Trastoy, J., Yakushiji, K., Fukushima, A., Kubota, H., et al. (2018). Vowel recognition with four coupled spin-torque nano-oscillators. Nature *563*, 230–234, https://doi.org/10.1038/s41586-018-0632-y.

Roy, K., Jaiswal, A., and Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. Nature *575*, 607–617, https://doi.org/10.1038/s41586-019-1677-2.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature *323*, 533–536, https://doi.org/10.1038/323533a0.

Sebastian, A., Gallo, M.L., Khaddam-Aljameh, R., and Eleftheriou, E. (2020). Memory devices and applications for in-memory computing. Nat. Nanotechnol. *15*, 529–544, https://doi.org/10.1038/s41565-020-0655-z.

Sebastian, A., Tuma, T., Papandreou, N., Le Gallo, M., Kull, L., Parnell, T., and Eleftheriou, E. (2017). Temporal correlation detection using computational phase-change memory. Nat. Commun. *8*, 1115, https://doi.org/10.1038/s41467-017-01481-9.

Shafiee, A., Nag, A., Muralimanohar, N., Balasubramonian, R., Strachan, J.P., Hu, M., Stanley Williams, R., and Srikumar, V. (2016). ISAAC: a convolutional neural network accelerator with in-situ analog arithmetic in crossbars. In Annual International Symposium on Computer Architecture (ISCA) (IEEE), pp. 14–26, https://doi.org/10.1109/ISCA.2016.12.

Shrestha, A., Fang, H., Wu, Q., and Qiu, Q. (2019). Approximating Back-propagation for a Biologically Plausible Local Learning Rule in Spiking Neural Networks. In Proceedings of the International Conference on Neuromorphic Systems (ACM). https://doi.org/10.1145/3354265.3354275.

Song, L., Qian, X., Li, H., and Chen, Y. (2017). PipeLayer: a pipelined ReRAM-based accelerator for deep learning. In 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA) (IEEE), pp. 541–552, https://doi.org/10.1109/HPCA.2017.55.

Stoliar, P., Yamada, H., Toyosaki, Y., and Sawa, A. (2019). Spike-shape dependence of the spike-timing dependent synaptic plasticity in ferroelectric-tunnel-junction synapses. Sci. Rep. *9*, 17740, https://doi.org/10.1038/s41598-019-54215-w.

Suri, M., Bichler, O., Querlioz, D., Cueto, O., Perniola, L., Sousa, V., Vuillaume, D., Gamrat, C.n, and DeSalvo, B. (2011). Phase change memory as synapse for ultra-dense neuromorphic systems: application to complex visual pattern extraction. In 2011 International Electron Devices Meeting (IEEE), pp. 4.4.1–4.4.4, https://doi.org/10.1109/IEDM.2011.6131488.

Tang, J., Douglas, B., Kim, S., Copel, M., Gokmen, T., Todorov, T., Shin, S., Lee, K., Solomon, P., Chan, K., et al. (2018). ECRAM as scalable synaptic cell for high-speed, low-power neuromorphic computing. In 2018 IEEE International Electron Devices Meeting (IEDM) (IEEE), pp. 13.1.1–13.1.4.

Tuma, T., Pantazi, A., Le Gallo, M., Sebastian, A., and Eleftheriou, E. (2016). Stochastic phase-change neurons. Nat. Nanotechnol. *11*, 693–699, https://doi.org/10.1038/nnano.2016.70.

del Valle, J., Salev, P., Tesler, F., Vargas, N.M., Kalcheim, Y., Wang, P., Trastoy, J., Lee, M.H., Kassabian, G., Ramírez, J.G., et al. (2019). Subthreshold firing in Mott nanodevices. Nature *569*, 388–392, https://doi.org/10.1038/s41586-019-1159-6.

VanRullen, R., Guyonneau, R., and Thorpe, S.J. (2005). Spike times make sense. Trends Neurosci. *28*, 1–4, https://doi.org/10.1016/j.tins.2004.10.010.

Wang, Z., Joshi, S., Savel'ev, S.E., Jiang, H., Midya, R., Lin, P., Hu, M., Ge, N., Strachan, J.P., Li, Z., et al. (2017). Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. Nat. Mater. *16*, 101–108, https://doi.org/10.1038/nmat4756.

Wan, W., Kubendran, R., Eryilmaz, S.B., Zhang, W., Liao, Y., Wu, D., Stephen, D., Gao, B., Raina, P., Joshi, S., et al. (2020). A 74 TMACS/W CMOS-RRAM neurosynaptic core with dynamically reconfigurable dataflow and in-situ transposable weights for probabilistic graphical models. In 2020 IEEE International Solid- State Circuits Conference - (ISSCC) (IEEE), pp. 498–500.

