


# In-memory computing with emerging memory devices: Status and outlook


Cite as: APL Mach. Learn. 1, 010902 (2023); <https://doi.org/10.1063/5.0136403>

Submitted: 25 November 2022 • Accepted: 24 January 2023 • Published Online: 14 February 2023

 P. Mannocci,  M. Farronato,  N. Lepri, et al.

## COLLECTIONS

 This paper was selected as Featured

 This paper was selected as Scilight



View Online



Export Citation



CrossMark

## ARTICLES YOU MAY BE INTERESTED IN

[Benchmarking energy consumption and latency for neuromorphic computing in condensed matter and particle physics](#)

APL Machine Learning 1, 016101 (2023); <https://doi.org/10.1063/5.0116699>

[Addressing the memory bottleneck with in-memory computing](#)

Scilight 2023, 071106 (2023); <https://doi.org/10.1063/10.0017431>

[Deep language models for interpretative and predictive materials science](#)

APL Machine Learning 1, 010901 (2023); <https://doi.org/10.1063/5.0134317>

# In-memory computing with emerging memory devices: Status and outlook

Cite as: APL Mach. Learn. 1, 010902 (2023); doi: 10.1063/5.0136403

Submitted: 25 November 2022 • Accepted: 24 January 2023 •

Published Online: 14 February 2023



View Online



Export Citation



CrossMark

P. Mannocci,<sup>1</sup>  M. Farronato,<sup>1</sup>  N. Lepri,<sup>1</sup>  L. Cattaneo,<sup>1</sup>  A. Glukhov,<sup>1</sup>  Z. Sun,<sup>2</sup> and D. Ielmini<sup>1,a)</sup> 

## AFFILIATIONS

<sup>1</sup>Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano and IUNET, Piazza L. da Vinci 32, 20133 Milano, Italy

<sup>2</sup>Beijing Advanced Innovation Center for Integrated Circuits, School of Integrated Circuits, Institute for Artificial Intelligence, Peking University, 100871 Beijing, China

<sup>a)</sup>Author to whom correspondence should be addressed: [daniele.ielmini@polimi.it](mailto:daniele.ielmini@polimi.it)

## ABSTRACT

In-memory computing (IMC) has emerged as a new computing paradigm able to alleviate or suppress the memory bottleneck, which is the major concern for energy efficiency and latency in modern digital computing. While the IMC concept is simple and promising, the details of its implementation cover a broad range of problems and solutions, including various memory technologies, circuit topologies, and programming/processing algorithms. This Perspective aims at providing an orientation map across the wide topic of IMC. First, the memory technologies will be presented, including both conventional complementary metal-oxide-semiconductor-based and emerging resistive/memristive devices. Then, circuit architectures will be considered, describing their aim and application. Circuits include both popular crosspoint arrays and other more advanced structures, such as closed-loop memory arrays and ternary content-addressable memory. The same circuit might serve completely different applications, e.g., a crosspoint array can be used for accelerating matrix-vector multiplication for forward propagation in a neural network and outer product for backpropagation training. The different algorithms and memory properties to enable such diversification of circuit functions will be discussed. Finally, the main challenges and opportunities for IMC will be presented.

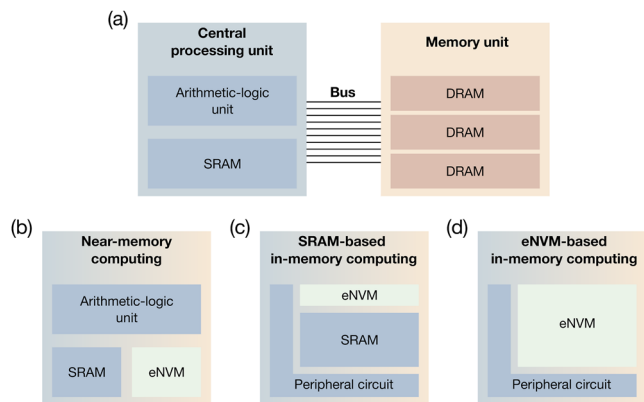
© 2023 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0136403>

## I. INTRODUCTION

Data-intensive computing tasks, such as data analytics, machine learning, and artificial intelligence (AI), require frequent access to the memory to exchange data input, output, and commands. Since the high-density memory is generally off-chip with respect to the central processing unit (CPU), data movement represents a significant overhead in the computation, largely exceeding the energy required for on-chip digital data processing.<sup>1,2</sup> There are two possible directions to tackle this memory bottleneck: one is the optimization of the data throughput in a multi-chip approach, such as the high bandwidth memory (HBM)<sup>3</sup> or the hybrid memory cube (HMC).<sup>4</sup> The second approach is to radically change the computing paradigm by enabling *in situ* computation of data within the memory, which goes by the name of in-memory computing (IMC).<sup>5–8</sup>

Various concepts of IMC have been proposed depending on the degree of integration of memory and processing, as illustrated in Fig. 1. On the one hand, a conventional von Neumann

architecture depicted in Fig. 1(a) has physically separate memory and computing unit sitting on distinct chips, where the movement of input/output/instructions causes significant latency and excess energy consumption. One solution to mitigate these issues is the concept of near-memory computing (NMC) shown in Fig. 1(b), where the embedded nonvolatile memory (eNVM) is integrated on the same chip as the computing unit to minimize the latency.<sup>9,10</sup> Note that eNVM serves as pure data storage for parameters and instructions in NMC, while the static random access memory (SRAM) is used as a cache memory storing intermediate input/output data. A further degree of integration consists of the true IMC approach shown in Fig. 1(c), where the SRAM is used directly as a computational engine, e.g., to accelerate matrix-vector multiplication (MVM).<sup>8</sup> An additional overhead is the need to move the computational parameters from the local eNVM [or an off-chip dynamic random access memory (DRAM)] to the volatile SRAM every time the computation is needed. To mitigate this drawback, the ultimate concept to maximize the integration of memory and processing is IMC within the eNVM, as shown in Fig. 1(d).<sup>7</sup> This



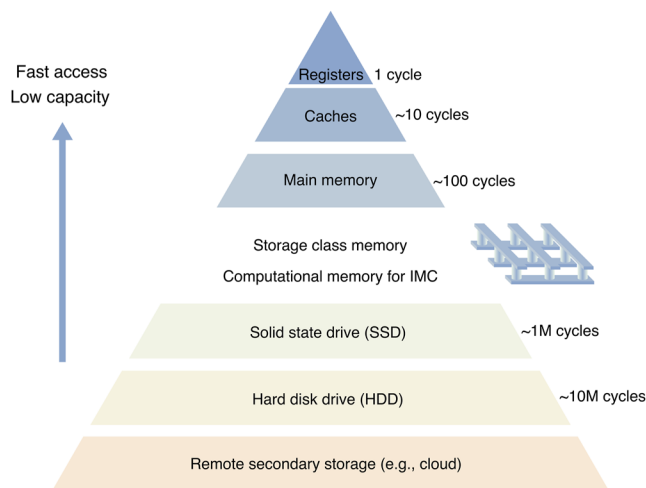
**FIG. 1.** Various degrees of integration between memory and computing units. (a) von Neumann architecture for computing systems, where the central processing unit and the memory unit are physically separated and connected through a data bus. (b) Near-memory computing architecture, where the processing unit is complemented with an eNVM unit to store commands and parameters. (c) SRAM-based in-memory computing architecture, where computation is performed directly within the SRAM unit via dedicated peripherals, while eNVM serves as storage for computational parameters. (d) eNVM-based in-memory computing architecture, where eNVM provides both the nonvolatile storage of computational parameters and the computation.

approach appears as the most promising concept to minimize the data movement, hence energy consumption and latency, although there are significant challenges and trade-offs in terms of throughput, energy efficiency, and accuracy of the processing. Emerging memories represent a promising approach for eNVM in IMC, given several attractive properties of scaling, 3D integration of back-end processing, and nonvolatile storage of computing parameters. The interplay of device technologies, circuit engineering, and algorithms thus requires a strong effort in terms of co-design across multiple disciplines.<sup>11</sup>

This Perspective provides an overview of IMC, including the status of the memory device technologies and the circuit architectures for a broad portfolio of applications. Section II describes the state-of-the-art memory devices for IMC, including both two-terminal and three-terminal emerging memory technologies. Section III presents the concept of analog IMC, highlighting the main challenges from a memory array point of view. Section IV addresses matrix-vector multiplication, which is a fundamental computing primitive at the basis of most IMC applications. Section V reviews the state-of-the-art of closed-loop IMC, which enables highly complex algebraic operations with reduced complexity. Section VI presents an overview of the field of content-addressable memories. Section VII focuses on accelerators for the training of neural networks based on in-memory outer product. Section VIII addresses brain-inspired neuromorphic computing leveraging device physics to reproduce neurobiological processes of sensing and learning. Finally, Sec. IX provides an outlook on the next urgent challenges and opportunities that need to be addressed.

## II. EMERGING MEMORY TECHNOLOGIES

Charge-storage memories based on the complementary metal-oxide-semiconductor (CMOS) technology provide the mainstream

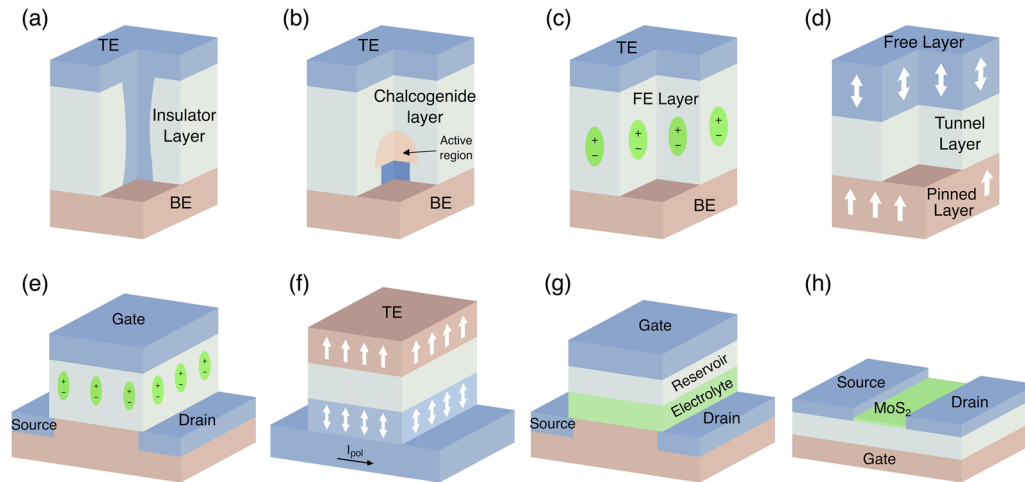


**FIG. 2.** Schematic illustration of the memory hierarchy in traditional CMOS-based computing systems. Registers and cache memories have relatively fast access and low capacity. Moving away from the CPU (top), memories increasingly display slower access and larger capacity. The storage class memory can bridge the gap between high-performance working memory and low-cost storage devices.

memory technology for digital computing systems. Figure 2 illustrates the memory hierarchy of CMOS-based computing systems, including (from top to bottom) on-chip registers and static random access memory (SRAM), followed by off-chip dynamic random access memory (DRAM) and nonvolatile Flash storage. While performance (e.g., access time) decreases from top to bottom, the area density and cost decrease from bottom to top, with NAND flash representing the highest density thanks to 3D integration.<sup>12,13</sup> Within this scenario, emerging memories based on *material* storage have been developed in an effort to provide a better trade-off between performance, area, and cost. In particular, emerging memory devices show unique storage principles relying on the physics of the active materials and offer advantages in terms of scalability,<sup>14</sup> integration in 3D structures,<sup>15,16</sup> and energy efficiency. These properties are also attractive for application as embedded memories in systems-on-chip, where flash memory faces additional integration difficulties due to the high- $\kappa$ /metal-gate process of the silicon front-end circuits.<sup>17</sup> Emerging memories have also attracted a considerable interest for IMC applications thanks to the nonvolatile storage of computing weights, high density, and fast programming/read. Figure 3 shows a summary of the main emerging memories, including two-terminal and three-terminal devices. Table I shows a summary of the properties of emerging memories compared to other nonvolatile memory technologies.<sup>18</sup>

### A. Resistive switching memory (RRAM)

Figure 3(a) schematically shows the resistive random-access memory (RRAM), consisting of a metal-insulator-metal (MIM) stack where the insulating layer serves as the active switching material. The bottom electrode (BE) typically consists of a relatively inert metal, such as Pt or TiN, while the top electrode (TE) is generally a more reactive metal, such as Ti or Ta.<sup>19–21</sup> In most cases, the switching layer is made of a metal oxide<sup>22</sup> although also other materials



**FIG. 3.** Schematic illustration of the emerging memory technologies considered for IMC, including both two-terminal and three-terminal devices. (a) Resistive random access memory (RRAM). (b) Phase change memory (PCM). (c) Ferroelectric resistive random access memory (FeRAM). (d) Spin-transfer torque magnetic random-access memory (STT-MRAM). (e) Ferroelectric field-effect transistor (FeFET). (f) Spin-orbit torque magnetic random-access memory (SOT-MRAM). (g) Electro-chemical random access memory (ECRAM). (h) Memtransistor based on the MoS<sub>2</sub> channel.

**TABLE I.** Comparison of different memory technologies suited for in-memory computing. Reproduced with permission from D. Ielmini and S. Ambrogio, *Nanotechnology* **31**(9), 092001 (2020). Copyright 2020 Author(s), licensed under a Creative Commons Attribution 4.0 License.

Technology	NOR flash	NAND flash	RRAM	PCM	STT-MRAM	FeRAM	FeFET	SOT-MRAM	Li-ion
On/off ratio	10 <sup>4</sup>	10 <sup>4</sup>	10–10 <sup>2</sup>	10 <sup>2</sup> –10 <sup>4</sup>	1.5–2	10 <sup>2</sup> –10 <sup>3</sup>	5–50	1.5–2	40–10 <sup>3</sup>
Multilevel operation	2 bit	4 bit	2 bit	2 bit	1 bit	1 bit	5 bit	1 bit	10 bit
Write voltage (V)	<10	10	<3	<3	<1.5	<3	<5	<1.5	<1
Write time	1–10 μs	0.1–1 ms	<10 ns	~50 ns	<10 ns	~30 ns	~10 ns	<10 ns	<10 ns
Read time	~50 ns	~10 μs	<10 ns	<10 ns	<10 ns	<10 ns	~10 ns	<10 ns	<10 ns
Stand-by power	Low	Low	Low	Low	Low	Low	Low	Low	Low
Write energy [J/bit]	~100 pJ	~10 fJ	0.1–1 pJ	10 pJ	~100 fJ	~100 fJ	<1 fJ	<100 fJ	~100 fJ
Linearity	Low	Low	Low	Low	None	None	Low	None	High
Drift	No	No	Weak	Yes	No	No	No	No	No
Integration density	High	Very high	High	High	High	Low	High	High	Low
Retention	Long	Long	Medium	Long	Medium	Long	Long	Medium	...
Endurance	10 <sup>5</sup>	10 <sup>4</sup>	10 <sup>5</sup> –10 <sup>8</sup>	10 <sup>6</sup> –10 <sup>9</sup>	10 <sup>15</sup>	10 <sup>10</sup>	>10 <sup>5</sup>	>10 <sup>15</sup>	>10 <sup>5</sup>
Suitability for DNN training	No	No	No	No	No	No	Moderate	No	Yes
Suitability for DNN inference	Yes	Yes	Moderate	Yes	No	No	Yes	No	Yes
Suitability for SNN applications	Yes	No	Yes	Yes	Moderate	Yes	Yes	Moderate	Moderate

have been used, such as nitrides,<sup>23</sup> ternary oxides,<sup>24</sup> chalcogenides,<sup>25</sup> or 2D materials.<sup>26,27</sup> Organic materials have been also explored, taking advantage of the low switching energies, wide-range of tunability, and facile ion-migration.<sup>28–30</sup> However, limitations in the writing speed, scaling, and reliability remain open challenges. The forming operation generates a conductive filament (CF) across the switching layer. The CF resistance is changed by electrically induced chemical redox reactions, where the set operation causes the transition to the low-resistance state (LRS), while the reset operation causes the transition to the high-resistance state (HRS). These transitions can occur either by operating the device under the same polarity in unipolar RRAM<sup>31</sup> or by alternating polarities in bipolar

RRAM.<sup>32</sup> Uniform switching RRAM where the resistance can change without any forming operation has also been proposed.<sup>33</sup>

### B. Phase change memory (PCM)

Figure 3(b) schematically shows the phase change memory (PCM), which is based on the ability of specific phase change materials to switch reversibly between the amorphous and the crystalline phases exhibiting different electrical resistivity.<sup>34–36</sup> The phase change material typically consists of chalcogenides, such as Ge<sub>2</sub>Sb<sub>2</sub>Te<sub>5</sub><sup>37</sup> where phase transition can be triggered by the applied voltage pulse via Joule heating. The PCM offers the ability

to store intermediate states by modulating the crystalline fraction within the active material<sup>38</sup> although the stability of the memory state is potentially affected by temperature-dependent retention, caused by the recrystallization of the amorphous region,<sup>39</sup> and drift, caused by the structural relaxation of the amorphous structure.<sup>40</sup> These issues can be handled by materials engineering to improve the high-temperature stability<sup>41</sup> and device engineering to reduce the resistance drift.<sup>42</sup> The PCM technology has also been demonstrated in relatively advanced technology nodes, such as 28<sup>43</sup> and 18 nm.<sup>44</sup> The very high maturity level of development and the higher endurance compared to other non-volatile memory devices<sup>45</sup> make PCM an ideal candidate for in-memory computing.

### C. Ferroelectric random-access memory (FeRAM)

Figure 3(c) schematically shows a ferroelectric random access memory (FeRAM) device based on the ability of a ferroelectric layer to display a remnant electric polarization after the application of voltage pulses.<sup>46</sup> The most typical ferroelectric materials include perovskites with structure  $ABO_3$ , where A and B are cations, e.g.,  $BaTiO_3$  (BTO)<sup>47</sup> and  $PbZr_xTi_{1-x}O_3$  (PZT).<sup>48</sup> Most recently, FeRAM has seen a revival since ferroelectricity was reported in pure and doped hafnium oxides  $HfO_2$  with an orthorhombic structure.<sup>49</sup> While being a CMOS-compatible oxide,  $HfO_2$  has a lower dielectric constant compared to perovskite materials, thus enabling the development of ferroelectric layers with a small thickness between 5 and 30 nm, which is suitable for memory device scaling and 3D integration.<sup>50,51</sup> However, a topic of intense research remains the realization of ferroelectric layer thickness well below 10 nm with good uniformity.<sup>52</sup> FeRAM is probed by measuring the displacement current during ferroelectric switching and thus is a destructive operation that is not always practical for in-memory computing applications. To solve this issue, the ferroelectric tunnel junction (FTJ) has been developed in which the ferroelectric polarization is reflected by the device resistance thanks to bilayer stack device engineering.<sup>53</sup>

### D. Spin-transfer torque magnetic random access memory (STT-MRAM)

Figure 3(d) schematically shows the spin-transfer torque magnetic random access memory (STT-MRAM), consisting of a magnetic tunnel junction (MTJ) composed of a thin insulator sandwiched between two ferromagnetic (FM) layers. In one of the two FM layers, the ferromagnetic polarization is pinned by the presence of adjacent magnetic layers, such as a synthetic antiferromagnetic stack,<sup>54,55</sup> thus acting as a reference for the polarization. The other layer is free and can change its polarization via electrical pulses. The free layer magnetization can thus be programmed by applying a current pulse directly across the MTJ via spin torque.<sup>56,57</sup> Two STT-MRAM states can thus be obtained, namely, a parallel state with relatively low resistance and an antiparallel state with relatively high resistance for equal and opposite directions, respectively, of the magnetic polarization in the pinned and free layers. STT-MRAM features fast switching and good cycling endurance.<sup>58</sup> On the other hand, the resistance window is generally quite limited (less than a factor 2) and multilevel operation is hard to achieve.<sup>59</sup>

### E. Ferroelectric field-effect transistor (FeFET)

In addition to two-terminal FeRAM and FTJ, a three-terminal ferroelectric device has been proposed, namely, the ferroelectric field-effect transistor (FeFET) in Fig. 3(e). The FeFET consists of a field-effect transistor where the gate dielectric is a ferroelectric layer.<sup>60,61</sup> The ferroelectric polarization thus affects the threshold voltage  $V_T$ , which can be used as a monitor of the memory state, similar to a floating-gate memory. Contrary to FeRAM devices, the reading operation of the FeFET device is non-destructive, which is highly favorable for IMC. In addition, FeFET can be integrated in vertical 3D architectures<sup>62</sup> and can display multilevel operation by multilayered stack engineering.<sup>63</sup> An important challenge is the limited cycling endurance of FeFET, which is typically in the range of  $10^5$  cycles, too small for most of applications.

### F. Spin-orbit transfer magnetic random access memory (SOT-MRAM)

Figure 3(f) schematically shows the spin-orbit torque magnetic random access memory (SOT-MRAM). Similar to the STT-MRAM device, SOT-MRAM consists of an MTJ structure deposited on top of a metallic line made of a heavy metal, such as Pt or Ta.<sup>64,65</sup> To program the SOT-MRAM device, a current pulse is applied along the heavy metal line, causing a polarity-dependent accumulation of spin-polarized electrons, thus inducing the magnetization switching in the free layer.<sup>65</sup> The read operation is conducted by probing the MTJ resistance, similar to STT-MRAM. The separation between programming and reading paths allows minimizing the MTJ degradation, thus improving the cycling endurance with respect to STT-MRAM devices. Recently, the integration of SOT-MRAM with the CMOS technology has been demonstrated.<sup>66</sup> Similar to MTJ devices, STT-MRAM suffers from a relatively small resistance window and difficult multilevel operation. Another potential issue is the need for an external magnetic field to support the free-layer switching, which can be overcome by advanced structures with built-in magnetic fields.<sup>67</sup>

### G. Electrochemical random-access memory (ECRAM)

Figure 3(g) schematically shows the electro-chemical random access memory (ECRAM), where the conductivity of a metal-oxide transistor channel can be changed by ionized defects injection across the vertical stack, consisting of a reservoir layer and a solid-state electrolyte layer.<sup>68–70</sup> Defects might consist of oxygen vacancies,<sup>71</sup> Li ions,<sup>72</sup> or protons.<sup>73</sup> Organic materials have also been explored,<sup>74,75</sup> demonstrating various synaptic and neuronal functionalities. Similar to SOT-MRAM, the three-terminal ECRAM structure allows decoupling the read and write paths, thus improving cycling endurance and reducing energy consumption thanks to the extremely low conductivity of the metal oxide channel, e.g.,  $WO_3$ .<sup>69</sup> Controllable and linear potentiation characteristics were reported, which makes ECRAM a promising technology for synaptic devices in neuromorphic devices capable of learning and training.<sup>70</sup> 3D vertical ECRAM has also been demonstrated,<sup>76</sup> paving the way for ECRAM-based high-density cross-point arrays.

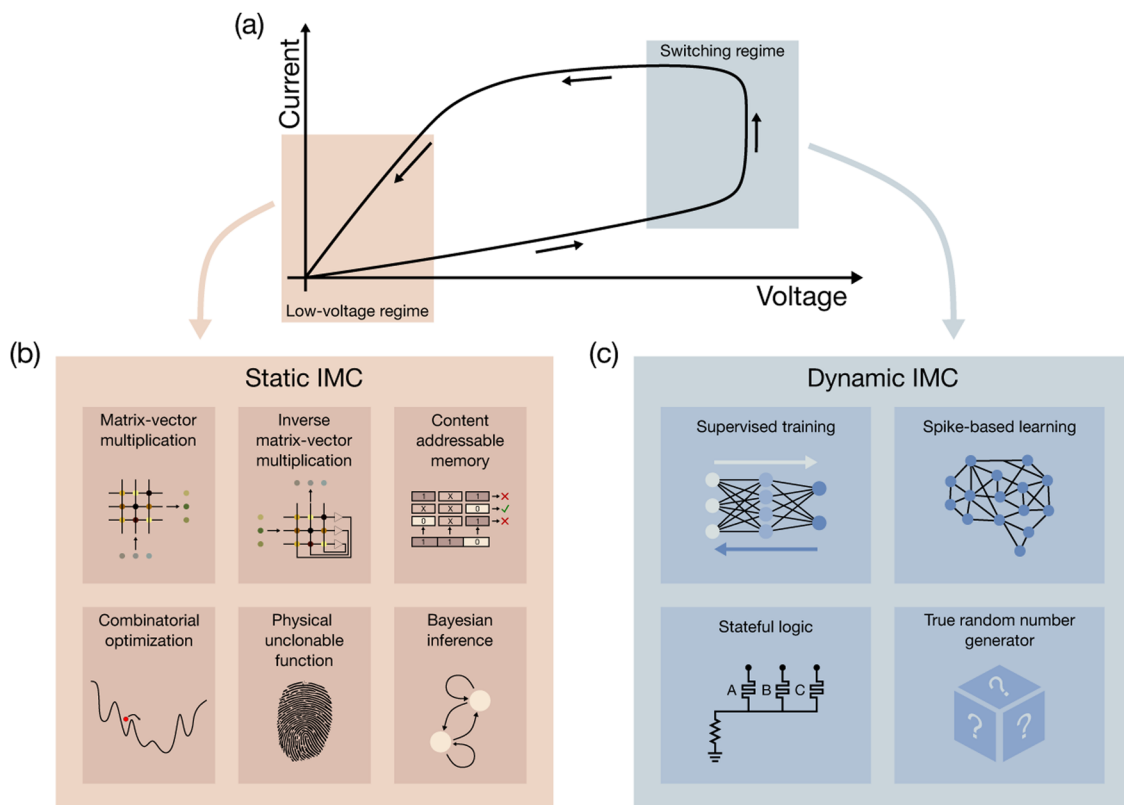
## H. Memtransistor

Memtransistor devices combine the three-terminal transistor structure with the memristor-like ability to change the channel conductance by the application of an in-plane drain–source voltage.<sup>77–79</sup> Typical memtransistors consist of a FET with a 2D semiconductor channel, such as MoS<sub>2</sub>. The memory behavior is obtained by applying large source–drain voltages, which can induce the resistance change by various physical mechanisms, such as field-induced dislocation migration in the polycrystalline MoS<sub>2</sub> channel,<sup>77,78</sup> the dynamic tuning of the Schottky barrier at the metal–semiconductor contact,<sup>80</sup> or the direct cation migration from the electrodes on the surface of a 2D semiconductor.<sup>79,81</sup> Other implementations of memtransistors exploit the optical properties of the 2D material (typically, a transition metal dichalcogenide) to develop devices with neural properties.<sup>82,83</sup> Similar neuromorphic devices were obtained exploiting the ionic diffusion on amorphous oxides, such as ZnO or indium tungsten oxide (IWO).<sup>84–86</sup> The major advantage of the memtransistor is the three-terminal structure, the atomically thin channel, and the 3D integration in the back end. However, compared to all the other reported technologies, memtransistors are still in their early stage of development, with significant challenges on materials, device structures, and reliability.

## III. IN-MEMORY COMPUTING

IMC development has achieved significant progress in the last 10 years, ranging from novel theoretical approaches to experimental IMC hardware demonstrations in silicon-verified test vehicles. The range of applications where IMC can offer improved energy efficiency, performance, and scaling opportunities can be divided into the two macro-categories of *static* and *dynamic* IMC, as shown in Fig. 4(a).

Static IMC, schematically shown in Fig. 4(b), consists of a physical computing concept where the emerging memories are used to store data and perform computation without changing or updating their programmed state.<sup>6</sup> Generally, memory devices in static IMC are first programmed to a desired state to encode pre-trained computing parameters in the form of conductance levels. Random states can also be used in some applications, such as the physical unclonable function (PUF)<sup>87</sup> and reservoir computing (RC) where the stochastic conductance resulting from the fabrication process is directly used in the computation.<sup>88</sup> The programmed memory arrays are then used as physical matrices to execute *in situ* vectorial operations with high parallelism, such as matrix-vector multiplication (MVM).<sup>89</sup> Low voltages are applied to prevent any perturbation



**FIG. 4.** IMC macro-categories and corresponding applications. (a) Schematic current–voltage ( $I$ – $V$ ) curve of an emerging memory device, highlighting the low-voltage and high-voltage/switching regimes, corresponding to static and dynamic IMC, respectively. (b) Examples of static IMC, where the memory stores pre-trained data and executes the computation, e.g., MVM. (c) Examples of dynamic IMC, in which the switching regime allows reproducing dynamic features, such as adaptation and learning.

to the conductive states during computation,<sup>90</sup> thus resulting in a low power consumption, which is attractive for decentralized computing architectures, such as edge<sup>91</sup> and fog<sup>92</sup> computing. The high degree of parallelism allows reducing the number of operations needed to carry out a given task, thus achieving  $\mathcal{O}(1)$  computational complexity.<sup>93,94</sup> Examples of static IMC include matrix-vector-multiplication (MVM, Sec. IV), inverse-matrix-vector multiplication (IMVM, Sec. V), and content-addressable memories (CAMs, Sec. VI).

Dynamic IMC, schematically shown in Fig. 4(c), generally combines all the opportunities of static IMC with the additional strength of enabling controlled switching of the memory devices to reproduce additional functions, such as neuron activation,<sup>95</sup> stateful Boolean logic,<sup>96,97</sup> and learning in supervised/unsupervised neural networks.<sup>98–101</sup> A wide range of physical mechanisms can be used for the controlled switching, such as filament plasticity in RRAM devices,<sup>102</sup> gradual crystallization in PCM devices,<sup>95</sup> charge trapping in MoS<sub>2</sub> memtransistors,<sup>103</sup> and magnetic polarization for true-random number generation (TRNG).<sup>104</sup> Dynamic IMC provides a promising avenue for reducing latency, energy, and circuit area by leveraging the intrinsic device physics of the device instead of emulating the desired characteristics via the analog/digital design of CMOS-based networks.<sup>105</sup> Dynamic and static IMC are generally combined in the same platform to provide energy-efficient computing systems capable of learning and adaptation.<sup>95,106</sup> Applications of dynamic IMC include outer product accelerators for neural network training (Sec. VII) and neuromorphic systems for brain-inspired computing (Sec. VIII).

#### IV. MATRIX-VECTOR MULTIPLICATION

##### A. Concepts and implementation

MVM can be executed in a crosspoint memory array by universal circuit laws, such as Kirchoff's current law for summation and Ohm's law for multiplication.<sup>7,107</sup> The crosspoint array consists of a matrix of programmable memory elements whose top and bottom electrodes are, respectively, tied to common columns and rows, as shown in Fig. 5(a).<sup>108,109</sup> According to the IMC concept, the crosspoint array acts as a physical matrix mapping computational

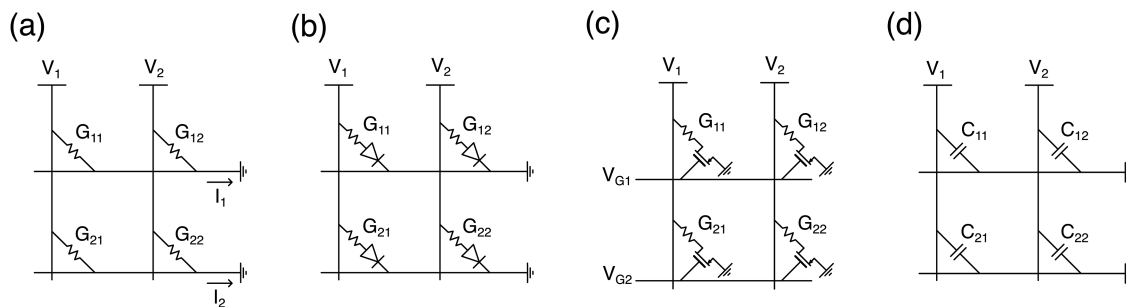
parameters, e.g., synaptic weights in a neural network, to compute the MVM physically in the analog domain. This is schematically shown in Fig. 5(a), where the application of a voltage  $V_j$  at the  $j$ th column results in a current at the  $i$ th row, connected to ground, given by

$$I_i = \sum_j^N G_{ij} \cdot V_j, \tag{1}$$

where  $G_{i,j}$  is the conductance of the memory element at position  $i, j$  and  $N$  is the number of rows and columns.<sup>7,107</sup> Equation (1) can be written in the compact matrix form  $\mathbf{i} = \mathbf{G}\mathbf{v}$ , thus evidencing the multiplication of the conductance matrix  $\mathbf{G}$  with the voltage vector  $\mathbf{v}$ .

The MVM operation of Fig. 5(a) is carried out without moving the matrix parameters, in line with *in situ* processing paradigm of IMC. In addition, the operation is performed in just one step, thus minimizing the latency and maximizing the throughput thanks to a computational complexity of  $\mathcal{O}(1)$ . Such a massive parallelism of MVM allows for achieving outstanding area and energy efficiency, compared to traditional digital multiply-and-accumulate (MAC) operations. Finally, the crosspoint array is generally integrated in the back end of the line (BEOL) of the CMOS process, thus taking advantage of 3D stacking and of a small cell area of only  $4F^2/N$ , where  $F$  is the lithographic feature size and  $N$  is the number stacked layers.<sup>110</sup> Despite the advantages of parallelism, density, and latency, the MVM concept is an analog computing process that is critically sensitive to device variability,<sup>111,112</sup> noise,<sup>113</sup> drift of conductance,<sup>40</sup> and parasitic IR drop along wires,<sup>114</sup> all affecting the accuracy of computation. To deal with these parasitic effects, several mitigation and compensation techniques have been proposed at device,<sup>115</sup> algorithm,<sup>114,116–120</sup> and architectural levels.<sup>121,122</sup>

The MVM concept can be extended to virtually all types of memory devices and cell structures in the array. The one-resistance (1R) structure of Fig. 5(a) is affected by crosstalk and sneak path issues during programming and reading.<sup>123</sup> These issues can be prevented by adding a selector device in series to the memory element, resulting in the one-selector/one-resistor (1S1R) structure<sup>124–126</sup> or the one-transistor/one-resistor (1T1R) structure,<sup>127–129</sup> illustrated in Figs. 5(b) and 5(c), respectively. The 1S1R configuration avoids



**FIG. 5.** Various cell structures for crosspoint array circuits. (a) One-resistor (1R) structure where the cell consists of a passive resistive device. (b) One-selector/one-resistor (1S1R) structure where the sneak path problem is circumvented by a non-linear selector device without affecting the integration density. (c) One-transistor/one-resistor (1T1R) structure allows for the selection of individual cells during programming and reading at the cost of a lower integration density. (d) One-capacitor (1C) structure, which prevents static leakage during MVM.

sneak path currents during the programming phase by introducing a highly non-linear two-terminal device<sup>109,130,131</sup> that suppresses the current of unselected and half-selected cells in the array while maintaining the small  $4F^2$  area of the 1R cell structure.<sup>109</sup> The 1T1R structure ensures tight control of the programming current while allowing sophisticated program/verify algorithms<sup>132</sup> at the cost of a larger cell area and a higher complexity introduced by the third terminal. In addition to resistive memory cells, where the computation parameter is stored in the conductance, capacitive memories can be adopted with the one capacitance (1C) structure in Fig. 5(d). Here, the small-signal capacitance can be tuned<sup>133</sup> and used in MVM operations via the charge-voltage capacitor law  $Q = CV$ .

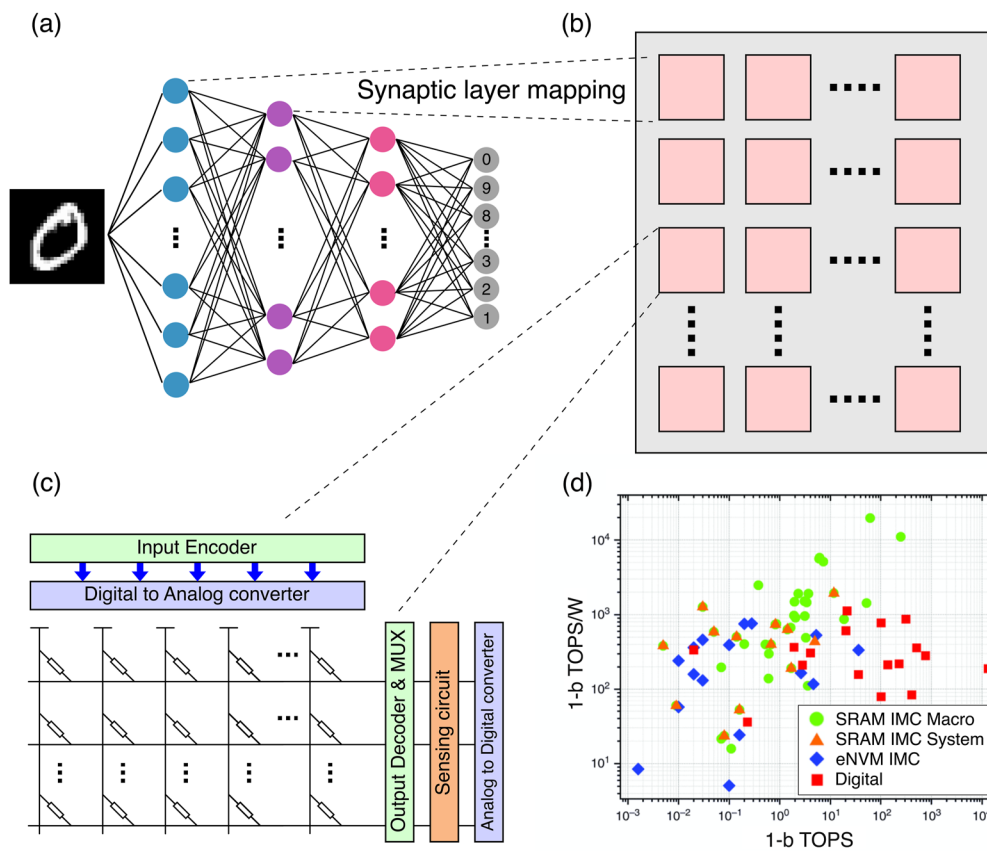
From the computational viewpoint, MVM requires the input vector to be encoded in voltage amplitudes, usually by means of a digital-to-analog converter (DAC). The output analog current can be sensed by using a transimpedance amplifier (TIA)<sup>134,135</sup> and then converted by using an analog-to-digital converter (ADC) for further processing in the digital domain. Alternatively, the input vector can be encoded as the time duration  $t_j$  by pulse-width modulation (PWM).<sup>136</sup> This approach is typically implemented in the 1T1R

array, where the time-encoded signal can be applied to the transistors gates, while a constant read voltage  $V_{read}$  is applied across the cells. PWM requires that the analog current at each row is integrated to yield the charge  $Q_i$  according to

$$Q_i = \sum_j^N V_{read} G_{i,j} \cdot t_j, \tag{2}$$

thus providing an alternate MVM operation yielding vector  $\mathbf{q} = V_{read} \mathbf{Gt}$  similar to Eq. (1).

Note that, while MVM is strongly accelerated thanks to the array parallelism, memory programming might require a relatively long time, especially when a high equivalent-bit precision is needed. However, the programming time can be generally amortized for applications where the computational parameters remain fixed for most of the MVM operations. This is the case for discrete cosine transform (DCT) for extracting frequency components from a data sequence.<sup>137</sup> DCT is routinely applied for image compression, thus providing an ideal application for IMC.<sup>134</sup>



**FIG. 6.** MVM for neural network accelerators. (a) Sketch of a fully connected DNN for image classification. (b) Multi-core architecture where each tile performs MVM between activation and synaptic weights. (c) Individual core consisting of a crosspoint array with peripheral circuits for input/output communication and conversion. (d) Correlation plot of energy efficiency as a function of throughput for different hardware accelerators of DNN inference, including eNVM-based, SRAM-based IMC, and a fully digital approach. Reproduced with permission from Seo *et al.*, IEEE Solid-State Circuits Mag. **14**(3), 65–79 (2022). Copyright 2022 IEEE.



## B. Application to neural network inference

Another application where computational parameters remain constant throughout computation is the forward propagation during the inference phase in a deep neural network (DNN).<sup>138,139</sup>

Figure 6(a) shows a sketch of a fully connected neural network (FCNN) for image classification with three synaptic layers. Each synaptic layer can be viewed as a MVM where synaptic weights are mapped in the conductance matrix, while activations are used as the input vector. The inference operation can thus be mapped in several MVMs occurring in distinct crosspoint arrays, each mapping a different synaptic layer or a region of the DNN. Figure 6(b) shows a possible multi-core IMC architecture where each computational unit performs the assigned computation independently, as illustrated in Fig. 6(c), while a logic unit collects output data from the cores and submits activation signals to them. Given the sequential operation of DNN inference, the architecture and computational cores can be optimized to maximize the data throughput.

Inference accelerators have been proposed with a variety of implementations, differing by the adopted memory technologies;<sup>98,127,140</sup> the number of quantized levels of input, weight, and output;<sup>141,142</sup> the peripheral circuits;<sup>136,143</sup> the amount of possible reconfiguration;<sup>143</sup> and the possibility of implementing back-propagation training in addition to forward-propagation inference.<sup>99,144</sup> Similar to FCNN layers, IMC has been shown to accelerate convolutional layers<sup>99,127</sup> and recurrent neural networks<sup>145</sup> by changing the MVM partition and computation technique.<sup>146</sup>

IMC can largely improve the energy efficiency and the throughput of MVM for DNN inference. Figure 6(d) shows the power efficiency and throughput of the state-of-the-art IMC accelerators based on nonvolatile memories compared to IMC based on static random access memory (SRAM) or fully digital accelerators.<sup>147</sup> SRAMs feature faster access time and better robustness to variability and disturbs thanks to their digital nature and fully silicon-based CMOS technology. However, SRAM has a larger cell area due to the 6T or 8T bit-cell structure, cannot implement multilevel operations, and cannot provide nonvolatile storage, thus requiring the upload of computational parameters at the power-on phase. The latter issue is

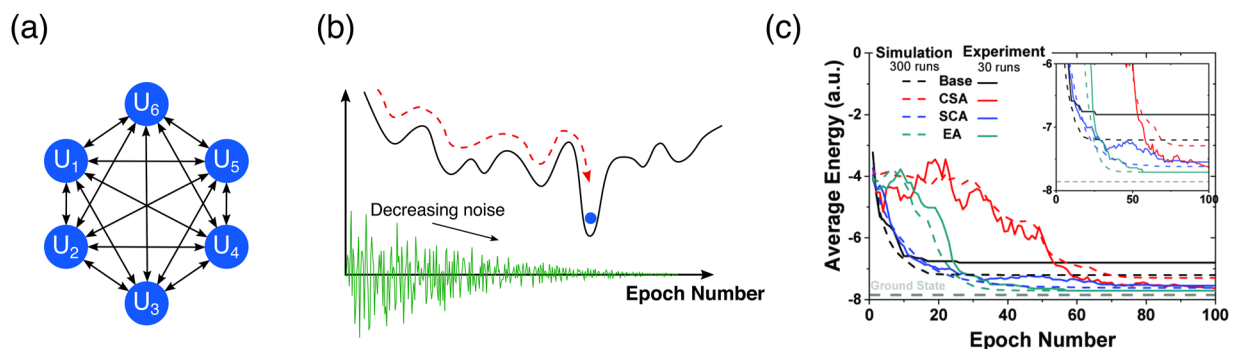
a significant drawback in applications where the neural accelerator frequently switches between stand-by and computing phase, which is typical in low-power edge-computing applications.