Wang, Z., Ambrogio, S., Balatti, S., and Ielmini, D. (2015). A 2-transistor/1-resistor artificial synapse capable of communication and stochastic learning in neuromorphic systems. Front. Neurosci. *8*, 438, https://doi.org/10.3389/fnins.2014.00438.

Wang, W., Bricalli, A., Laudato, M., Ambrosi, E., Covi, E., and Ielmini, D. (2018a). Physics-based modeling of volatile resistive switching memory (RRAM) for crosspoint selector and neuromorphic computing. In IEEE International Electon Device Meeting (IEEE), pp. 40.3.1–40.3.4, https://doi.org/10.1109/IEDM.2018.8614556.

Wang, Z., Li, C., Lin, P., Rao, M., Nie, Y., Song, W., Qiu, Q., Li, Y., Yan, P., Strachan, J.P., et al. (2019c). In situ training of feed-forward and recurrent convolutional memristor networks. Nat. Machine Intelligence *1*, 434–442, https://doi.org/10.1038/s42256-019-0089-1.

Wang, W., Pedretti, G., Milo, V., Carboni, R., Calderoni, A., Ramaswamy, N., Spinelli, A.S., and Ielmini, D. (2018b). Learning of spatiotemporal patterns in a spiking neural network with resistive switching synapses. Sci. Adv. *4*, eaat4752, https://doi.org/10.1126/sciadv.aat4752.

Wang, Z., Rao, M., Han, J.-W., Zhang, J., Lin, P., Li, Y., Li, C., Song, W., Asapu, S., Midya, R., et al. (2018d). Capacitive neural network with neuro-transistors. Nat. Commun. *9*, 3208, https://doi.org/10.1038/s41467-018-05677-5.

Wang, W., Pedretti, G., Milo, V., Carboni, R., Calderoni, A., Ramaswamy, N., Spinelli, A.S., and Ielmini, D. (2019a). Computing of temporal information in spiking neural networks with ReRAM synapses. Faraday Discuss. *213*, 453–469, https://doi.org/10.1039/C8FD00097B.

Wang, W., Wang, M., Ambrosi, E., Bricalli, A., Laudato, M., Sun, Z., Chen, X., and Ielmini, D. (2019b). Surface diffusion-limited lifetime of silver and copper nanofilaments in resistive switching devices. Nat. Commun. *10*, 81, https://doi.org/10.1038/s41467-018-07979-0.

Wang, Z., Joshi, S., Savel'ev, S., Song, W., Midya, R., Li, Y., Rao, M., Yan, P., Asapu, S., Zhuo, Y., et al. (2018c). Fully memristive neural networks for pattern classification with unsupervised learning. Nat. Electron. *1*, 137–145, https://doi.org/10.1038/s41928-018-0023-2.

Wang, Z., Wu, H., Burr, G.W., Hwang, C.S., Wang, K.L., Xia, Q., and Yang, J.J. (2020c). Resistive switching materials for information processing. Nat. Rev. Mater. *5*, 173–195, https://doi.org/10.1038/s41578-019-0159-3.

Wang, Y., Yu, L., Wu, S., Huang, R., and Yang, Y. (2020a). Memristor-based biologically plausible memory based on discrete and continuous attractor networks for neuromorphic systems. Adv. Intell. Syst. *2*, 2000001, https://doi.org/10.1002/aisy.202000001.

Wang, Z., Zeng, T., Ren, Y., Lin, Y., Xu, H., Zhao, X., Liu, Y., and Ielmini, D. (2020b). Toward a generalized bienenstock-cooper-munro rule for spatiotemporal learning via triplet-STDP in memristive devices. Nat. Commun. *11*, 1510, https://doi.org/10.1038/s41467-020-15158-3.

Werbos, P.J. (1988). Generalization of backpropagation with application to a recurrent gas market model. Neural Networks *1*, 339–356, https://doi.org/10.1016/0893-6080(88)90007-X.

Woo, J., Moon, K., Song, J., Lee, S., Kwak, M., Park, J., and Hwang, H. (2016). Improved synaptic behavior under identical pulses using AlO_x/HfO_2 bilayer RRAM array for neuromorphic systems. IEEE Electron Device Lett. *37*, 994–997, https://doi.org/10.1109/LED.2016.2582859.

Wu, Q., Wang, H., Luo, Q., Banerjee, W., Cao, J., Zhang, X., Wu, F., Qi, L., Li, L., and Liu, M. (2018). Full imitation of synaptic metaplasticity based on memristor devices. Nanoscale *10*, 5875–5881, https://doi.org/10.1039/C8NR00222C.