## C. Application to combinatorial optimization

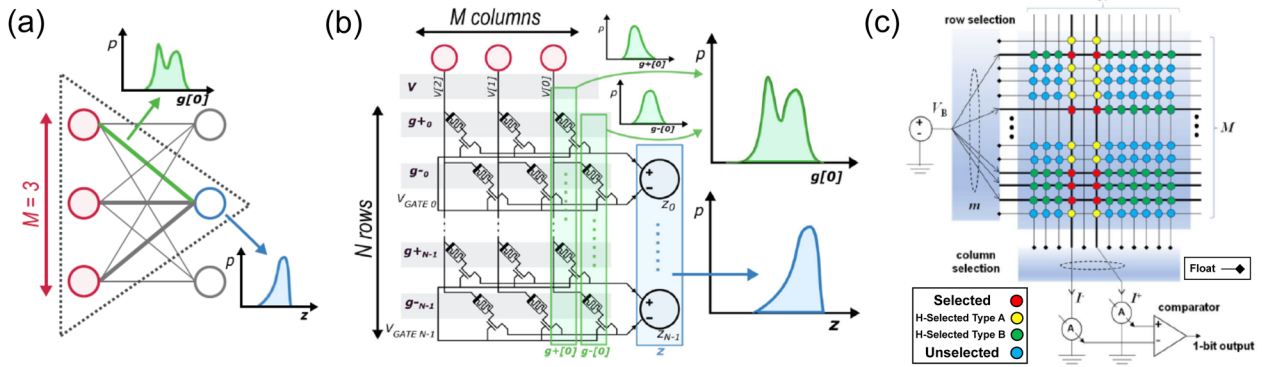
MVM represents the core operation of combinatorial optimization tasks.<sup>148</sup> Here, emerging memories can provide both the MVM operation via the crosspoint array and the stochastic physical noise, which is generally needed to navigate among the local minima of the cost function. Indeed, metaheuristic optimization techniques, such as chaotic simulated annealing or stochastic simulated annealing, require massive MVM and tunable sources of noise. These computing strategies typically rely on recurrent stochastic networks, such as the Hopfield neural network, sketched in Fig. 7(a),<sup>95,106,149</sup> or restricted Boltzmann machine (RBM).<sup>150–152</sup> In these approaches, the network is characterized by a certain energy (or cost) function  $E$  that depends on the state of the neurons, which in turn depends on the synaptic spike stimulations and the injected noise. By properly tuning the injected noise, it is possible to control the ability of the neurons to escape from local minima of  $E$ , as depicted in Fig. 7(b). By gradually decreasing the injected noise, the search takes the shape of a simulated annealing algorithm, where the effective temperature is slowly decreased in analogy with the cooling phase of physical annealing. This is shown in Fig. 7(c), where the network manages to find thermal equilibrium at the global minimum of  $E$ , thus solving the optimization task.<sup>145</sup> This approach finds application in several key workloads in logistics, scheduling, and other NP-hard problems, such as the traveling salesperson problem.

## D. Application to stochastic computing and security

Programming variability is a major issue in deterministic DNNs by affecting the weight precision, hence the accuracy of inference. On the other hand, programming variation can provide a source of stochasticity for specific computing applications, such as stochastic computing and hardware security. For instance, Bayesian inference relies on neural networks where the model parameters are probability distributions. In this scenario, transferring the



**FIG. 7.** MVM for combinatorial optimization. (a) Sketch of a Hopfield-type recurrent neural network, characterized by a system energy  $E$ . (b) System energy  $E$  and iterative search of the global minimum, representing the optimal solution of the combinatorial task. The decreasing noise allows for reaching the global minimum by escaping local minima. (c) Evolution of the average energy of a RRAM-based Hopfield RNN for various optimization strategies. Reprinted with permission from Mahmoodi *et al.*, 2019 *International Electron Devices Meeting (IEDM)* (IEEE, 2019), pp. 14.7.1–14.7.4. Copyright 2019 IEEE.

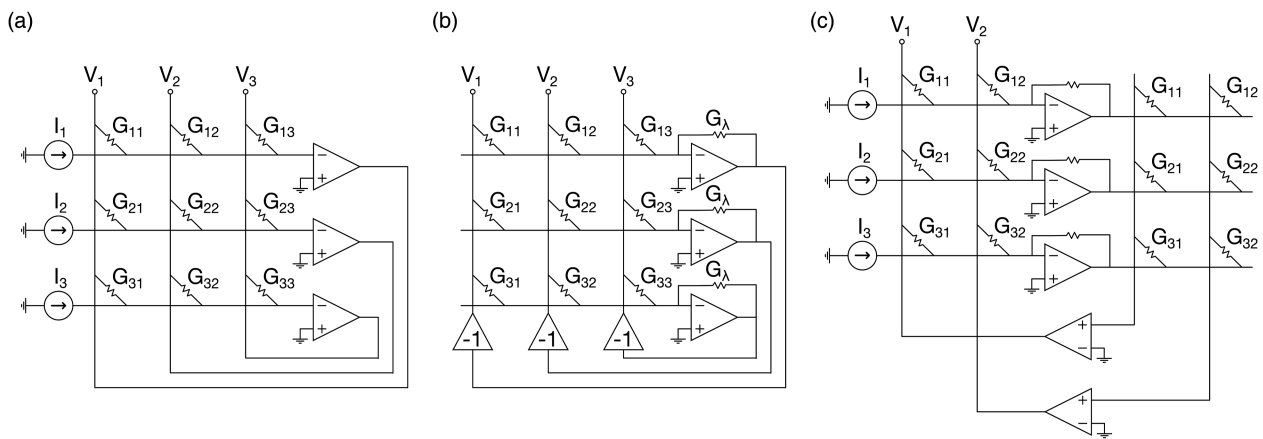


**FIG. 8.** MVM for stochastic computing. (a) Sketch of a Bayesian neural network where synapses and neurons are represented by probability distributions. Reproduced with permission from Dalgaty *et al.*, *Adv. Intell. Syst.* **3**(8), 2000103 (2021). Copyright 2021 Author(s), licensed under a Creative Commons Attribution 4.0 License. (b) RRAM-based realization of the Bayesian neural network, where each column describes the distribution of a synaptic parameter. Reproduced with permission from Dalgaty *et al.*, *Adv. Intell. Syst.* **3**(8), 2000103 (2021). Copyright 2021 Author(s), licensed under a Creative Commons Attribution 4.0 License. (c) NVM-based PUF circuit based on a passive crosspoint array of stochastic memory devices for the generation of a response as the input of a submitted challenge. Reproduced with permission from M. R. Mahmoodi, D. B. Strukov, and O. Kavehei, *IEEE Trans. Electron Devices* **66**(12), 5050–5059 (2019). Copyright 2019 IEEE.

ex-situ trained model to the hardware network is less critical since a probability distribution can be naturally modeled by the physical distribution of conductance states.<sup>153</sup> Figure 8(a) shows the conceptual scheme of an RRAM-based Bayesian network where each synaptic weight belongs to a certain distribution. Figure 8(b) shows a possible implementation in an  $N \times M$  array of RRAM synapses with 1T1R structures.<sup>153</sup> Here, the distribution of a synaptic parameter is modeled by the distribution of conductance states of  $N$  devices in a column, while the input voltages to each column are the outputs generated by  $M$  neurons in the previous layer. By applying a voltage vector across  $M$  columns, each row yields a current that flows into a neuron circuit, resulting in a distribution of  $N$  neuron activation voltages, namely, the output distribution of the neuron. Based on

the same approach, Monte Carlo Markov chain (MCMC) networks have been demonstrated with stochastic RRAM arrays.<sup>154</sup>

The stochastic properties of emerging memories can also provide the foundation for developing novel security primitive circuits.<sup>104</sup> Figure 8(c) shows the conceptual idea for implementing a memory-based physical unclonable function (PUF) for chip authentication.<sup>87</sup> An input challenge encodes the information to select specific rows and columns of the crosspoint memory array, thus generating a single-bit unique response by current comparison. A 1R crosspoint array is adopted to take advantage of circulating sneak path currents, enabling the participation and interaction of all memory devices in the array, thus increasing the complexity of the solution and robustness to external attacks.<sup>87</sup>



**FIG. 9.** Closed-loop IMC circuits for IMVM. (a) Circuit for the solution of linear systems<sup>155</sup> of the form  $\mathbf{Ax} = \mathbf{b}$ . (b) Circuit for the eigenvector computation, i.e., for the solution of the secular equation<sup>156</sup>  $\mathbf{Ax} = \lambda\mathbf{x}$ . (c) Pseudoinverse matrix computing circuit for the solution of the linear regression problems<sup>157,158</sup> of the form  $\mathbf{X}\beta + \epsilon = \mathbf{y}$ .

## V. INVERSE MATRIX-VECTOR MULTIPLICATION

Crosspoint memory arrays with closed-loop circuit topology can accelerate inverse-matrix vector multiplication (IMVM), such as linear system solution, matrix inversion, and linear/regularized regression.<sup>155,157,158</sup> Figure 9(a) shows a typical IMVM circuit for the solution of a linear system, where the array is complemented with an array of operational amplifiers (OAs). In this circuit, currents are provided as row input, while the voltages that satisfy Eq. (1) are automatically established by the OAs via the closed-loop feedback connection, thus allowing for the solution for the set of linear equations by

$$\mathbf{v} = -\mathbf{G}^{-1}\mathbf{i} \quad (3)$$

Similar to open-loop MVM of Sec. IV, closed-loop IMC (CL-IMC) can achieve the  $\mathcal{O}(1)$  solution of algebra problems with polynomial complexity  $\mathcal{O}(n^\alpha)$ , where  $n$  is the number of linear equations and  $\alpha$  is between 2 and 3.<sup>156</sup> CL-IMC appears thus as one of the most promising candidates for accelerating complex linear algebra tasks via IMC.

Figure 10(a) shows the experimental output of a hardware implementation of the circuit in Fig. 9(a) to yield the elements of a  $3 \times 3$  inverse matrix  $\mathbf{A}^{-1}$  as a function of the analytical solution.<sup>155</sup> In-memory matrix inversion might find application in a number of machine learning tasks, such as Markov chain<sup>159</sup> and numerical solution of differential equations.<sup>155</sup> With errors as low as 3%, feedback-based crossbar circuits can provide a viable alternative to bulky digital processors for linear system solution tasks, serving as a potential cornerstone of IMC-based analog processing units.

### A. Application to ranking algorithms

The CL-IMC prototype topology of Fig. 9(a) can be extended to eigenvector computation by the circuit of Fig. 9(b).<sup>155</sup> Here, the output is directly fed as input after sign inversion, thus resulting in a self-sustaining architecture. OAs are used in the transimpedance amplifier (TIA) configuration, where the feedback conductance  $G_\lambda$

encodes the principal matrix eigenvalue  $\lambda$ . Kirchhoff's law at the virtual ground nodes thus reads

$$\mathbf{G}\mathbf{v} = G_\lambda\mathbf{v}, \quad (4)$$

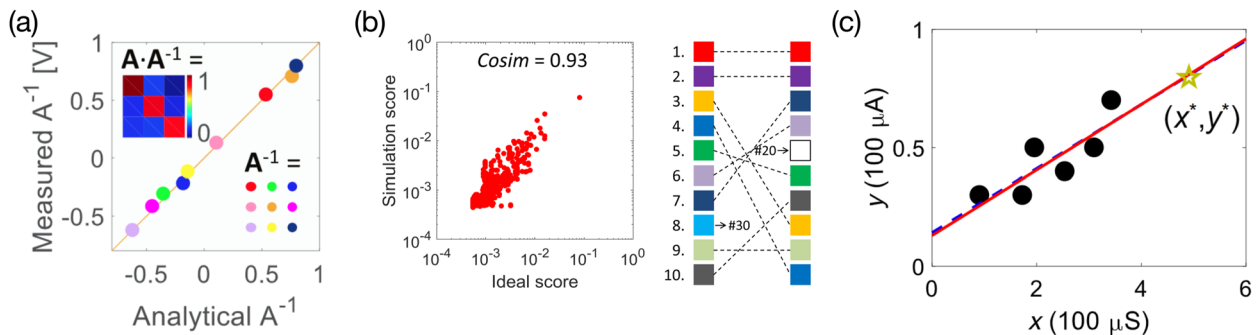
which electrically matches the secular equation  $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ . For negative  $\lambda$ , the analog inversion buffers are removed and the absolute value  $|\lambda|$  is encoded as the conductance  $G_\lambda$ . Differently from the linear system solver in Fig. 9(a), the eigenvector circuit operates in a positive feedback regime, thus allowing for self-sustaining operation. Due to the positive feedback, only the eigenvectors of the largest positive and negative eigenvalues can be solved. In addition,  $G_\lambda$  should slightly deviate from the ideal  $\lambda$  to initiate the self-sustained dynamic response.<sup>160</sup> Figure 10(b) shows the results of a website ranking task according to Google's PageRank algorithm, which is a typical application of eigenvector computation,<sup>156</sup> together with similar ranking algorithms.<sup>161</sup> It has been estimated that the solution of PageRank with CL-IMC can provide up to  $100\times$  throughput improvement with respect to a digital computer.<sup>156</sup>

### B. Application to data regression

The CL-IMC concept can be further extended to non-square matrices as in the computation of the Moore-Penrose inverse or pseudoinverse.<sup>156</sup> Figure 9(c) shows the CL-IMC circuit for matrix pseudoinverse computation or linear regression. The circuit features two  $m \times n$  crosspoint memory arrays, each encoding a given matrix dataset, and two OA arrays. A simple analysis shows that, by injecting the input current at the virtual grounds of the first  $m$  OAs, the output voltages at the second array of OAs are given by

$$\mathbf{v} = -(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{i} = -\mathbf{G}^+\mathbf{i} \quad (5)$$

Figure 10(c) shows experimental results for a two-dimensional linear regression problem on a relatively small scale.<sup>156</sup> Note that this circuit also eliminates the stability constraints of the linear system solver in Fig. 9(a),<sup>90,158</sup> which is limited to positive-definite matrices only as a requirement for ensuring poles to lie in the left-half-plane. Furthermore, by using a matrix  $\mathbf{F}$  instead of simple local-feedback



**FIG. 10.** Results of closed-loop IMC for IMVM problems. (a) Correlation plot of the experimental results of the inversion of a  $3 \times 3$  matrix as a function of ideal analytical results. Reproduced with permission from Sun *et al.*, Proc. Natl. Acad. Sci. U. S. A. **116**(10), 4123–4128 (2019). Copyright 2019 National Academy of Sciences. (b) Correlation plot of the circuit output for a PageRank algorithm of the Harvard 500 dataset as a function of the ideal analytical results. Reproduced with permission from Sun *et al.*, IEEE Trans. Electron Devices **67**(4), 1466–1470 (2020). Copyright 2020 IEEE. (c) Experimental demonstration of linear regression on RRAM devices. Reproduced with permission from Sun *et al.*, Sci. Adv. **6**(5), eaay2378 (2020). Copyright 2020 Author(s), licensed under a Creative Commons Attribution 4.0 License.

conductance for the first  $m$  OAs, the same circuit can execute a generalized regression according to

$$\mathbf{v} = -(\mathbf{G}^T \mathbf{F}^{-1} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{F}^{-1} \mathbf{i} = -\mathbf{G}_F^+ \mathbf{i}, \quad (6)$$

where  $\mathbf{F}$  is a generalization matrix for the given dataset.<sup>158</sup> Among the applications of the Moore–Penrose inverse are linear/logistic regression and prediction, which play an important role in data analytics and machine learning.<sup>157</sup>

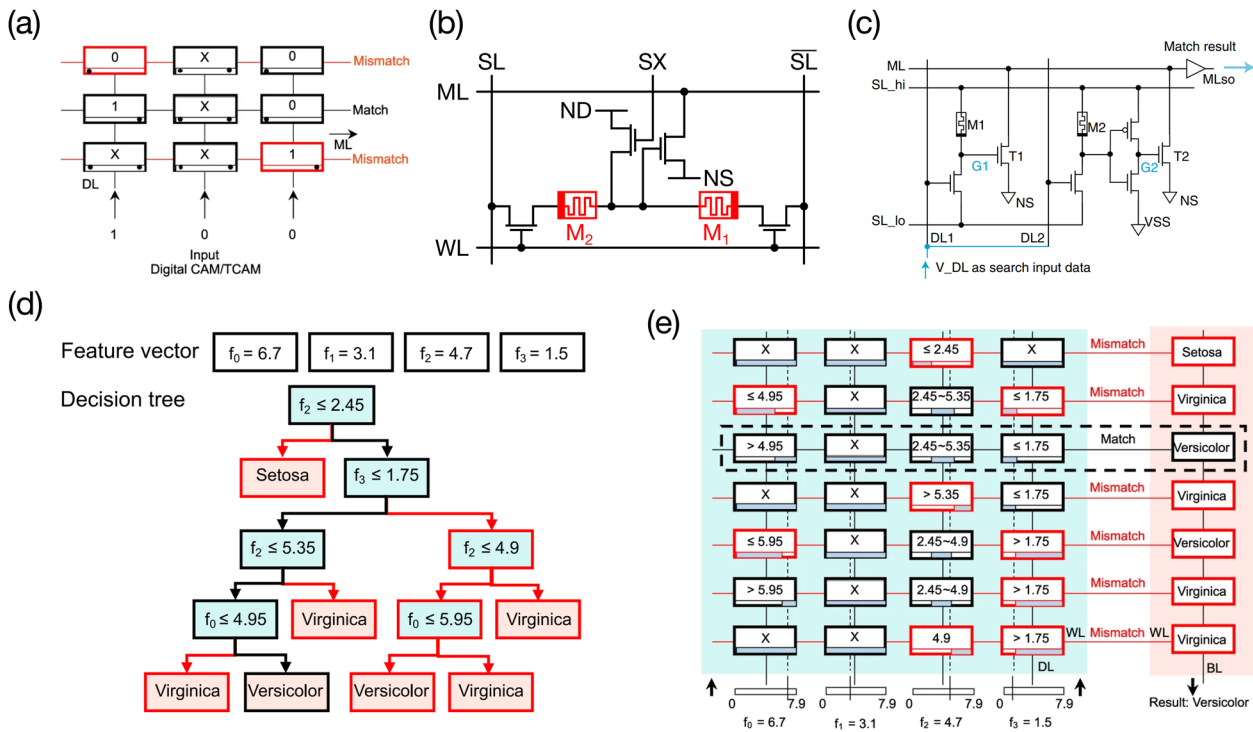
C. Discussion

CL-IMC allows for the acceleration of several IMVM operations with reduced complexity, which is attractive for large-scale general-purpose machine learning accelerators. On the other hand, CL-IMC also faces considerable challenges, such as the reduced precision with respect to a floating-point computers, owing to the increased sensitivity of the analog domain.<sup>159</sup> Circuit non-ideality affecting the computing accuracy includes the parasitic interconnect resistances,<sup>114</sup> electronic noise from circuit components,<sup>90</sup> and conductance variations.<sup>158</sup> The effect of non-ideality can be mitigated by compensation schemes, array tiling, signal range

increase, and fine-tuned programming algorithms, thus resulting in a complex trade-off with the overall throughput, area, and energy consumption.<sup>90,122,162</sup> On the other hand, error-tolerant applications, such as massive multiple-input/multiple-output (MIMO) decoding in 6G networks, allow for better robustness to circuit non-ideality.<sup>163</sup> Finally, the medium-precision solution obtained by analog IMC might be used as a seed for high-precision digital solvers,<sup>164</sup> allowing for orders-of-magnitude improvements in energy consumption and execution time.

VI. COMPUTING WITH CONTENT ADDRESSABLE MEMORY

The content-addressable memory (CAM) is a specialized memory structure where stored data are accessed by inputting the desired data content and extracting their address as the output, which is the opposite compared to conventional memories.<sup>165</sup> Figure 11(a) shows a schematic structure of a typical ternary content addressable memory (TCAM), where the third option *don't care* or “X” is available in addition to binary 0 and 1 values in the memory array. Here, an input pattern presented to the CAM from data lines (DLs)



**FIG. 11.** Content-addressable memory based on emerging memories. (a) Schematic of a digital TCAM, where binary data are matched against patterns stored in a ternary array. Reproduced with permission from Pedretti *et al.*, Nat. Commun. **12**(1), 5806 (2021). Copyright 2015 Author(s), licensed under a Creative Commons Attribution 4.0 License. (b) RRAM-based TCAM cell, where memory devices  $M_1, M_2$  store the ternary value as a suitable combination of *HRS* and *LRS* states. (c) Memristor-based analog CAM cell, where the analog input pattern is encoded as the voltage amplitude on the Data Line (DL). Reproduced with permission from Li *et al.*, Nat. Commun. **11**(1), 1638 (2020). Copyright 2015 Author(s), licensed under a Creative Commons Attribution 4.0 License. (d) Decision tree for the Iris dataset classification. Reproduced with permission from Pedretti *et al.*, Nat. Commun. **12**(1), 5806 (2021). Copyright 2015 Author(s), licensed under a Creative Commons Attribution 4.0 License. (e) Analog-CAM implementation of the decision tree in (d), where each root-to-leaf path corresponds to a row of the memory array. Reproduced with permission from Pedretti *et al.*, Nat. Commun. **12**(1), 5806 (2021). Copyright 2015 Author(s), licensed under a Creative Commons Attribution 4.0 License.

is compared with the stored data and the corresponding match line (ML) is asserted if a match is found. Due to its inherently high parallelism, CAM/TCAM is naturally suited to accelerate pattern matching,<sup>166,167</sup> branch prediction,<sup>168</sup> and lookup operations<sup>169</sup> *in situ* within the memory, thus minimizing data movement.

TCAM parallelism comes at the expense of relatively large area and power consumption as every memory cell must be equipped with a dedicated comparison circuit. When implemented using SRAM memories, a single CAM cell may use up to 16 transistors,<sup>165</sup> thus adding significant area, latency, and power overhead for the search operation and preventing large-scale integration. By replacing conventional SRAM with emerging memories, leakage power can be reduced and cell density can be improved. Figure 11(b) shows a differential RRAM-based CAM cell, where memory devices  $M_1$  and  $M_2$  are programmed to either state *LRS/HRS* or *HRS/LRS* to reproduce values “1” or “0,” respectively.<sup>167</sup> State “X” is instead obtained by programming both RRAM devices to either *HRS* or *LRS*. Depending on the relative ratio of the two conductances (stored data) and the voltage at the wordline (WL) (input data), the matchline ML is either asserted low or left high, thus realizing CAM operation. RRAM-based TCAMs were shown to accelerate regular expression matching and genomic sequencing with up to 25× improvement in energy efficiency.<sup>167</sup>

The analog tunability of emerging memories allows for realizing analog CAMs capable of analog pattern matching with stored data. Figure 11(c) shows an analog CAM cell<sup>170</sup> where value intervals, rather than binary values, can be stored and compared with analog input patterns. In this case, the match line is asserted when all values of the input pattern fall within the ranges stored in the corresponding row of the memory array. Analog memory-based CAMs are naturally suited to accelerate more-than-binary tree-based algorithms, which represent the foundation of many machine learning tasks. Figure 11(d) shows a proposed implementation<sup>171</sup> of tree-based inference applied to the classification of the Iris dataset. By mapping each root-to-leaf path into a corresponding row of the memory array, input data can be instantly

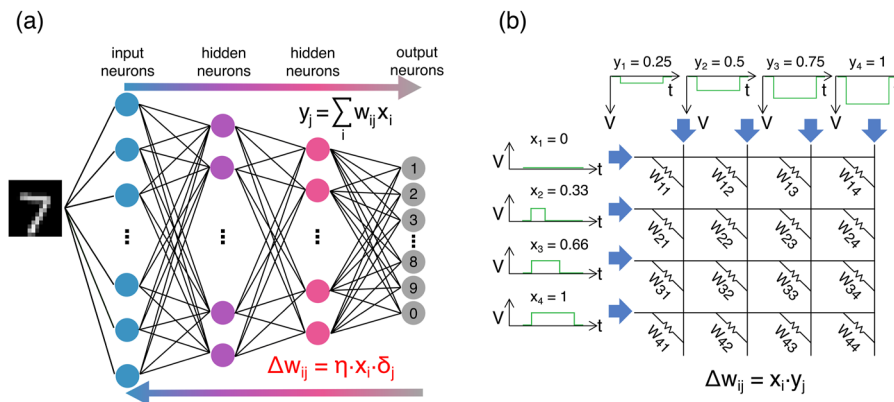
classified by coupling the analog CAM to a label array, as shown in Fig. 11(e), with a  $\times 10^3$  throughput improvement with respect to digital implementations.<sup>171</sup>

### VII. ONLINE TRAINING BY IN-MEMORY OUTER PRODUCT

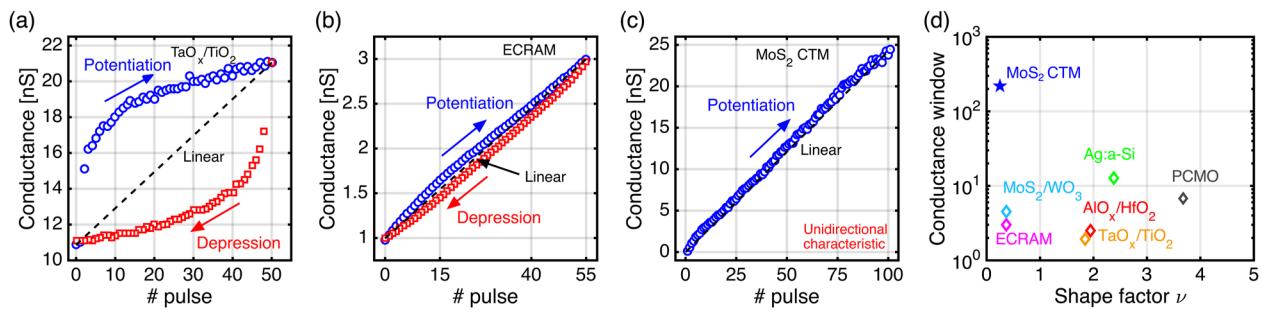
While MVM can efficiently accelerate forward propagation for DNN inference, it only partially supports the execution of the training process. In fact, DNN training is by far the most energy- and time-consuming operation in a DNN.<sup>172</sup> The most typical training methodology relies on the gradient-descent algorithm, such as the backpropagation approach, which requires a multiple synaptic weight update and data transferring.<sup>138</sup> DNN training requires several days/weeks of iterations in multicore supercomputers, such as the graphical processing unit (GPU) or the tensor processing unit (TPU), to update billions of synaptic parameters in the network. This is mainly because all data and synaptic parameters must be transferred between the memory and the processing unit, which results in a major memory bottleneck. Figure 12(a) shows a DNN with the typical training approach, including (i) forward propagation of data for generating an output neuron, (ii) calculation of the error  $\delta_j$  between the  $j$ th neuron current output and the ideal output also known as the *label*, and (iii) backpropagation of the error for the weight update according to

$$\Delta w_{ij} = \eta x_i \delta_j, \tag{7}$$

where  $\Delta w_{ij}$  is the weight update,  $\eta$  is the learning rate, and  $x_i$  is the input of the pre-synaptic neuron. The operation in Eq. (7) is an outer product, where the input vectors  $\mathbf{x}$  and  $\delta$  generate a matrix of weight update  $\Delta \mathbf{W}$  to be applied to the whole synaptic layer. The vector–vector outer product  $\Delta \mathbf{W} = \mathbf{x} \otimes \mathbf{y}$  can be accelerated within the crosspoint array, as shown in Fig. 12(b), where  $\mathbf{x}$  is mapped as the pulse-width of the row voltage pulses, while  $\mathbf{y}$  is mapped as the amplitude of the column voltage pulses.<sup>173</sup>



**FIG. 12.** IMC training by an outer product. (a) Schematic representation of an artificial neural network, where backpropagation training relies on the weight update according to an outer product of the error  $\delta_j$  and the signal  $x_i$ . (b) Crosspoint implementation of the outer product. The weight  $w_{ij}$  is updated by a value  $\Delta w_{ij} = x_i \cdot y_j$ . The multiplicative effect is obtained by encoding  $x_i$  as the pulse width of the row voltage pulse and  $y_j$  as the amplitude of the column voltage pulse. From Agarwal *et al.*, 2016 *International Joint Conference on Neural Networks (IJCNN)*. Copyright 2016 IEEE. Reproduced with permission from IEEE.



**FIG. 13.** Experimental weight-update characteristics by pulses of equal amplitude and pulse-width for potentiation and depression. (a) Update characteristics of  $\text{TaO}_x/\text{TiO}_2$  RRAM. Reprinted with permission from Yu *et al.*, 2015 *IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2015), pp. 17.3.1–17.3.4. Copyright 2015 IEEE. (b) Update characteristics of Li-based ECRAM. Reprinted with permission from Tang *et al.*, 2018 *IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2018), pp. 13.1.1–13.1.4. Copyright 2018 IEEE. (c) Update characteristics of  $\text{MoS}_2$ -based CTM. Reprinted with permission from Farronato *et al.*, 2022 *IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (IEEE, 2022), pp. 1–4. Copyright 2022 IEEE. (d) Correlation plot of the non-linearity factor  $\nu$  and normalized conductance window  $(G_{\max} - G_{\min})/G_{\min}$  for various synaptic devices. Reprinted with permission from Farronato *et al.*, 2022 *IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (IEEE, 2022), pp. 1–4. Copyright 2022 IEEE.

The key requirement for the in-memory outer product of Fig. 12(b) is the linearity of the conductance change with both pulse voltage and time or at least one of the two. Conductance update can be physically obtained by potentiation or depression of the memory conductance by applying suitable pulses to the devices. The linear update must be obtained by an open-loop operation, where the same conductance change is achieved at a given voltage and pulse-width, irrespective of the initial state. Unfortunately, potentiation and depression of emerging memories are generally non-linear with applied voltage as a result of the exponential time–voltage relationship of ion migration, tunneling, and other fundamental physical processes of set/reset.<sup>128</sup>

To support the linearity of potentiation/depression with time, Fig. 13 shows measured conductance update characteristics for emerging memory devices. The RRAM device in Fig. 13(a) displays a non-linear increase with the number of pulses, or equivalently time, with an initially steep change followed by a saturation regime.<sup>15,174</sup> Figure 13(b) shows the weight update characteristics for an ECRAM device, where an improved linearity can be seen thanks to the three-terminal structure separating the read and program paths.<sup>69</sup> Figure 13(c) shows the potentiation characteristic for a  $\text{MoS}_2$  charge trap memory (CTM) under drain voltage pulses of equal amplitude.<sup>103,175</sup> The conductance update characteristics can be described by the empirical formula as follows:

$$G = G_{\min} + (G_{\max} - G_{\min})(1 - e^{-\nu p}), \quad (8)$$

where  $G_{\max}$  and  $G_{\min}$  are the initial and final conductance values,  $p$  is the normalized number of pulses, and  $\nu$  is a shape factor describing the linearity of the weight update.

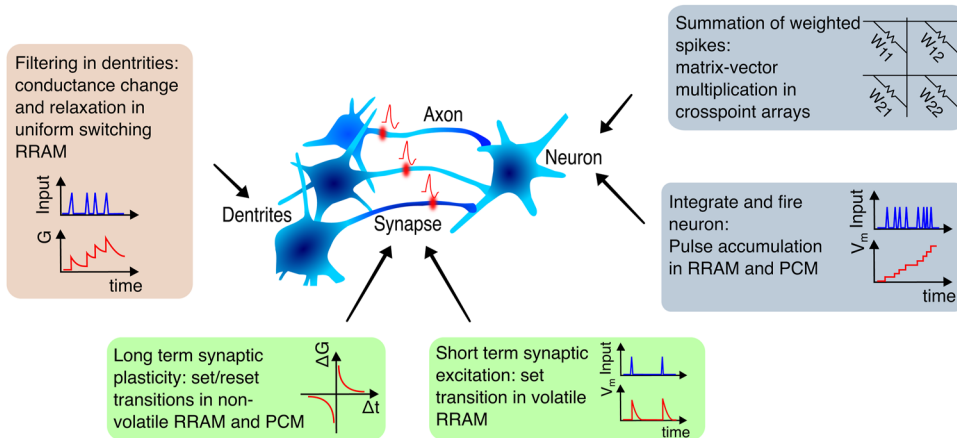
Figure 13(d) summarizes the metrics for synaptic memory devices, reporting the normalized conductance window  $(G_{\max} - G_{\min})/G_{\min}$ , describing the full-scale range of the synaptic weight as a function of the shape factor  $\nu$ , and describing linearity for various synaptic devices.<sup>15,69,175–179</sup> Among all the memory technologies, the CTM device combines excellent linearity of the weight update curve with a large conductance window. Note that the CTM device has a unidirectional characteristic, i.e., depression

is spontaneous and generally non-linear. However, this limitation is mitigated by a differential synapse scheme where two CTM devices are combined in the same synapse to map positive and negative weights.<sup>18</sup> CTM also offers extremely low conductance thanks to the sub-threshold operation, which is useful to suppress the IR drop and enable the training of large synaptic arrays.  $\text{MoS}_2$  also displays excellent scaling properties thanks to the atomically thin 2D semiconductor and the capability of 3D integration, thus providing a promising avenue for high-density 3D crosspoint arrays for training accelerators.<sup>180</sup>

## VIII. NEUROMORPHIC COMPUTING

Neuromorphic engineering aims at developing computing systems by using design principles that are based on those of the biological nervous systems.<sup>105,181</sup> By mimicking the human brain, the objective is to achieve a high energy efficiency, large parallelism, and the capacity to solve cognitive tasks, such as object recognition, association, adaptation, and learning.<sup>18</sup> Most importantly, the brain provides a blueprint for non-von Neumann computation, where information and memory are co-located in the same neurobiological network.<sup>182</sup> The neuromorphic term and concept were originally introduced in the early 1990s<sup>181</sup> and later revived in the early 2000s,<sup>183</sup> when the fast growth of online generated data started to spur the investigation of alternative computing paradigms. Recently, the neuromorphic engineering topic has seen a new wave of research interest in view of the added potential to embrace emerging memories as an enabling technology to implement brain-inspired processes.<sup>184–186</sup>

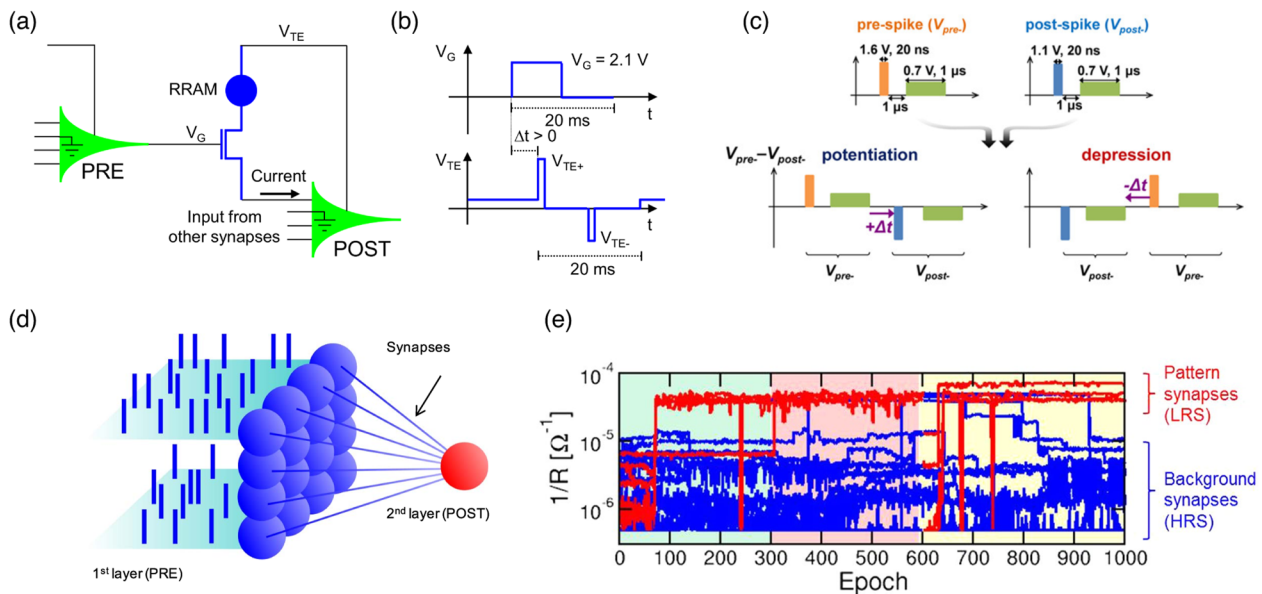
Figure 14 shows a summary of the main neurobiological features that can be implemented in a neuromorphic system, including synapses and neurons, the latter composed of a soma, an axon, and several dendrites.<sup>187,188</sup> Information is exchanged among neurons in the form of temporal spikes, which are weighted by synaptic connections and collected by the neuron soma. Synapses display synaptic plasticity, where the synaptic weight is changed upon spiking stimulation. Both long-term plasticity<sup>189,190</sup> and



**FIG. 14.** Schematic illustration of the main neuro-biological processes involved in neuromorphic brain-inspired computing, including neuron summation, integration and fire, dendritic filtering, and synaptic long- and short-term plasticity. Reproduced with permission from Ielmini *et al.*, *APL Mater.* **9**(5), 050702 (2021). Copyright 2021 AIP Publishing LLC.

short-term plasticity<sup>191</sup> have been evidenced by experiments. Over the years, several plasticity rules have been proposed, including paired-pulse facilitation (PPF),<sup>192,193</sup> spike-timing dependent plasticity (STDP),<sup>191,194–196</sup> triplet-based plasticity,<sup>197,198</sup> and spike-rate dependent plasticity (SRDP).<sup>199,200</sup> The hardware implementation of each element in Fig. 14 in CMOS technology generally requires complicated transistor-based circuits and large-area capacitors to

match the dynamic temporal evolution of the brain processes. From this standpoint, emerging memories offer a technology platform for providing nonvolatile synaptic weights capable of short- and long-term plasticity, increasing the area density of synapses and featuring unique dynamic properties with neuro-plausible time constant by the physical device mechanism.<sup>187,188</sup> For instance, synaptic long-term plasticity by STDP has been demonstrated in both



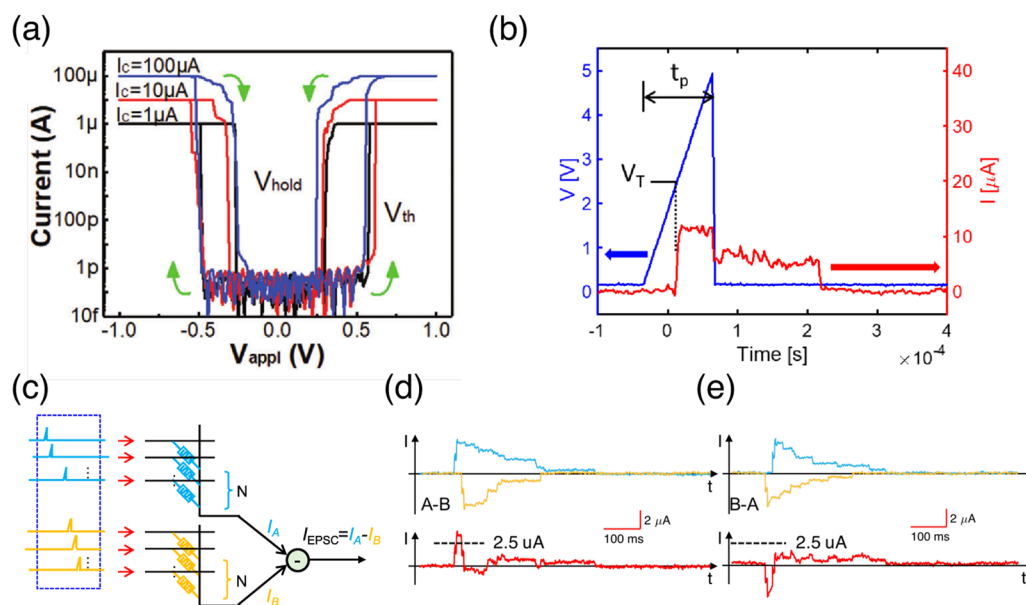
**FIG. 15.** Long-term plasticity in memory-based artificial synapses. (a) Structure of an STDP synapse based on RRAM with the 1T1R structure. Reproduced with permission from Ambrogio *et al.*, *IEEE Trans. Electron Devices* **63**(4), 1508–1515 (2016). Copyright 2016 Author(s), licensed under a Creative Commons Attribution 4.0 License. (b) Typical overlapping gate and TE voltages applied to the synapse for the case of synaptic potentiation with  $\Delta t > 0$ . Reproduced with permission from Ambrogio *et al.*, *IEEE Trans. Electron Devices* **63**(4), 1508–1515 (2016). Copyright 2016 Author(s), licensed under a Creative Commons Attribution 4.0 License. (c) Conceptual scheme of the STDP via non-overlapping spikes in a second-order memristor based on  $\text{Ta}_2\text{O}_{5-x}/\text{TaO}_x$ . Reproduced with permission from Kim *et al.*, *Nano Lett.* **15**(3), 2203–2211 (2015). Copyright 2015 American Chemical Society. (d) Schematic illustration of a perceptron-like neuromorphic network capable of unsupervised learning via STDP in memory-based synapses. Reproduced with permission from Pedretti *et al.*, *IEEE J. Emerging Sel. Top. Circuits Syst.* **8**(1), 77–85 (2017). Copyright 2017 Author(s), licensed under a Creative Commons Attribution 4.0 License. (e) Measured synaptic weights  $1/R$  as a function of spike number (epoch) for the perceptron in (d), indicating potentiation of stimulated synapses and depression of non-stimulated synapses. Reproduced with permission from Pedretti *et al.*, *Sci. Rep.* **7**(1), 5288 (2017). Copyright 2015 Author(s), licensed under a Creative Commons Attribution 4.0 License.