Yan, B., Li, B., Qiao, X., Xue, C.-X., Chang, M.-F., Chen, Y., and Li, H.H. (2019a). "Resistive memory-based in-memory computing: from device and large-scale integration system perspectives. Adv. Intell. Syst. *1*, 1900068, https://doi.org/10.1002/aisy.201900068.

Yan, B., Yang, Q., Chen, W.-H., Chang, K.-T., Su, J.-W., Hsu, C.-H., Li, S.-H., Lee, H.-Y., Sheu, S.-S., Ho, M.-S., et al. (2019b). RRAM-based spiking nonvolatile computing-in-memory processing engine with precision-configurable in situ nonlinear activation. In 2019 Symposium on VLSI Technology (IEEE), pp. T86–T87.

Yang, R., Li, H., Kirby, K., Smithe, H., Kim, T.R., Okabe, K., Pop, E., Fan, J.A., and Wong, H.-S.P. (2019). Ternary content-addressable memory with MoS2 transistors for massively parallel data search. Nat. Electron. *2*, 108–114, https://doi.org/10.1038/s41928-019-0220-7.

Yang, J.J., Strukov, D.B., and Stewart, D.R. (2013). Memristive devices for computing. Nat. Nanotechnol. *8*, 13–24, https://doi.org/10.1038/nnano.2012.240.

Yao, P., Wu, H., Gao, B., Burc Eryilmaz, S., Huang, X., Zhang, W., Zhang, Q., Deng, N., Shi, L., Wong, H.P., et al. (2017). Face classification using electronic synapses. Nat. Commun. *8*, 15199, https://doi.org/10.1038/ncomms15199.

Yao, P., Wu, H., Gao, B., Tang, J., Zhang, Q., Zhang, W., Joshua Yang, J., and He, Q. (2020). Fully hardware-implemented memristor convolutional neural network. Nature *577*, 641–646, https://doi.org/10.1038/s41586-020-1942-4.

Yi, W., Tsang, K.K., Lam, S.K., Bai, X., Crowell, J.A., and Flores, E.A. (2018). Biological plausibility and stochasticity in scalable VO_2 active memristor neurons. Nat. Commun. *9*, 4661, https://doi.org/10.1038/s41467-018-07052-w.

Yu, S., Gao, B., Zheng, F., Yu, H., Kang, J., and Wong, H.-S.P. (2013). Stochastic learning in oxide binary synaptic device for neuromorphic computing. Front. Neurosci. *7*, 186, https://doi.org/10.3389/fnins.2013.00186.

Zhang, W., Peng, X., Wu, H., Gao, B., Hu, H., Zhang, Y., Yu, S., and He, Q. (2019). Design guidelines of RRAM based neural-processing-unit. In Proceedings of the 56th Annual Design Automation Conference 2019 (ACM), pp. 1–6, https://doi.org/10.1145/3316781.3317797.

Zhang, Y., He, W., Wu, Y., Huang, K., Shen, Y., Su, J., Wang, Y., Zhang, Z., Ji, X., Li, G., et al. (2018b). Highly compact artificial memristive neuron with low energy consumption. Small *14*, 1802188, https://doi.org/10.1002/smll.201802188.

Zhang, W., Gao, B., Tang, J., Yao, P., Yu, S., Chang, M.-F., Yoo, H.-J., He, Q., and Wu, H. (2020a). Neuro-inspired computing chips. Nat. Electronics *3*, 371–382, https://doi.org/10.1038/s41928-020-0435-7.

Zhang, X., Wang, W., Qi, L., Zhao, X., Wei, J., Cao, R., Yao, Z., Zhu, X., Zhang, F., Lv, H., et al. (2018a). An artificial neuron based on a threshold switching memristor. IEEE Electron Device Lett. *39*, 308–311, https://doi.org/10.1109/LED.2017.2782752.

Zhang, X., Ye, Z., Luo, Q., Wu, Z., Midya, R., Wang, Z., Song, W., Wang, R., Upadhyay, N.K., Fang, Y., et al. (2020b). An artificial spiking afferent nerve based on Mott memristors for neurorobotics. Nat. Commun. *11*, 51, https://doi.org/10.1038/s41467-019-13827-6.

Zidan, M.A., Jeong, Y.J., and Lu, W.D. (2017). Temporal learning using second-order memristors. IEEE Trans. Nanotechnol. *16*, 721–723, https://doi.org/10.1109/TNANO.2017.2710158.

Zuloaga, S., Liu, R., Chen, P.-Y., and Yu, S. (2015). Scaling 2-layer RRAM cross-point array towards 10 nm node: a device-circuit co-design. In 2015 IEEE International Symposium on Circuits and Systems (ISCAS) (IEEE), pp. 193–196.