PCM<sup>201,202</sup> and RRAM.<sup>177,203–205</sup> Learning was shown to occur both by properly overlapping the pre- and post-synaptic spikes across the memory element<sup>205,206</sup> or by the physical interaction between thermal and electrical stimulations in the so-called second-order memristors.<sup>207</sup> Figures 15(a) and 15(b) show the 1 T1R synapse circuit with the typical pulses applied to the gate and TE. This circuit demonstrated both the synaptic weight update according to STDP and the communication between the PRE- and POST-neurons. Figure 15(c) shows instead the programming pulses and pre/post-spikes for STDP in a  $\text{Ta}_2\text{O}_{5-x}/\text{TaO}_y$  second-order memristor. By applying the pre- and post-spikes at the TE and BE, the interaction between the applied electric field and the local temperature leads to a  $\Delta t$ -dependent conductance change. Multi-synaptic circuits with 1T1R RRAM devices capable of STDP were shown to display unsupervised learning,<sup>101,208</sup> which is extremely promising for the development of the perceptron-like network capable of autonomous learning and adaptation [Figs. 15(d) and 15(e)].

### A. Brain-inspired computing with volatile memories

Volatile memory devices, while lacking a clear application in digital systems due to insufficient retention, provide an ideal

technology for reproducing short-term memory (STM) behavior in neuromorphic systems.<sup>193</sup> Volatile switching can be displayed in a class of filamentary RRAM devices where Ag or Cu are used as TE materials<sup>20,209</sup> or dispersed in the switching layer.<sup>210</sup> Figure 16(a) shows the typical  $I$ - $V$  characteristics of a volatile RRAM device based on Ag nanodots.<sup>211</sup> The volatile behavior is generally attributed to the filamentary switching and spontaneous rediffusion of Ag atoms to minimize the total energy of the filament.<sup>209</sup> Volatile RRAMs were initially proposed as selector elements in crosspoint memory arrays thanks to their large on/off ratio and low leakage current.<sup>212–214</sup> Later, these devices attracted interest from the neuromorphic community in view of their relatively long retention time similar to the biological time constants for STM.<sup>193,215</sup> For instance, Fig. 16(b) shows a typical pulsed characteristic of an Ag-based RRAM, stimulated by a triangular pulse. After the pulse, the current persists for a retention time of about 150  $\mu\text{s}$ , revealing the time decay of the filamentary path within the active material. Volatile switching of RRAM devices can be used as the fire function in an integrate-and-fire neuron circuit, thus avoiding the use of area-consuming amplifiers and pulse generators.<sup>216</sup> Volatile RRAMs have also been used for replicating PPF induced by paired spikes, where the pulsed-induced potentiation of the synaptic weight is enhanced by the application of two identical stimuli.<sup>217,218</sup> Most importantly,



**FIG. 16.** Short-term memory in artificial synapses based on volatile memories. (a) Measured  $I$ - $V$  characteristics of an RRAM device based on Ag nanodots, indicating the set transition to the on-state at  $V_{th}$  and spontaneous decay to the off-state at  $V_{hold}$ . Reproduced with permission from Li *et al.*, *Adv. Sci.* 7(22), 2002251 (2020). Copyright 2020 Author(s), licensed under a Creative Commons Attribution 4.0 License. (b) Pulsed characteristic of a volatile RRAM device, indicating the spontaneous decay to the off-state after spiking stimulation with a retention time of about 150  $\mu\text{s}$ . Reproduced with permission from Covi *et al.*, *IEEE Trans. Electron Devices* 68(9), 4335–4341 (2021). Copyright 2021 Author(s), licensed under a Creative Commons Attribution 4.0 License. (c) Schematic circuit for spatiotemporal recognition, where the EPSC is obtained as the comparison of excitatory and inhibitory synaptic currents. Reproduced with permission from Wang *et al.*, *Adv. Intell. Syst.* 3(4), 2000224 (2020). Copyright 2020 Author(s), licensed under a Creative Commons Attribution 4.0 License. (d) Measured EPSC for the case of preferred sequence A-B in (c), resulting in a positive current. Reproduced with permission from Wang *et al.*, *Adv. Intell. Syst.* 3(4), 2000224 (2020). Copyright 2020 Author(s), licensed under a Creative Commons Attribution 4.0 License. (e) Measured EPSC for the case of non-preferred sequence B-A in (c), resulting in a negative current. Reproduced with permission from Wang *et al.*, *Adv. Intell. Syst.* 3(4), 2000224 (2020). Copyright 2020 Author(s), licensed under a Creative Commons Attribution 4.0 License.



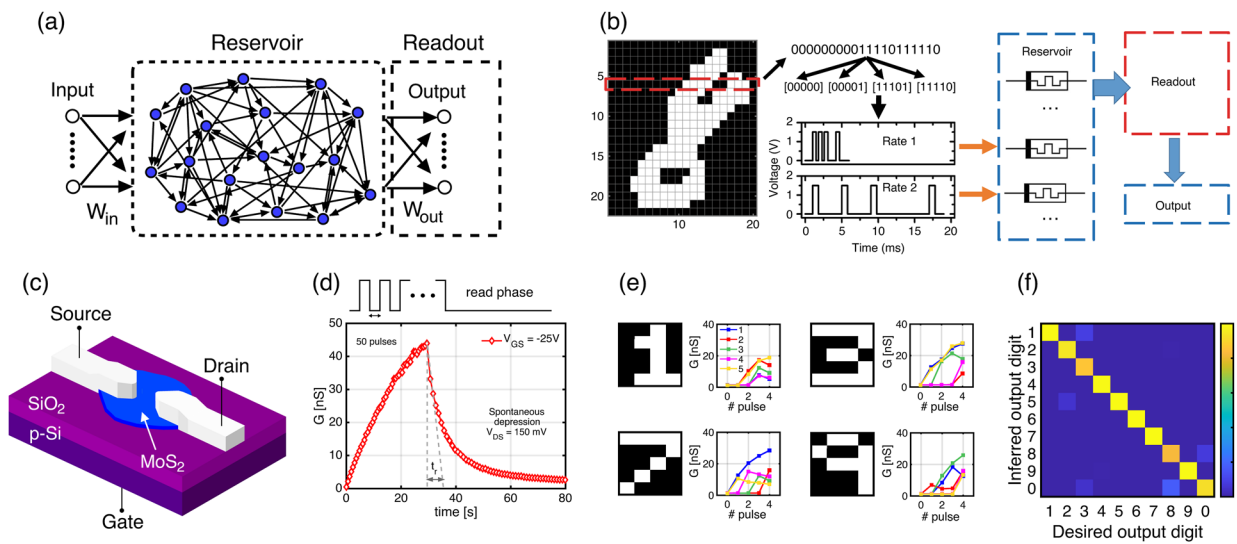
the dynamic STM effect can be useful to mimic sensing, learning, and processing of spatiotemporal patterns, such as audio and video sequences.

Figure 16(c) shows an example of spatiotemporal pattern recognition via volatile RRAM.<sup>219</sup> Two volatile synapses, serving as excitatory and inhibitory synapses, respectively, are stimulated by spikes A and B. Each synapse consists of several Ag-based volatile RRAM devices, where the spike stimulation and the persistent current cause an overall exponentially decaying response of each synapse as a result of Kirchhoff's law summation of each RRAM current contribution. The excitatory current  $I_{exc}$  and the inhibitory current  $I_{inh}$  are subtracted from each other to yield the excitatory postsynaptic current (EPSC) given by  $I_{EPSC} = I_{exc} - I_{inh}$ . Figures 16(d) and 16(e) show the synaptic currents and the EPSC for the case of the preferred sequence, namely, A–B, and the non-preferred sequence, namely, B–A. Due to the delay between the synaptic currents, the preferred sequence yields a positive EPSC, while the non-preferred sequence yields a negative EPSC. By comparing the EPSC with a threshold current, e.g.,  $I_{th} = 2.5 \mu A$  in Figs. 16(d) and 16(e) allows us to easily discriminate between the two patterns. This concept was applied to realize a retina-inspired artificial vision system capable of motion detection. In the biological retina, motion detection is achieved by direction-selective (DS) ganglion cells,<sup>220</sup> where excitatory and inhibitory synapses occupy

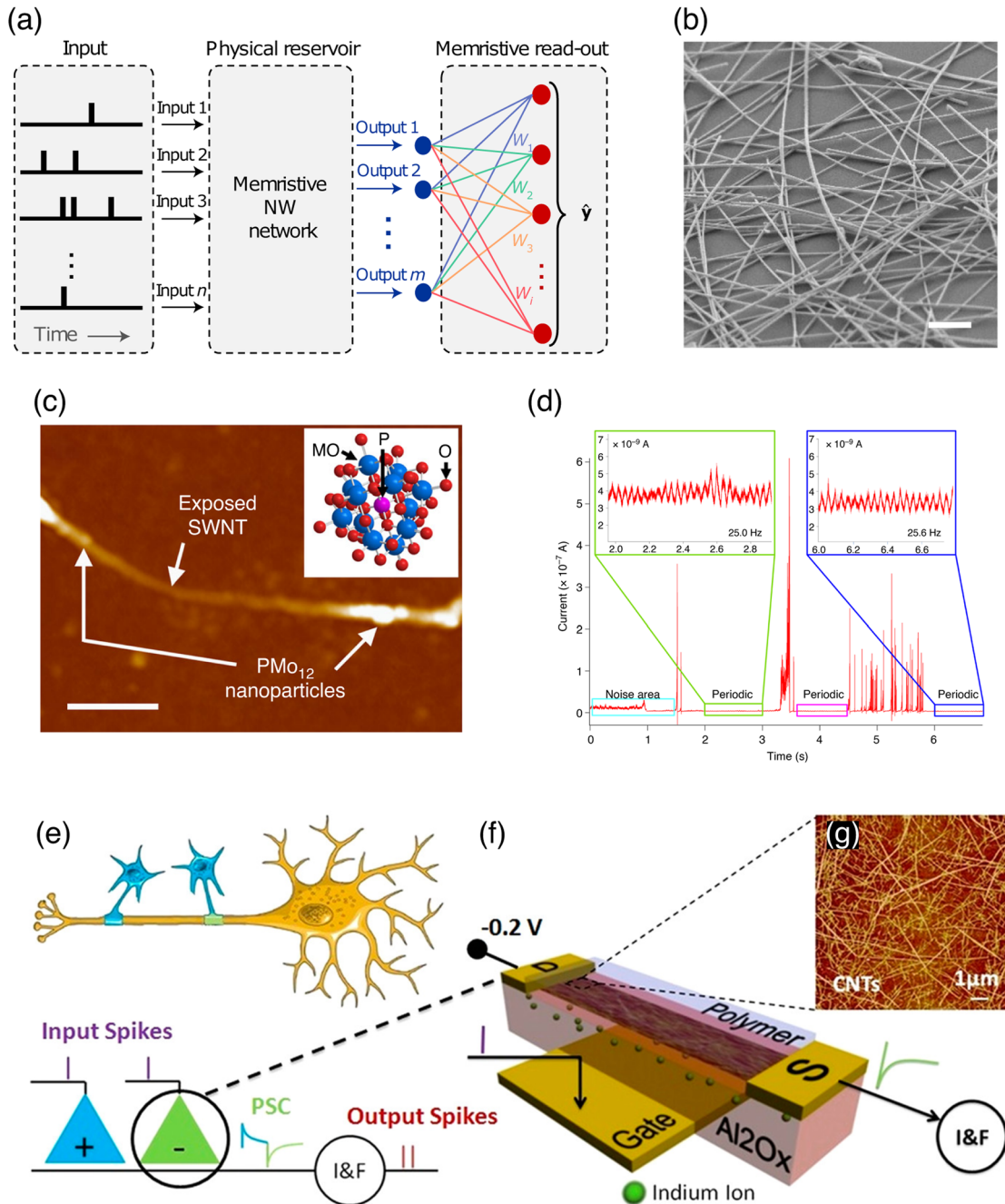
adjacent areas within the receptive field [Fig. 16(c)]. An image moving across the ganglion cell might stimulate the excitatory synapses followed by the inhibitory synapses, or vice versa, depending on the direction [Figs. 16(d) and 16(e)]. The EPSC of the ganglion cell thus allows us to recognize the direction of the image. The same concept can be extended to multiple directions by mimicking the starburst amacrine cell (SAC) structure in the retina, thus enabling a fast, low-power direction sensitivity in the analog domain.<sup>219,221</sup>

**B. Reservoir computing with volatile memories**

Reservoir computing (RC) is a modern machine learning technique, which is particularly suited to temporal/sequential information processing.<sup>222</sup> Figure 17(a) schematically shows the RC concept, which was originally conceived as an alternative approach to recurrent neural network (RNN) design and training, such as liquid state machines<sup>223</sup> and echo state networks.<sup>224</sup> In general, an RC network transforms sequential input data into a high-dimensional dynamical state via a reservoir layer. The output of the reservoir network is then processed by a readout layer to provide recognition and classification. The reservoir layer generally features random weights and connections, thus limiting the need for training to the readout layer and overcoming the complexity of multi-layer gradient-descent training techniques. Hardware RC networks are attracting interest



**FIG. 17.** Reservoir computing (RC) based on volatile memory devices. (a) Conceptual scheme of an RC system, composed of a random reservoir layer and a trained readout layer. Adapted from the work of Tanaka *et al.*, *Neural Networks* **115**, 100–123 (2019). Copyright 2019 Author(s), licensed under a Creative Commons Attribution 4.0 License. (b) RC system for handwritten digit recognition. The image is converted into a spatiotemporal pattern fed to the memory-based reservoir layer. The readout network processes the reservoir states for classification. Reproduced with permission from Du *et al.*, *Nat. Commun.* **8**(1), 2204 (2017). Copyright 2015 Author(s), licensed under a Creative Commons Attribution 4.0 License. (c) Illustration of the MoS<sub>2</sub>-based charge trap memory (CTM). Reproduced with permission from Farronato *et al.*, *Adv. Mater.* (published online) (2022). Copyright 2022 Author(s), licensed under a Creative Commons Attribution 4.0 License. (d) Measured characteristics of a MoS<sub>2</sub>-based CTM device showing pulse-induced potentiation followed by spontaneous decay. Reproduced with permission from Farronato *et al.*, *Adv. Mater.* (published online) (2022). Copyright 2022 Author(s), licensed under a Creative Commons Attribution 4.0 License. (e) Input patterns and corresponding reservoir states for a MoS<sub>2</sub>-based reservoir layer. Reproduced with permission from Farronato *et al.*, *Adv. Mater.* (published online) (2022). Copyright 2022 Author(s), licensed under a Creative Commons Attribution 4.0 License. (f) Confusion matrix for the MoS<sub>2</sub>-based RC system, demonstrating the classification results for digit images. Reproduced with permission from Farronato *et al.*, *Adv. Mater.* (published online) (2022). Copyright 2022 Author(s), licensed under a Creative Commons Attribution 4.0 License.



**FIG. 18.** In-materia neuromorphic computing. (a) Schematic of a general RC network with a random reservoir layer and a properly trained readout network for the recognition of spatiotemporal patterns. Reproduced with permission from Milano *et al.*, *Nat. Mater.* **21**(2), 195–202 (2022). Copyright 2021 Springer Nature Limited. (b) SEM image of a memristive nanowire network used as the reservoir layer. Scale bar is 2  $\mu$ m. Reproduced with permission from Milano *et al.*, *Nat. Mater.* **21**(2), 195–202 (2022). Copyright 2021 Springer Nature Limited. (c) Atomic force microscopy (AFM) image of a single-walled carbon nanotube (SWCNT) within a SWCNT-based transistor. Reproduced with permission from Tanaka *et al.*, *Nat. Commun.* **9**(1), 2693 (2018). Copyright 2015 Author(s), licensed under a Creative Commons Attribution 4.0 License. (d) Measured response of a SWCNT transistor, including noisy and periodic dynamics. Reproduced with permission from Tanaka *et al.*, *Nat. Commun.* **9**(1), 2693 (2018). Copyright 2015 Author(s), licensed under a Creative Commons Attribution 4.0 License. (e) Schematic of a neurobiological model with two pre-synaptic spikes and integrate-and-fire neurons. Reproduced with permission from Shen *et al.*, *ACS Nano* **7**(7), 6117–6122 (2013). Copyright 2013 American Chemical Society. (f) Synaptic transistor based on an SWCNT network. Reproduced with permission from Shen *et al.*, *ACS Nano* **7**(7), 6117–6122 (2013). Copyright 2013 American Chemical Society. (g) AFM image of a random SWCNT network in the transistor channel. Reproduced with permission from Shen *et al.*, *ACS Nano* **7**(7), 6117–6122 (2013). Copyright 2013 American Chemical Society.

thanks to their potential in energy efficiency, high versatility, and fast learning.<sup>225–227</sup>

Figure 17(b) schematically shows an IMC-based RC network for image recognition.<sup>228</sup> First, the input pattern, e.g., the image of a handwritten digit, is converted into a spatiotemporal pattern, where rows represent the sequential spikes and columns represent the  $N$  input channels. The resulting spatiotemporal is fed to  $N$  volatile RRAM devices where the STM response provides a physical reservoir layer. The dynamic reservoir layer yields a unique output response, e.g., the output transient current, to each input pattern, which can then be classified by the readout layer, consisting of a properly trained fully connected network.

RC was demonstrated by using charge-trap memory (CTM) devices based on a MoS<sub>2</sub>-based channel.<sup>103</sup> Figure 17(c) shows the device structure with source/drain contacts deposited on a MoS<sub>2</sub> channel, where inversion and depletion were controlled by a back gate. In this device, a positive or negative gate voltage results in the trapping of electrons or holes, respectively, at the interface between MoS<sub>2</sub> and SiO<sub>2</sub>, the latter serving as gate dielectric layer. Electron/hole trappings cause a shift of threshold voltage, thus resulting in a change in the channel conductivity. This is shown in Fig. 17(d), where a train of negative gate pulses leads to an increase in conductance, which spontaneously decays at the end of the stimulation. The dynamic response in Fig. 17(d) was used as a physical reservoir process in an RC network for image recognition with 5 CTM devices as the reservoir layer.<sup>103</sup> Figure 17(e) shows examples of the reservoir output, indicating potentiation and spontaneous decay as a result of the spatiotemporal stimulation. After training the readout layer by the logistic regression,<sup>157</sup> a good classification accuracy was achieved, as shown by the confusion diagram in Fig. 17(f). Compared to DNNs, RC networks employ fewer devices by leveraging the rich analog, dynamic response of the CTM device, thus resulting in a significantly smaller classification network.<sup>229</sup> In addition, power consumption can be minimized in the RC layer by operating the CTM device in the subthreshold regime.<sup>103</sup> Similar spatiotemporal RC networks were used for solving second-order nonlinear equations,<sup>228</sup> spoken-digit recognition,<sup>229</sup> and autonomous chaotic time-series forecasting,<sup>229</sup> thus supporting the wide application scenario for RC-based IMC circuits.

### C. In-materia computing

The principle of using device physics to achieve smart computing functions is further extended from devices to materials in the so-called *in-materia computing*.<sup>230,231</sup> In-materia computing relies on the ability of certain materials, such as nanoparticles, nanostructures, or even randomly-doped semiconductors, to act as a distributed, random network of physical dynamical nodes for computation.<sup>232</sup> In-materia computing systems include nanostructures based on carbon nanotubes (CNTs),<sup>233,234</sup> nanowires (NWs),<sup>235–237</sup> and metallic nanoparticles.<sup>238</sup> Indeed, programming, stimulating, and controlling the individual nodes in the computing materials are a challenging task since the materials can exhibit dynamic fluctuations.<sup>239,240</sup> However, nanostructures are ideally suited to serve as the randomly connected reservoir layer of an RC network.<sup>225,236</sup> Figure 18(a) shows a fully memristive RC system where the RC layer is made of a network of silver nanowires (NWs), which is shown in Fig. 18(b).<sup>236</sup> The electrical stimulation of the NW network induces a change

in the NW cross-point junctions,<sup>235</sup> thus resulting in a dynamic potentiation of the local connection, hence the local effective conductance. The output of the reservoir, i.e., the output current or the node potential of the NW network, is then processed by the readout layer, e.g., a fully connected network of RRAM devices. By properly training the readout network, tasks such as image recognition and spatiotemporal pattern prediction can be carried out.<sup>236</sup> This approach to computation has distinct advantages in terms of scaling and easy manufacturing thanks to the bottom-up technology for developing the physical NW network. Figure 18(c) shows a neuromorphic device composed of a single-walled carbon nanotube (SWCNT) complexed with polyoxometalate (POM).<sup>234,241</sup> When arranged in a network, SWCNT can spontaneously generate spikes and noise thanks to multi-redox activities at the crossing points.<sup>242</sup> Both periodic and aperiodic current spikes are generated under a constant-voltage bias, as shown in Fig. 18(d). The applied bias causes the conductance to switch between POMs and SWCNTs, thus mimicking the potentiation behavior of a neurobiological synapse. Chemical reaction phenomena, such as aggregation and dissociation of counter-cations, play an additional role, thus leading to spike generation. Similar to the NW network of Fig. 18(b), the POM/SWCNT network can serve as a reservoir layer in an RC system thanks to its nonlinear dynamic.<sup>234</sup>

SWCNT networks were also used as analog synapses in the neuromorphic module of Fig. 18(e).<sup>233</sup> The module consists of a single neuron connected with other neurons through synapses. The synapses are emulated by transistors based on a random CNT network, while the axon in the neuron is realized by Si-based transistors. Figure 18(f) shows the CNTs-based synaptic transistor, with the random SWCNT network in the inset. Electron trapping in the dielectric layer due to the application of gate pulses results in an increase of current in the p-type SWCNT channel. Potentiation is followed by decay due to the tunneling of electrons out from the dielectric layer. The SWCNT-based synapse also shows inhibitory characteristics under the negative voltage of the gate. Potentiation/depression allows for the emulation of biological STDP and PPF, which is promising for the development of in-materia neuromorphic computing systems.

## IX. OUTLOOK

The main enablers of IMC are emerging memory devices, whose distinct advantages, such as nonvolatile behavior, make them more appealing than SRAM<sup>243</sup> or DRAM<sup>244</sup> although at the expense of increased programming energy and times.<sup>245,246</sup> For tasks where computational parameters must be frequently updated, such as stateful Boolean logic circuits,<sup>96,97</sup> the programming overhead may overshadow the advantages of IMC. Moreover, given the fundamentally different characteristics of emerging memories in terms of linearity, power consumption, conductance window, noise, and CMOS compatibility,<sup>245,247,248</sup> it is difficult to identify a best-in-class technology with universal applicability across all IMC applications.<sup>100,134,170,249–253</sup> As an example, combinatorial optimization tasks<sup>254–256</sup> inherently require controllable, device-level randomness<sup>148</sup> as an enabling feature for simulated annealing.<sup>106</sup> On the other hand, scientific computing applications show extremely narrow tolerance to perturbation and noise,<sup>257</sup> relying on high-precision data storage to provide high-quality results.<sup>249</sup> The search

for a *universal memory*, capable of satisfying the requirements of many applications at the same time, is thus still open. One of the main pathways for the implementation of in-memory computing is the reduction of the power consumption of memory devices to allow for the operation of extremely large arrays with an affordable cost. Another key challenge is the improvement of reliability, e.g., the realization of self-selecting, multilevel memory devices with a large endurance and low variability. At the present time, these requirements can be partially solved by proper programming approaches (program and verify algorithms) or device implementation (1T1R structures, etc.) at the cost of operation slowness and decrease of integration density.

Many of the advantages of IMC derive from the collective behavior of densely packed memory cells in an array configuration. Common parasitics, such as line resistance and capacitance,<sup>258</sup> can limit the accuracy of both write and read operations, thus affecting the reliability of IMC.<sup>114,259</sup> While selector devices alleviate the issue during the programming phase, they have limited impact during computation as all cells are simultaneously selected. Schemes for parasitic compensation<sup>164,260,261</sup> may help mitigate the issue at the expense of increased pre-processing overhead and reduced effectiveness for large array size. For error-tolerant or adaptive applications, optimization frameworks can be developed<sup>262,263</sup> with negligible loss of accuracy. Another approach is to use three-terminal devices with ultra-low conductance, such as ECRAM and MoS<sub>2</sub> CTM devices,<sup>175</sup> to minimize both the IR drop and the line capacitances of the array. However, large-scale crosspoint arrays of two-terminal devices have been exhaustively demonstrated in academia and industry,<sup>264–268</sup> whereas the same maturity level is currently lacking for arrays of three-terminal emerging memory devices.<sup>76,80</sup>

Power consumption is another key consideration imposing constraints on the individual array size.<sup>122,269,270</sup> Power can be handled by arranging the IMC system with tiled architecture<sup>7</sup> where multiple replicas of a fundamental computing macro, or *core*, work in parallel for the execution of a computing task. Core architecture design is another open quest in the field of IMC, where computational efficiency and robustness must be balanced with analog-to-digital and digital-to-analog conversion overheads.<sup>248</sup> On the one hand, IMC-specific conversion front-ends<sup>271,272</sup> should balance accuracy, latency, energy, and area consumption. On the other hand, various approaches to data encoding, such as amplitude modulation<sup>134,273</sup> or pulse-width modulation,<sup>136</sup> require conversion circuits to be flexible and reconfigurable. Finally, proper design of the inter-core communication is crucial to maintain the IMC advantage and allow for the solution of large-scale problems.<sup>274</sup> Co-optimization of the device, architecture, and application seems to be the most promising concept to fully unleash the IMC potential in overcoming the von Neumann bottleneck.<sup>269,275</sup>

Finally, to allow for widespread IMC adoption, it is essential to bridge the gap between hardware and software by implementing an electronic design automation (EDA) toolchain. On the one hand, IMC-specific design tools<sup>276</sup> are useful for system designers and engineers to develop large-scale, highly accurate IMC hardware and software systems. On the other hand, end users operating at a higher level of abstraction need a software stack capable of transparently compiling a given problem for a target IMC architecture optimization.<sup>277–279</sup> This challenge should be tackled by the

codesign and co-development of a full set of hardware and software tools to elevate the maturity of IMC for real-life applications.

## X. CONCLUSIONS

This Perspective provides a review of the status and outlook of IMC with emerging memory devices. The candidate alternatives to the conventional von Neumann architecture are presented and compared in terms of their degree of integration between memory and computing units. Two-terminal and three-terminal emerging memory devices are reviewed. By distinguishing two general operating regimes of emerging devices, low-voltage static IMC and high-voltage dynamic IMC are identified as the main IMC macro-categories. Correspondingly, the most relevant computing primitives are explored in view of their real-world applications. For static IMC, MVM and IMVM accelerators, as well as TCAMs, are presented together with their applications in machine learning, hardware security, and data classification. Similarly, for dynamic IMC, outer-product accelerators for neural network training and brain-inspired systems for reservoir computing are discussed. Finally, challenges for the *in silico* implementation of an IMC architecture are outlined. Owing to the overarching nature of IMC, encompassing device, computing core, and the EDA toolchain, a strongly multidisciplinary approach is needed to co-optimize all components and fully unleash the IMC potential.

## ACKNOWLEDGMENTS

This work received funding from the Italian Ministry of University and Research (MUR) and the European Union (EU) under the PON/REACT program and the Horizon 2020 Research and Innovation program (Grant Agreement Nos. 824164 and 899559). This work also received funding from ECSEL Joint Undertaking (JU) under Grant Agreement No. 101007321. The JU receives support from the European Union's Horizon 2020 Research and Innovation program and France, Belgium, Czech Republic, Germany, Italy, Sweden, Switzerland, and Turkey.

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

### Author Contributions

P.M. and M.F. contributed equally to this work.

**P. Mannocci:** Writing – original draft (lead). **M. Farronato:** Writing – original draft (lead). **N. Lepri:** Writing – original draft (equal). **L. Cattaneo:** Writing – original draft (equal). **A. Glukhov:** Writing – original draft (equal). **Z. Sun:** Writing – original draft (equal). **D. Ielmini:** Conceptualization (lead); Funding acquisition (lead); Project administration (lead); Resources (lead); Supervision (lead); Writing – review & editing (lead).

## DATA AVAILABILITY

The data that support the findings of this study are openly available in <https://zenodo.org/record/7378087#.Y4Y8xHbMKCo>.

## REFERENCES

- <sup>1</sup>W. A. Wulf and S. A. McKee, "Hitting the memory wall: Implications of the obvious," *ACM SIGARCH Comput. Archit. News* **23**(1), 20–24 (1995).
- <sup>2</sup>M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (IEEE, San Francisco, CA, 2014), pp. 10–14.
- <sup>3</sup>H. Jun, J. Cho *et al.*, "HBM (high bandwidth memory) DRAM technology and architecture," in *2017 IEEE International Memory Workshop (IMW)* (IEEE, Monterey, CA, 2017), pp. 1–4.
- <sup>4</sup>J. Jeddeloh and B. Keeth, "Hybrid memory cube new DRAM architecture increases density and performance," in *2012 Symposium on VLSI Technology (VLSIT)* (IEEE, Honolulu, HI, 2012), pp. 87–88.
- <sup>5</sup>D. Patterson, T. Anderson *et al.*, "A case for intelligent RAM," *IEEE Micro* **17**(2), 34–44 (1997).
- <sup>6</sup>M. A. Zidan, J. P. Strachan, and W. D. Lu, "The future of electronics based on memristive systems," *Nat. Electron.* **1**(1), 22–29 (2018).
- <sup>7</sup>D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nat. Electron.* **1**(6), 333–343 (2018).
- <sup>8</sup>S. Mittal, G. Verma *et al.*, "A survey of SRAM-based in-memory computing techniques and applications," *J. Syst. Archit.* **119**, 102276 (2021).
- <sup>9</sup>H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nat. Nanotechnol.* **10**(3), 191–194 (2015).
- <sup>10</sup>J. Wang, X. Wang *et al.*, "A 28-nm compute SRAM with bit-serial logic/arithmetic operations for programmable in-memory vector computing," *IEEE J. Solid-State Circuits* **55**(1), 76–86 (2020).
- <sup>11</sup>W. Wang, W. Song *et al.*, "Integration and co-design of memristive devices and algorithms for artificial intelligence," *iScience* **23**(12), 101809 (2020).
- <sup>12</sup>C. M. Compagnoni, A. Goda *et al.*, "Reviewing the evolution of the NAND flash technology," *Proc. IEEE* **105**(9), 1609–1633 (2017).
- <sup>13</sup>R. Micheloni, L. Crippa *et al.*, *Inside NAND Flash Memories* (Springer Science & Business Media, 2010).
- <sup>14</sup>B. Govoreanu, G. S. Kar *et al.*, "10×10nm<sup>2</sup> Hf/HfO<sub>x</sub> crossbar resistive RAM with excellent performance, reliability and low-energy operation," in *2011 IEEE International Electron Devices Meeting (IEDM)* (IEEE, Washington, DC, 2011), pp. 31.6.1–31.6.4.
- <sup>15</sup>C.-W. Hsu, C.-C. Wan *et al.*, "3D vertical TaO<sub>x</sub>/TiO<sub>2</sub> RRAM with over 10<sup>3</sup> self-rectifying ratio and sub-μA operating current," in *2013 IEEE International Electron Devices Meeting (IEDM)* (IEEE, Washington, DC, 2013), pp. 10.4.1–10.4.4.
- <sup>16</sup>Q. Luo, X. Xu *et al.*, "8-layers 3D vertical RRAM with excellent scalability towards storage class memory applications," in *2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2017), pp. 2.7.1–2.7.4.
- <sup>17</sup>J. A. Kittl, K. Opsomer *et al.*, "High-k dielectrics for future generation memory devices," *Microelectron. Eng.* **86**(7–9), 1789–1795 (2009).
- <sup>18</sup>D. Ielmini and S. Ambrogio, "Emerging neuromorphic devices," *Nanotechnol.ogy* **31**(9), 092001 (2020).
- <sup>19</sup>S. Brivio, S. Spiga, and D. Ielmini, "HfO<sub>2</sub>-based resistive switching memory devices for neuromorphic computing," *Neuromorph. Comput. Eng.* **2**, 042001 (2022).
- <sup>20</sup>A. Bricalli, E. Ambrosi *et al.*, "Resistive switching device technology based on silicon oxide for improved ON-OFF ratio—Part II: Select devices," *IEEE Trans. Electron Devices* **65**(1), 122–128 (2018).
- <sup>21</sup>W.-G. Kim and S.-W. Rhee, "Effect of the top electrode material on the resistive switching of TiO<sub>2</sub> thin film," *Microelectron. Eng.* **87**(2), 98–103 (2010).
- <sup>22</sup>H. Akinaga and H. Shima, "Resistive random access memory (ReRAM) based on metal oxides," *Proc. IEEE* **98**(12), 2237–2251 (2010).
- <sup>23</sup>Z. Zhang, B. Gao *et al.*, "All-metal-nitride RRAM devices," *IEEE Electron Device Lett.* **36**(1), 29–31 (2015).
- <sup>24</sup>J. Chen, C.-Y. Lin *et al.*, "LiSiO<sub>x</sub>-based analog memristive synapse for neuromorphic computing," *IEEE Electron Device Lett.* **40**(4), 542–545 (2019).
- <sup>25</sup>U. Russo, D. Kamalanathan *et al.*, "Study of multilevel programming in programmable metallization cell (PMC) memory," *IEEE Trans. Electron Devices* **56**(5), 1040–1047 (2009).
- <sup>26</sup>C. Pan, Y. Ji *et al.*, "Coexistence of grain-boundaries-assisted bipolar and threshold resistive switching in multilayer hexagonal boron nitride," *Adv. Funct. Mater.* **27**(10), 1604811 (2017).
- <sup>27</sup>Y. Shen, W. Zheng *et al.*, "Variability and yield in h-BN-based memristive circuits: The role of each type of defect," *Adv. Mater.* **33**(41), 2103656 (2021).
- <sup>28</sup>S. Goswami, A. J. Matula *et al.*, "Robust resistive memory devices using solution-processable metal-coordinated azo aromatics," *Nat. Mater.* **16**(12), 1216–1224 (2017).
- <sup>29</sup>S. Goswami, S. P. Rath *et al.*, "Charge disproportionate molecular redox for discrete memristive and memcapacitive switching," *Nat. Nanotechnol.* **15**(5), 380–389 (2020).
- <sup>30</sup>S. Goswami, R. Pramanick *et al.*, "Decision trees within a molecular memristor," *Nature* **597**(7874), 51–56 (2021).
- <sup>31</sup>D. Ielmini, R. Bruchhaus, and R. Waser, "Thermochemical resistive switching: Materials, mechanisms, and scaling projections," *Phase Transitions* **84**(7), 570–602 (2011).
- <sup>32</sup>H. Y. Lee, P. S. Chen *et al.*, "Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO<sub>2</sub> based RRAM," in *2008 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2008), pp. 1–4.
- <sup>33</sup>A. Sawa, "Resistive switching in transition metal oxides," *Mater. Today* **11**(6), 28–36 (2008).
- <sup>34</sup>S. Raoux, W. Welnic, and D. Ielmini, "Phase change materials and their application to nonvolatile memories," *Chem. Rev.* **110**(1), 240–267 (2010).
- <sup>35</sup>G. W. Burr, M. J. Breitwisch *et al.*, "Phase change memory technology," *J. Vac. Sci. Technol. B* **28**(2), 223–262 (2010).
- <sup>36</sup>D. Ielmini and A. L. Lacaíta, "Phase change materials in non-volatile storage," *Mater. Today* **14**(12), 600–607 (2011).
- <sup>37</sup>M. Wuttig and N. Yamada, "Phase-change materials for rewriteable data storage," *Nat. Mater.* **6**(11), 824–832 (2007).
- <sup>38</sup>D. Ielmini, A. L. Lacaíta *et al.*, "Analysis of phase distribution in phase-change nonvolatile memories," *IEEE Electron Device Lett.* **25**(7), 507–509 (2004).
- <sup>39</sup>U. Russo, D. Ielmini, and A. L. Lacaíta, "Analytical modeling of chalcogenide crystallization for PCM data-retention extrapolation," *IEEE Trans. Electron Devices* **54**(10), 2769–2777 (2007).
- <sup>40</sup>D. Ielmini, A. L. Lacaíta, and D. Mantegazza, "Recovery and drift dynamics of resistance and threshold voltages in phase-change memories," *IEEE Trans. Electron Devices* **54**(2), 308–315 (2007).
- <sup>41</sup>P. Zuliani, E. Varesi *et al.*, "Overcoming temperature limitations in phase change memories with optimized Ge<sub>x</sub>Sb<sub>y</sub>Te<sub>z</sub>," *IEEE Trans. Electron Devices* **60**(12), 4020–4026 (2013).
- <sup>42</sup>S. Kim, N. Sosa *et al.*, "A phase change memory cell with metallic surfactant layer as a resistance drift stabilizer," in *2013 IEEE International Electron Devices Meeting (IEDM)* (IEEE, Washington, DC, 2013), pp. 30.7.1–30.7.4.
- <sup>43</sup>F. Arnaud, P. Zuliani *et al.*, "Truly innovative 28nm FDSOI technology for automotive micro-controller applications embedding 16MB phase change memory," in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2018), pp. 18.4.1–18.4.4.
- <sup>44</sup>D. Min, J. Park *et al.*, "18nm FDSOI technology platform embedding PCM & innovative continuous-active construct enhancing performance for leading-edge MCU applications," in *2021 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2021), pp. 13.1.1–13.1.4.
- <sup>45</sup>P. Zuliani, A. Conte, and P. Cappelletti, "The PCM way for embedded non volatile memories applications," in *2019 Symposium on VLSI Technology* (IEEE, Kyoto, Japan, 2019), pp. T192–T193.
- <sup>46</sup>T. Mikolajick, C. Dehm *et al.*, "FeRAM technology for high density applications," *Microelectron. Reliab.* **41**(7), 947–950 (2001).
- <sup>47</sup>D. J. Kim, J. Y. Jo *et al.*, "Polarization relaxation induced by a depolarization field in ultrathin ferroelectric BaTiO<sub>3</sub> capacitors," *Phys. Rev. Lett.* **95**(23), 237602 (2005).

- <sup>48</sup>J. F. Scott, "Applications of modern ferroelectrics," *Science* **315**(5814), 954–959 (2007).
- <sup>49</sup>T. S. Böscke, J. Müller *et al.*, "Ferroelectricity in hafnium oxide thin films," *Appl. Phys. Lett.* **99**(10), 102903 (2011).
- <sup>50</sup>J.-M. Koo, B.-S. Seo *et al.*, "Fabrication of 3D trench PZT capacitors for 256Mbit FRAM device application," in *2005 IEEE International Electron Devices Meeting (IEDM). IEDM Technical Digest* (IEEE, 2005), pp. 340–343.
- <sup>51</sup>P. Polakowski, S. Riedel *et al.*, "Ferroelectric deep trench capacitors based on Al:HfO<sub>2</sub> for 3D nonvolatile memory applications," in *2014 IEEE 6th International Memory Workshop (IMW)* (IEEE, 2014), pp. 1–4.
- <sup>52</sup>M. Pešić, F. P. G. Fengler *et al.*, "Physical mechanisms behind the field-cycling behavior of HfO<sub>2</sub>-based ferroelectric capacitors," *Adv. Funct. Mater.* **26**(25), 4601–4612 (2016).
- <sup>53</sup>A. Chanthbouala, A. Crassous *et al.*, "Solid-state memories based on ferroelectric tunnel junctions," *Nat. Nanotechnol.* **7**(2), 101–104 (2012).
- <sup>54</sup>D. Wang, C. Nordman *et al.*, "70% TMR at room temperature for SDT sandwich junctions with CoFeB as free and reference layers," *IEEE Trans. Magn.* **40**(4), 2269–2271 (2004).
- <sup>55</sup>C. Chappert, A. Fert, and F. N. Van Dau, "The emergence of spin electronics in data storage," *Nat. Mater.* **6**(11), 813–823 (2007).
- <sup>56</sup>J. C. Sankey, Y.-T. Cui *et al.*, "Measurement of the spin-transfer-torque vector in magnetic tunnel junctions," *Nat. Phys.* **4**(1), 67–71 (2008).
- <sup>57</sup>S. Ikeda, K. Miura *et al.*, "A perpendicular-anisotropy CoFeB–MgO magnetic tunnel junction," *Nat. Mater.* **9**(9), 721–724 (2010).
- <sup>58</sup>S. Sakhare, M. Perumkunnil *et al.*, "Enablement of STT-MRAM as last level cache for the high performance computing domain at the 5nm node," in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2018), pp. 18.3.1–18.3.4.
- <sup>59</sup>J. Grollier, D. Querlioz, and M. D. Stiles, "Spintronic nanodevices for bioinspired computing," *Proc. IEEE* **104**(10), 2024–2039 (2016).
- <sup>60</sup>A. I. Khan, A. Keshavarzi, and S. Datta, "The future of ferroelectric field-effect transistor technology," *Nat. Electron.* **3**(10), 588–597 (2020).
- <sup>61</sup>K. Sugibuchi, Y. Kurogi, and N. Endo, "Ferroelectric field-effect memory device using Bi<sub>4</sub>Ti<sub>3</sub>O<sub>12</sub> film," *J. Appl. Phys.* **46**(7), 2877–2881 (1975).
- <sup>62</sup>K. Florent, M. Pesic *et al.*, "Vertical ferroelectric HfO<sub>2</sub> FET based on 3-D NAND architecture: Towards dense low-power memory," in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2018), pp. 2.5.1–2.5.4.
- <sup>63</sup>K. A. Aabrar, J. Gomez *et al.*, "BEOL compatible superlattice FerroFET-based high precision analog weight cell with superior linearity and symmetry," in *2021 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2021), pp. 19.6.1–19.6.4.
- <sup>64</sup>I. M. Miron, K. Garello *et al.*, "Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection," *Nature* **476**(7359), 189–193 (2011).
- <sup>65</sup>K. Garello, C. O. Avci *et al.*, "Ultrafast magnetization switching by spin-orbit torques," *Appl. Phys. Lett.* **105**(21), 212402 (2014).
- <sup>66</sup>T. Endoh, H. Honjo *et al.*, "Recent progresses in STT-MRAM and SOT-MRAM for next generation MRAM," in *2020 IEEE Symposium on VLSI Technology* (IEEE, Honolulu, HI, 2020), pp. 1–2.
- <sup>67</sup>H. Wu, J. Zhang *et al.*, "Field-free approaches for deterministic spin-orbit torque switching of the perpendicular magnet," *Mater. Futures* **1**(2), 022201 (2022).
- <sup>68</sup>E. J. Fuller, S. T. Keene *et al.*, "Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing," *Science* **364**(6440), 570–574 (2019).
- <sup>69</sup>J. Tang, D. Bishop *et al.*, "ECRAM as scalable synaptic cell for high-speed, low-power neuromorphic computing," in *2018 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2018), pp. 13.1.1–13.1.4.
- <sup>70</sup>S. Kim, T. Todorov *et al.*, "Metal-oxide based, CMOS-compatible ECRAM for deep learning accelerator," in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2019), pp. 35.7.1–35.7.4.
- <sup>71</sup>Y. Li, E. J. Fuller *et al.*, "Filament-free bulk resistive memory enables deterministic analogue switching," *Adv. Mater.* **32**(45), 2003984 (2020).
- <sup>72</sup>E. J. Fuller, F. E. Gabaly *et al.*, "Li-ion synaptic transistor for low power analog computing," *Adv. Mater.* **29**(4), 1604310 (2017).
- <sup>73</sup>Y. van de Burgt, E. Lubberman *et al.*, "A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing," *Nat. Mater.* **16**(4), 414–418 (2017).
- <sup>74</sup>M. Berggren, X. Crispin *et al.*, "Ion electron-coupled functionality in materials and devices based on conjugated polymers," *Adv. Mater.* **31**(22), 1805813 (2019).
- <sup>75</sup>P. C. Harikeśh, C.-Y. Yang *et al.*, "Organic electrochemical neurons and synapses with ion mediated spiking," *Nat. Commun.* **13**(1), 901 (2022).
- <sup>76</sup>H. Lee, D. G. Ryu *et al.*, "Vertical metal-oxide electrochemical memory for high-density synaptic array based high-performance neuromorphic computing," *Adv. Electron. Mater.* **8**(8), 2200378 (2022).
- <sup>77</sup>V. K. Sangwan, D. Jariwala *et al.*, "Gate-tunable memristive phenomena mediated by grain boundaries in single-layer MoS<sub>2</sub>," *Nat. Nanotechnol.* **10**(5), 403–406 (2015).
- <sup>78</sup>V. K. Sangwan, H.-S. Lee *et al.*, "Multi-terminal memtransistors from polycrystalline monolayer molybdenum disulfide," *Nature* **554**(7693), 500–504 (2018).
- <sup>79</sup>M. Farronato, M. Melegari *et al.*, "Memtransistor devices based on MoS<sub>2</sub> multilayers with volatile switching due to Ag cation migration," *Adv. Electron. Mater.* **8**(8), 2101161 (2022).
- <sup>80</sup>H. S. Lee, V. K. Sangwan *et al.*, "Dual-gated MoS<sub>2</sub> memtransistor crossbar array," *Adv. Funct. Mater.* **30**(45), 2003683 (2020).
- <sup>81</sup>S. Hao, X. Ji *et al.*, "A monolayer leaky integrate-and-fire neuron for 2D memristive neuromorphic networks," *Adv. Electron. Mater.* **6**(4), 1901335 (2020).
- <sup>82</sup>R. A. John, F. Liu *et al.*, "Synergistic gating of electro-iono-photoactive 2D chalcogenide neuristors: Coexistence of hebbian and homeostatic synaptic metaplasticity," *Adv. Mater.* **30**(25), 1800220 (2018).
- <sup>83</sup>R. A. John, J. Acharya *et al.*, "Optogenetics inspired transition metal dichalcogenide neuristors for in-memory deep recurrent neural networks," *Nat. Commun.* **11**(1), 3211 (2020).
- <sup>84</sup>E. Fortunato, P. Barquinha, and R. Martins, "Oxide semiconductor thin-film transistors: A review of recent advances," *Adv. Mater.* **24**(22), 2945–2986 (2012).
- <sup>85</sup>R. A. John, N. Tiwari *et al.*, "Ultralow power dual-gated subthreshold oxide neuristors: An enabler for higher order neuronal temporal correlations," *ACS Nano* **12**(11), 11263–11273 (2018).
- <sup>86</sup>R. A. John *et al.*, "Self healable neuromorphic memtransistor elements for decentralized sensory signal processing in robotics," *Nat. Commun.* **11**(1), 4030 (2020).
- <sup>87</sup>M. R. Mahmoodi, D. B. Strukov, and O. Kavehei, "Experimental demonstrations of security primitives with nonvolatile memories," *IEEE Trans. Electron Devices* **66**(12), 5050–5059 (2019).
- <sup>88</sup>M. Baldo, O. Melnic *et al.*, "Modeling of virgin state and forming operation in embedded phase change memory (PCM)," in *2020 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2020), pp. 13.3.1–13.3.4.
- <sup>89</sup>G. Pedretti and D. Ielmini, "In-memory computing with resistive memory circuits: Status and outlook," *Electronics* **10**(9), 1063 (2021).
- <sup>90</sup>S. Wang, Z. Sun *et al.*, "Optimization schemes for in-memory linear regression circuit with memristor arrays," *IEEE Trans. Circuits Syst., I* **68**(12), 4900–4909 (2021).
- <sup>91</sup>O. Krestinskaya, A. P. James, and L. O. Chua, "Neuromemristive circuits for edge computing: A review," *IEEE Trans. Neural Networks Learn. Syst.* **31**(1), 4–23 (2020).
- <sup>92</sup>L. M. Vaquero and L. Rodero-Merino, "Finding your way in the fog: Towards a comprehensive definition of fog computing," *ACM SIGCOMM Comput. Commun. Rev.* **44**(5), 27–32 (2014).
- <sup>93</sup>Z. Sun, G. Pedretti *et al.*, "Time complexity of in-memory solution of linear systems," *IEEE Trans. Electron Devices* **67**(7), 2945–2951 (2020).
- <sup>94</sup>Z. Sun and R. Huang, "Time complexity of in-memory matrix-vector multiplication," *IEEE Trans. Circuits Syst., II* **68**(8), 2785–2789 (2021).
- <sup>95</sup>G. Pedretti, P. Mannocci *et al.*, "A spiking recurrent neural network with phase-change memory neurons and synapses for the accelerated solution of constraint satisfaction problems," *IEEE J. Explor. Solid-State Comput. Devices Circuits* **6**(1), 89–97 (2020).
- <sup>96</sup>J. Borghetti, Z. Li *et al.*, "A hybrid nanomemristor/transistor logic circuit capable of self-programming," *Proc. Natl. Acad. Sci. U. S. A.* **106**(6), 1699–1703 (2009).

- <sup>97</sup>Z. Sun, E. Ambrosi *et al.*, “Logic computing with stateful neural networks of resistive switches,” *Adv. Mater.* **30**(38), 1802554 (2018).
- <sup>98</sup>G. W. Burr, R. M. Shelby *et al.*, “Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element,” *IEEE Trans. Electron Devices* **62**(11), 3498–3507 (2015).
- <sup>99</sup>S. Ambrogio, P. Narayanan *et al.*, “Equivalent-accuracy accelerated neural-network training using analogue memory,” *Nature* **558**(7708), 60–67 (2018).
- <sup>100</sup>M. Prezioso, F. Merrih-Bayat *et al.*, “Training and operation of an integrated neuromorphic network based on metal-oxide memristors,” *Nature* **521**(7550), 61–64 (2015).
- <sup>101</sup>G. Pedretti, V. Milo *et al.*, “Memristive neural network for on-line learning and tracking with brain-inspired spike timing dependent plasticity,” *Sci. Rep.* **7**(1), 5288 (2017).
- <sup>102</sup>V. Milo, G. Pedretti *et al.*, “Demonstration of hybrid CMOS/RRAM neural networks with spike time/rate-dependent plasticity,” in *2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016), pp. 16.8.1–16.8.4.
- <sup>103</sup>M. Farronato, P. Mannonci *et al.*, “Reservoir computing with charge-trap memory based on a MoS<sub>2</sub> channel for neuromorphic engineering,” *Adv. Mater.* (published online) (2022).
- <sup>104</sup>R. Carboni and D. Ielmini, “Stochastic memory devices for security and computing,” *Adv. Electron. Mater.* **5**(9), 1900198 (2019).
- <sup>105</sup>G. Indiveri, B. Linares-Barranco *et al.*, “Neuromorphic silicon neuron circuits,” *Front. Neurosci.* **5**, 73 (2011).
- <sup>106</sup>F. Cai, S. Kumar *et al.*, “Power-efficient combinatorial optimization using intrinsic noise in memristor Hopfield neural networks,” *Nat. Electron.* **3**(7), 409–418 (2020).
- <sup>107</sup>S. N. Truong and K.-S. Min, “New memristor-based crossbar array architecture with 50-% area reduction and 48-% power saving for matrix-vector multiplication of analog neuromorphic computing,” *J. Semicond. Technol. Sci.* **14**(3), 356–363 (2014).
- <sup>108</sup>J. J. Yang, D. B. Strukov, and D. R. Stewart, “Memristive devices for computing,” *Nat. Nanotechnol.* **8**(1), 13–24 (2013).
- <sup>109</sup>D. Kau, S. Tang *et al.*, “A stackable cross point Phase Change Memory,” in *2009 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2009), pp. 1–4.
- <sup>110</sup>M.-C. Hsieh, Y.-C. Liao *et al.*, “Ultra high density 3D via RRAM in pure 28nm CMOS process,” in *2013 IEEE International Electron Devices Meeting (IEDM)* (IEEE, Washington, DC, 2013), pp. 10.3.1–10.3.4.
- <sup>111</sup>S. Balatti, S. Ambrogio *et al.*, “Set variability and failure induced by complementary switching in bipolar RRAM,” *IEEE Electron Device Lett.* **34**(7), 861–863 (2013).
- <sup>112</sup>S. Ambrogio, S. Balatti *et al.*, “Statistical fluctuations in HfO<sub>x</sub> resistive-switching memory: Part II—random telegraph noise,” *IEEE Trans. Electron Devices* **61**(8), 2920–2927 (2014).
- <sup>113</sup>S. Ambrogio *et al.*, “Statistical fluctuations in HfO<sub>x</sub> resistive-switching memory: Part I—set/reset variability,” *IEEE Trans. Electron Devices* **61**(8), 2912–2919 (2014).
- <sup>114</sup>N. Lepri, M. Baldo *et al.*, “Modeling and compensation of IR drop in crosspoint accelerators of neural networks,” *IEEE Trans. Electron Devices* **69**(3), 1575–1581 (2022).
- <sup>115</sup>B. Govoreanu, A. Redolfi *et al.*, “Vacancy-modulated conductive oxide resistive RAM (VMCO-RRAM): An area-scalable switching current, self-compliant, highly nonlinear and wide on/off-window resistive switching cell,” in *2013 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2013), pp. 10.2.1–10.2.4.
- <sup>116</sup>F. Zhang and M. Hu, “Mitigate parasitic resistance in resistive crossbar-based convolutional neural networks,” *ACM J. Emerging Technol. Comput. Syst.* **16**(3), 1–20 (2020).
- <sup>117</sup>C. Mackin, M. J. Rasch *et al.*, “Optimised weight programming for analogue memory-based deep neural networks,” *Nat. Commun.* **13**(1), 3765 (2022).
- <sup>118</sup>D. Joksas, E. Wang *et al.*, “Nonideality-aware training for accurate and robust low-power memristive neural networks,” *Adv. Sci.* **9**(17), 2105784 (2022).
- <sup>119</sup>D. Joksas, P. Freitas *et al.*, “Committee machines—a universal method to deal with non-idealities in memristor-based neural networks,” *Nat. Commun.* **11**(1), 4273 (2020).
- <sup>120</sup>M. L. Gallo, S. R. Nandakumar *et al.*, “Precision of bit slicing with in-memory computing based on analog phase-change memory crossbars,” *Neuromorphic Comput. Eng.* **2**(1), 014009 (2022).
- <sup>121</sup>F. L. Aguirre, N. M. Gomez *et al.*, “Minimization of the line resistance impact on memdiode-based simulations of multilayer perceptron arrays applied to pattern recognition,” *J. Low Power Electron. Appl.* **11**(1), 9 (2021).
- <sup>122</sup>N. Lepri, A. Glukhov, and D. Ielmini, “Mitigating read-program variation and IR drop by circuit architecture in RRAM-based neural network accelerators,” in *2022 IEEE International Reliability Physics Symposium (IRPS)* (IEEE, Dallas, TX, 2022), pp. 3C.2–1–3C.2–6.
- <sup>123</sup>A. Flocke and T. G. Noll, “Fundamental analysis of resistive nano-crossbars for the use in hybrid nano/CMOS-memory,” in *ESSCIRC 2007 - 33rd European Solid-State Circuits Conference* (IEEE, Muenchen, Germany, 2007), pp. 328–331.
- <sup>124</sup>F. Li, X. Yang *et al.*, “Evaluation of SiO<sub>2</sub> antifuse in a 3D-OTP memory,” *IEEE Trans. Device Mater. Reliab.* **4**(3), 416–421 (2004).
- <sup>125</sup>G. W. Burr, R. S. Shenoy *et al.*, “Access devices for 3D crosspoint memory,” *J. Vac. Sci. Technol. B* **32**(4), 040802 (2014).
- <sup>126</sup>M.-J. Lee, D. Lee *et al.*, “A plasma-treated chalcogenide switch device for stackable scalable 3D nanoscale memory,” *Nat. Commun.* **4**(1), 2629 (2013).
- <sup>127</sup>M. Hu, J. P. Strachan *et al.*, “Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication,” in *Proceedings of the 53rd Annual Design Automation Conference* (ACM, Austin, TX, 2016), pp. 1–6.
- <sup>128</sup>D. Ielmini, F. Nardi, and C. Cagli, “Physical models of size-dependent nanofilament formation and rupture in NiO resistive switching memories,” *Nanotechnology* **22**(25), 254022 (2011).
- <sup>129</sup>V. Milo, C. Zambelli *et al.*, “Multilevel HfO<sub>2</sub>-based RRAM devices for low-power neuromorphic networks,” *APL Mater.* **7**(8), 081120 (2019).
- <sup>130</sup>K. Gopalakrishnan, R. S. Shenoy *et al.*, “Highly-scalable novel access device based on mixed ionic electronic conduction (MIEC) materials for high density phase change memory (PCM) arrays,” in *2010 Symposium on VLSI Technology* (IEEE, Honolulu, 2010), pp. 205–206.
- <sup>131</sup>M. Son, J. Lee *et al.*, “Excellent selector characteristics of nanoscale VO<sub>2</sub> for high-density bipolar ReRAM applications,” *IEEE Electron Device Lett.* **32**(11), 1579–1581 (2011).
- <sup>132</sup>V. Milo, A. Glukhov *et al.*, “Accurate program/verify schemes of resistive switching memory (RRAM) for in-memory neural network circuits,” *IEEE Trans. Electron Devices* **68**(8), 3832–3837 (2021).
- <sup>133</sup>Y.-C. Luo, A. Lu *et al.*, “Design of non-volatile capacitive crossbar array for in-memory computing,” in *2021 IEEE International Memory Workshop (IMW)* (IEEE, 2021), pp. 1–4.
- <sup>134</sup>C. Li, M. Hu *et al.*, “Analogue signal and image processing with large memristor crossbars,” *Nat. Electron.* **1**(1), 52–59 (2018).
- <sup>135</sup>M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, and J. P. Strachan, *Adv. Mater.* **30**, 1705914 (2018).
- <sup>136</sup>P. Narayanan, S. Ambrogio *et al.*, “Fully on-chip MAC at 14 nm enabled by accurate row-wise programming of PCM-based weights and parallel vector-transport in duration-format,” *IEEE Trans. Electron Devices* **68**(12), 6629–6636 (2021).
- <sup>137</sup>S. N. Truong, S. Shin *et al.*, “New twin crossbar architecture of binary memristors for low-power image recognition with discrete cosine transform,” *IEEE Trans. Nanotechnol.* **14**(6), 1104–1111 (2015).
- <sup>138</sup>Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature* **521**(7553), 436–444 (2015).
- <sup>139</sup>B. Fleischer, S. Shukla *et al.*, “A scalable multi-TeraOPS deep learning processor core for AI training and inference,” in *2018 IEEE Symposium on VLSI Circuits* (IEEE, 2018), pp. 35–36.
- <sup>140</sup>H. Cai, Y. Guo *et al.*, “Proposal of analog in-memory computing with magnified tunnel magnetoresistance ratio and universal STT-MRAM cell,” *IEEE Trans. Circuits Syst., I* **69**(4), 1519–1531 (2022).
- <sup>141</sup>J. Chen, S. Wen *et al.*, “Highly parallelized memristive binary neural network,” *Neural Networks* **144**, 565–572 (2021).
- <sup>142</sup>X. Sun, S. Yin *et al.*, “XNOR-RRAM: A scalable and parallel resistive synaptic architecture for binary neural networks,” in *2018 Design, Automation and Test*

- in *Europe Conference and Exhibition (DATE)* (IEEE, Dresden, Germany, 2018), pp. 1423–1428.
- <sup>143</sup>W. Wan, R. Kubendran *et al.*, “A compute-in-memory chip based on resistive random-access memory,” *Nature* **608**(7923), 504–512 (2022).
- <sup>144</sup>C. Li, D. Belkin *et al.*, “Efficient and self-adaptive in-situ learning in multilayer memristor neural networks,” *Nat. Commun.* **9**(1), 2385 (2018).
- <sup>145</sup>M. R. Mahmoodi, H. Kim *et al.*, “An analog neuro-optimizer with adaptable annealing based on  $64 \times 64$  OTIR crossbar circuit,” in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2019), pp. 14.7.1–14.7.4.
- <sup>146</sup>L. Deng, L. Liang *et al.*, “SemiMap: A semi-folded convolution mapping for speed-overhead balance on crossbars,” *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **39**(1), 117–130 (2020).
- <sup>147</sup>J.-s. Seo, J. Saikia *et al.*, “Digital versus analog artificial intelligence accelerators: Advances, trends, and emerging designs,” *IEEE Solid-State Circuits Mag.* **14**(3), 65–79 (2022).
- <sup>148</sup>W. Maass, “Noise as a resource for computation and learning in networks of spiking neurons,” *Proc. IEEE* **102**(5), 860–880 (2014).
- <sup>149</sup>J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proc. Natl. Acad. Sci. U. S. A.* **79**(8), 2554–2558 (1982).
- <sup>150</sup>C. D. Wright, P. Hosseini, and J. A. V. Diodado, “Beyond von-Neumann computing with nanoscale phase-change memory devices,” *Adv. Funct. Mater.* **23**(18), 2248–2254 (2013).
- <sup>151</sup>M. R. Mahmoodi, M. Prezioso, and D. B. Strukov, “Versatile stochastic dot product circuits based on nonvolatile memories for high performance neurocomputing and neurooptimization,” *Nat. Commun.* **10**(1), 5113 (2019).
- <sup>152</sup>M. N. Bojnordi and E. Ipek, “Memristive Boltzmann machine: A hardware accelerator for combinatorial optimization and deep learning,” in *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (IEEE, Barcelona, Spain, 2016), pp. 1–13.
- <sup>153</sup>T. Dalgaty, E. Esmanhotto *et al.*, “Ex situ transfer of Bayesian neural networks to resistive memory-based inference hardware,” *Adv. Intell. Syst.* **3**(8), 2000103 (2021).
- <sup>154</sup>T. Dalgaty, N. Castellani *et al.*, “In situ learning using intrinsic memristor variability via Markov chain Monte Carlo sampling,” *Nat. Electron.* **4**(2), 151–161 (2021).
- <sup>155</sup>Z. Sun, G. Pedretti *et al.*, “Solving matrix equations in one step with cross-point resistive arrays,” *Proc. Natl. Acad. Sci. U. S. A.* **116**(10), 4123–4128 (2019).
- <sup>156</sup>Z. Sun, E. Ambrosi *et al.*, “In-memory PageRank accelerator with a cross-point array of resistive memories,” *IEEE Trans. Electron Devices* **67**(4), 1466–1470 (2020).
- <sup>157</sup>Z. Sun, G. Pedretti *et al.*, “One-step regression and classification with cross-point resistive memory arrays,” *Sci. Adv.* **6**(5), eaay2378 (2020).
- <sup>158</sup>P. Mannonci, G. Pedretti *et al.*, “A universal, analog, in-memory computing primitive for linear algebra using memristors,” *IEEE Trans. Circuits Syst., I* **68**, 4889 (2021).
- <sup>159</sup>G. Zoppo, A. Korkmaz *et al.*, “Analog solutions of discrete Markov chains via memristor crossbars,” *IEEE Trans. Circuits Syst., I* **68**(12), 4910–4923 (2021).
- <sup>160</sup>Z. Sun, G. Pedretti *et al.*, “In-memory eigenvector computation in time  $O(1)$ ,” *Adv. Intell. Syst.* **2**(8), 2000042 (2020).
- <sup>161</sup>P. Gupta, A. Goel *et al.*, “WTF: The who to follow service at twitter,” in *Proceedings of the 22nd international conference on World Wide Web - WWW'13* (ACM Press, Rio de Janeiro, Brazil, 2013), pp. 505–514.
- <sup>162</sup>G. Pedretti, P. Mannonci *et al.*, “Redundancy and analog slicing for precise in-memory machine learning—Part II: Applications and benchmark,” *IEEE Trans. Electron Devices* **68**(9), 4379–4383 (2021).
- <sup>163</sup>P. Mannonci, E. Melacarne, and D. Ielmini, “An analogue in-memory ridge regression circuit with application to massive MIMO acceleration,” *IEEE J. Emerging Sel. Top. Circuits Systems* **12**, 952 (2022).
- <sup>164</sup>B. Feinberg, R. Wong *et al.*, “An analog preconditioner for solving linear systems,” in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (IEEE, Seoul, South Korea, 2021), pp. 761–774.
- <sup>165</sup>K. Pagiamtzis and A. Sheikholeslami, “Content-addressable memory (CAM) circuits and architectures: A tutorial and survey,” *IEEE J. Solid-State Circuits* **41**(3), 712–727 (2006).
- <sup>166</sup>I. Sourdis and D. Pnevmatikatos, “Pre-decoded CAMs for efficient and high-speed NIDS pattern matching,” in *12th Annual IEEE Symposium on Field-Programmable Custom Computing Machines* (IEEE, Napa, CA, 2004), pp. 258–267.
- <sup>167</sup>C. E. Graves, C. Li *et al.*, “In-memory computing with memristor content addressable memories for pattern matching,” *Adv. Mater.* **32**(37), 2003437 (2020).
- <sup>168</sup>R. Karam, R. Puri *et al.*, “Emerging trends in design and applications of memory-based computing and content-addressable memories,” *Proc. IEEE* **103**(8), 1311–1330 (2015).
- <sup>169</sup>A. J. McAuley and P. Francis, “Fast routing table lookup using CAMs,” in *Proceedings of the IEEE INFOCOM'93 The Conference on Computer Communications* (IEEE Computer Society Press, San Francisco, CA, 1993), pp. 1382–1391.
- <sup>170</sup>C. Li, C. E. Graves *et al.*, “Analog content-addressable memories with memristors,” *Nat. Commun.* **11**(1), 1638 (2020).
- <sup>171</sup>G. Pedretti, C. E. Graves *et al.*, “Tree-based machine learning performed in-memory with memristive analog CAM,” *Nat. Commun.* **12**(1), 5806 (2021).
- <sup>172</sup>H. Tsai, S. Ambrogio *et al.*, “Recent progress in analog memory-based accelerators for deep learning,” *J. Phys. D: Appl. Phys.* **51**(28), 283001 (2018).
- <sup>173</sup>S. Agarwal, S. J. Plimpton *et al.*, “Resistive memory device requirements for a neural algorithm accelerator,” in *2016 International Joint Conference on Neural Networks (IJCNN)* (IEEE, Vancouver, BC, 2016), pp. 929–938.
- <sup>174</sup>S. Yu, P.-Y. Chen *et al.*, “Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect,” in *2015 IEEE International Electron Devices Meeting (IEDM)* (IEEE, Washington, DC, 2015), pp. 17.3.1–17.3.4.
- <sup>175</sup>M. Farronato, M. Melegari *et al.*, “Low-current, highly linear synaptic memory device based on MoS<sub>2</sub> transistors for online training and inference,” in *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)* (IEEE, Incheon, Republic of Korea, 2022), pp. 1–4.
- <sup>176</sup>J. Woo, K. Moon *et al.*, “Improved synaptic behavior under identical pulses using AlO<sub>x</sub>/HfO<sub>2</sub> bilayer RRAM array for neuromorphic systems,” *IEEE Electron Device Lett.* **37**(8), 994–997 (2016).
- <sup>177</sup>S. H. Jo, T. Chang *et al.*, “Nanoscale memristor device as synapse in neuromorphic systems,” *Nano Lett.* **10**(4), 1297–1301 (2010).
- <sup>178</sup>J.-W. Jang, S. Park *et al.*, “Optimization of conductance change in Pr<sub>1-x</sub>Ca<sub>x</sub>MnO<sub>3</sub>-based synaptic devices for neuromorphic systems,” *IEEE Electron Device Lett.* **36**(5), 457–459 (2015).
- <sup>179</sup>S. Hao, X. Ji *et al.*, “Monolayer MoS<sub>2</sub>/WO<sub>3</sub> heterostructures with sulfur anion reservoirs as electronic synapses for neuromorphic computing,” *ACS Appl. Nano Mater.* **4**(2), 1766–1775 (2021).
- <sup>180</sup>C. J. McClellan, A. C. Yu *et al.*, “Vertical sidewall MoS<sub>2</sub> growth and transistors,” in *2019 Device Research Conference (DRC)* (IEEE, Ann Arbor, MI, 2019), pp. 65–66.
- <sup>181</sup>C. Mead, “Neuromorphic electronic systems,” *Proc. IEEE* **78**(10), 1629–1636 (1990).
- <sup>182</sup>G. Indiveri, F. Corradi, and N. Qiao, “Neuromorphic architectures for spiking deep neural networks,” in *2015 IEEE International Electron Devices Meeting (IEDM)* (IEEE, Washington, DC, 2015), pp. 4.2.1–4.2.4.
- <sup>183</sup>Y. Taigman, M. Yang *et al.*, “DeepFace: Closing the gap to human-level performance in face verification,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Columbus, OH, 2014), pp. 1701–1708.
- <sup>184</sup>G. Indiveri, B. Linares-Barranco *et al.*, “Integration of nanoscale memristor synapses in neuromorphic computing architectures,” *Nanotechnology* **24**(38), 384010 (2013).
- <sup>185</sup>C. Zamarreño-Ramos, L. A. Camuñas-Mesa *et al.*, “On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex,” *Front. Neurosci.* **5**, 26 (2011).
- <sup>186</sup>T. Serrano-Gotarredona, T. Masquelier *et al.*, “STDP and STDP variations with memristors for spiking neuromorphic learning systems,” *Front. Neurosci.* **7**, 2 (2013).
- <sup>187</sup>D. Ielmini, Z. Wang, and Y. Liu, “Brain-inspired computing via memory device physics,” *APL Mater.* **9**(5), 050702 (2021).
- <sup>188</sup>Y. D. Zhao, J. F. Kang, and D. Ielmini, “Materials challenges and opportunities for brain-inspired computing,” *MRS Bull.* **46**(10), 978–986 (2021).
- <sup>189</sup>D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (L. Erlbaum Associates, Mahwah, NJ, 2002).



- <sup>190</sup>T. V. P. Bliss and G. L. Collingridge, "A synaptic model of memory: Long-term potentiation in the hippocampus," *Nature* **361**(6407), 31–39 (1993).
- <sup>191</sup>H. Markram, J. Lübke *et al.*, "Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs," *Science* **275**(5297), 213–215 (1997).
- <sup>192</sup>A. Citri and R. C. Malenka, "Synaptic plasticity: Multiple forms, functions, and mechanisms," *Neuropsychopharmacology* **33**(1), 18–41 (2008).
- <sup>193</sup>T. Ohno, T. Hasegawa *et al.*, "Short-term plasticity and long-term potentiation mimicked in single inorganic synapses," *Nat. Mater.* **10**(8), 591–595 (2011).
- <sup>194</sup>G.-Q. Bi and M.-M. Poo, "Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type," *J. Neurosci.* **18**(24), 10464–10472 (1998).
- <sup>195</sup>M. Rahimi Azghadi, N. Iannella *et al.*, "Spike-based synaptic plasticity in silicon: Design, implementation, application, and challenges," *Proc. IEEE* **102**(5), 717–737 (2014).
- <sup>196</sup>L. F. Abbott and S. B. Nelson, "Synaptic plasticity: Taming the beast," *Nat. Neurosci.* **3**(S11), 1178–1183 (2000).
- <sup>197</sup>J.-P. Pfister, "Triplets of spikes in a model of spike timing-dependent plasticity," *J. Neurosci.* **26**(38), 9673–9682 (2006).
- <sup>198</sup>J. Gjorgjieva, C. Clopath *et al.*, "A triplet spike-timing-dependent plasticity model generalizes the Bienenstock–Cooper–Munro rule to higher-order spatiotemporal correlations," *Proc. Natl. Acad. Sci. U. S. A.* **108**(48), 19383–19388 (2011).
- <sup>199</sup>E. Bienenstock, L. Cooper, and P. Munro, "Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex," *J. Neurosci.* **2**(1), 32–48 (1982).
- <sup>200</sup>P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, "Rate, timing, and cooperativity jointly determine cortical synaptic plasticity," *Neuron* **32**(6), 1149–1164 (2001).
- <sup>201</sup>D. Kuzum, R. G. D. Jayasingh *et al.*, "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," *Nano Lett.* **12**(5), 2179–2186 (2012).
- <sup>202</sup>S. Ambrogio, N. Ciochini *et al.*, "Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses," *Front. Neurosci.* **10**, 56 (2016).
- <sup>203</sup>S. Ambrogio, S. Balatti *et al.*, "Spike-timing dependent plasticity in a transistor-selected resistive switching memory," *Nanotechnology* **24**(38), 384012 (2013).
- <sup>204</sup>Z. Wang, S. Ambrogio *et al.*, "A 2-transistor/1-resistor artificial synapse capable of communication and stochastic learning in neuromorphic systems," *Front. Neurosci.* **8**, 438 (2015).
- <sup>205</sup>S. Ambrogio, S. Balatti *et al.*, "Neuromorphic learning and recognition with one-transistor-one-resistor synapses and bistable metal oxide RRAM," *IEEE Trans. Electron Devices* **63**(4), 1508–1515 (2016).
- <sup>206</sup>V. Milo, G. Pedretti *et al.*, "A 4-transistors/1-resistor hybrid synapse based on resistive switching memory (RRAM) capable of spike-rate-dependent plasticity (SRDP)," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **26**(12), 2806–2815 (2018).
- <sup>207</sup>S. Kim, C. Du *et al.*, "Experimental demonstration of a second-order memristor and its ability to biorealistically implement synaptic plasticity," *Nano Lett.* **15**(3), 2203–2211 (2015).
- <sup>208</sup>G. Pedretti, V. Milo *et al.*, "Stochastic learning in neuromorphic hardware via spike timing dependent plasticity with RRAM synapses," *IEEE J. Emerging Sel. Top. Circuits Syst.* **8**(1), 77–85 (2018).
- <sup>209</sup>W. Wang, M. Wang *et al.*, "Surface diffusion-limited lifetime of silver and copper nanofilaments in resistive switching devices," *Nat. Commun.* **10**(1), 81 (2019).
- <sup>210</sup>Q. Hua, H. Wu *et al.*, "Threshold switching selectors: A threshold switching selector based on highly ordered Ag nanodots for X-point memory applications (Adv. Sci. 10/2019)," *Adv. Sci.* **6**(10), 1970058 (2019).
- <sup>211</sup>Y. Li, J. Tang *et al.*, "High-uniformity threshold switching HfO<sub>2</sub>-based selectors with patterned Ag nanodots," *Adv. Sci.* **7**(22), 2002251 (2020).
- <sup>212</sup>M. Wang, W. Wang *et al.*, "Enhancing the matrix addressing of flexible sensory arrays by a highly nonlinear threshold switch," *Adv. Mater.* **30**(33), 1802516 (2018).
- <sup>213</sup>J. Song, A. Prakash *et al.*, "Bidirectional threshold switching in engineered multilayer (Cu<sub>2</sub>O/Ag:Cu<sub>2</sub>O/Cu<sub>2</sub>O) stack for cross-point selector application," *Appl. Phys. Lett.* **107**(11), 113504 (2015).
- <sup>214</sup>R. Midya, Z. Wang *et al.*, "Anatomy of Ag/Hafnia-based selectors with 10<sup>10</sup> nonlinearity," *Adv. Mater.* **29**(12), 1604457 (2017).
- <sup>215</sup>E. Covi, W. Wang *et al.*, "Switching dynamics of Ag-based filamentary volatile resistive switching devices—Part I: Experimental characterization," *IEEE Trans. Electron Devices* **68**(9), 4335–4341 (2021).
- <sup>216</sup>Z. Wang, S. Joshi *et al.*, "Fully memristive neural networks for pattern classification with unsupervised learning," *Nat. Electron.* **1**(2), 137–145 (2018).
- <sup>217</sup>J. Zhu, Y. Yang *et al.*, "Ion gated synaptic transistors based on 2D van der Waals crystals with tunable diffusive dynamics," *Adv. Mater.* **30**(21), 1800195 (2018).
- <sup>218</sup>S. H. Sung, T. J. Kim *et al.*, "Simultaneous emulation of synaptic and intrinsic plasticity using a memristive synapse," *Nat. Commun.* **13**(1), 2811 (2022).
- <sup>219</sup>W. Wang, E. Covi *et al.*, "Neuromorphic motion detection and orientation selectivity by volatile resistive switching memories," *Adv. Intell. Syst.* **3**(4), 2000224 (2021).
- <sup>220</sup>H. B. Barlow, R. M. Hill, and W. R. Levick, "Retinal ganglion cells responding selectively to direction and speed of image motion in the rabbit," *J. Physiol.* **173**(3), 377 (1964).
- <sup>221</sup>J. Elstrott and M. B. Feller, "Vision and the establishment of direction-selectivity: A tale of two circuits," *Curr. Opin. Neurobiol.* **19**(3), 293–297 (2009).
- <sup>222</sup>M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Comput. Sci. Rev.* **3**(3), 127–149 (2009).
- <sup>223</sup>W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural Comput.* **14**(11), 2531–2560 (2002).
- <sup>224</sup>H. Jaeger, "The 'echo state' approach to analysing and training recurrent neural networks – with an Erratum note," Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, **148**(34), 13, 2001.
- <sup>225</sup>G. Tanaka, T. Yamane *et al.*, "Recent advances in physical reservoir computing: A review," *Neural Networks* **115**, 100–123 (2019).
- <sup>226</sup>K. Nakajima, "Physical reservoir computing—An introductory perspective," *Jpn. J. Appl. Phys.* **59**(6), 060501 (2020).
- <sup>227</sup>H. Jaeger, "A tutorial on training recurrent neural networks, covering BPTT, RTRL, EKF and the 'echo state network' approach," 2013, available at <https://www.ai.rug.nl/minds/uploads/ESNTutorialRev.pdf>.
- <sup>228</sup>C. Du, F. Cai *et al.*, "Reservoir computing using dynamic memristors for temporal information processing," *Nat. Commun.* **8**(1), 2204 (2017).
- <sup>229</sup>J. Moon, W. Ma *et al.*, "Temporal data classification and forecasting using a memristor-based reservoir computing system," *Nat. Electron.* **2**(10), 480–487 (2019).
- <sup>230</sup>S. Stepney, "The neglected pillar of material computation," *Physica D* **237**(9), 1157–1164 (2008).
- <sup>231</sup>M. Dale, S. Stepney *et al.*, "Reservoir computing in materio: A computational framework for in materio computing," in *2017 International Joint Conference on Neural Networks (IJCNN)* (IEEE, Anchorage, AK, 2017), pp. 2178–2185.
- <sup>232</sup>H.-C. Ruiz Euler, M. N. Boon *et al.*, "A deep-learning approach to realizing functionality in nanoelectronic devices," *Nat. Nanotechnol.* **15**(12), 992–998 (2020).
- <sup>233</sup>A. M. Shen, C.-L. Chen *et al.*, "Analog neuromorphic module based on carbon nanotube synapses," *ACS Nano* **7**(7), 6117–6122 (2013).
- <sup>234</sup>H. Tanaka, M. Akai-Kasaya *et al.*, "A molecular neuromorphic network device consisting of single-walled carbon nanotubes complexed with polyoxometalate," *Nat. Commun.* **9**(1), 2693 (2018).
- <sup>235</sup>G. Milano, G. Pedretti *et al.*, "Brain-inspired structural plasticity through reweighting and rewiring in multi-terminal self-organizing memristive nanowire networks," *Adv. Intell. Syst.* **2**(8), 2000096 (2020).
- <sup>236</sup>G. Milano, G. Pedretti *et al.*, "In materia reservoir computing with a fully memristive architecture based on self-organizing nanowire networks," *Nat. Mater.* **21**(2), 195–202 (2022).
- <sup>237</sup>K. Fu, R. Zhu *et al.*, "Reservoir computing with neuromemristive nanowire networks," in *2020 International Joint Conference on Neural Networks (IJCNN)* (IEEE, Glasgow, UK, 2020), pp. 1–8.

- <sup>238</sup>S. K. Bose, C. P. Lawrence *et al.*, "Evolution of a designless nanoparticle network into reconfigurable Boolean logic," *Nat. Nanotechnol.* **10**(12), 1048–1052 (2015).
- <sup>239</sup>T. Chen, J. van Gelder, B. van de Ven *et al.*, "Classification with a disordered dopant-atom network in silicon," *Nature* **577**, 341–345 (2020).
- <sup>240</sup>B. Kiraly, E. J. Knol *et al.*, "An atomic Boltzmann machine capable of self-adaption," *Nat. Nanotechnol.* **16**(4), 414–420 (2021).
- <sup>241</sup>L. Hong, H. Tanaka, and T. Ogawa, "Rectification direction inversion in a phosphododecamolybdc acid/single-walled carbon nanotube junction," *J. Mater. Chem. C* **1**(6), 1137–1143 (2013).
- <sup>242</sup>X. Guo, D.-J. Guo *et al.*, "Using phosphomolybdc acid ( $H_3PMo_{12}O_{40}$ ) to efficiently enhance the electrocatalytic activity and CO-tolerance of platinum nanoparticles supported on multi-walled carbon nanotubes catalyst in acidic medium," *J. Electroanal. Chem.* **638**(1), 167–172 (2010).
- <sup>243</sup>M. Sharifkhani and M. Sachdev, "SRAM cell stability: A dynamic perspective," *IEEE J. Solid-State Circuits* **44**(2), 609–619 (2009).
- <sup>244</sup>W. Shin, J. Choi *et al.*, "DRAM-latency optimization inspired by relationship between row-access time and refresh timing," *IEEE Trans. Comput.* **65**(10), 3027–3040 (2016).
- <sup>245</sup>D. Ielmini, "Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling," *Semicond. Sci. Technol.* **31**(6), 063002 (2016).
- <sup>246</sup>F. Zahoor, T. Z. Azni Zulkifli, and F. A. Khanday, "Resistive random access memory (RRAM): An overview of materials, switching mechanism, performance, multilevel cell (MLC) storage, modeling, and applications," *Nanoscale Res. Lett.* **15**(1), 90 (2020).
- <sup>247</sup>N. K. Upadhyay, H. Jiang *et al.*, "Emerging memory devices for neuromorphic computing," *Adv. Mater. Technol.* **4**(4), 1800589 (2019).
- <sup>248</sup>W. Zhang, B. Gao *et al.*, "Neuro-inspired computing chips," *Nat. Electron.* **3**(7), 371–382 (2020).
- <sup>249</sup>M. Le Gallo, A. Sebastian *et al.*, "Mixed-precision in-memory computing," *Nat. Electron.* **1**(4), 246–253 (2018).
- <sup>250</sup>M. A. Zidan, Y. Jeong *et al.*, "A general memristor-based partial differential equation solver," *Nat. Electron.* **1**(7), 411–420 (2018).
- <sup>251</sup>R. Cai, A. Ren *et al.*, "Memristor-based discrete Fourier transform for improving performance and energy efficiency," in *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (IEEE, Pittsburgh, PA, 2016), pp. 643–648.
- <sup>252</sup>G. Yuan, C. Ding *et al.*, "Memristor crossbar-based ultra-efficient next-generation baseband processors," in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)* (IEEE, Boston, MA, 2017), pp. 1121–1124.
- <sup>253</sup>W. Wang, G. Pedretti *et al.*, "Computing of temporal information in spiking neural networks with ReRAM synapses," *Faraday Discuss.* **213**, 453–469 (2019).
- <sup>254</sup>J. J. Hopfield and D. W. Tank, "'Neural' computation of decisions in optimization problems," *Biol. Cybern.* **52**, 141–152 (1985).
- <sup>255</sup>S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science* **220**(4598), 671–680 (1983).
- <sup>256</sup>C. Voudouris and E. P. K. Tsang, "Guided local search," in *Handbook of Metaheuristics*, edited by F. Glover and G. A. Kochenberger (Springer, Boston, MA, 2003), pp. 185–218.
- <sup>257</sup>S. K. Gonugondla, C. Sakr *et al.*, "Fundamental limits on the precision of in-memory architectures," in *Proceedings of the 39th International Conference on Computer-Aided Design* (ACM, 2020), pp. 1–9.
- <sup>258</sup>M. E. Fouda, A. M. Eltwail, and F. Kurdahi, "Modeling and analysis of passive switching crossbar arrays," *IEEE Trans. Circuits Syst., I* **65**(1), 270–282 (2018).
- <sup>259</sup>S. Agarwal, R. L. Schiek, and M. J. Marinella, "Compensating for parasitic voltage drops in resistive memory arrays," in *2017 IEEE International Memory Workshop (IMW)* (IEEE, 2017), pp. 1–4.
- <sup>260</sup>Y. Luo, S. Wang *et al.*, "Modeling and mitigating the interconnect resistance issue in analog RRAM matrix computing circuits," *IEEE Trans. Circuits Syst., I* **69**(11), 4367–4380 (2022).
- <sup>261</sup>B. Liu, H. Li *et al.*, "Reduction and IR-drop compensations techniques for reliable neuromorphic computing systems," in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (IEEE, 2014), pp. 63–70.
- <sup>262</sup>M. E. Fouda, S. Lee *et al.*, "IR-QNN framework: An IR drop-aware offline training of quantized crossbar arrays," *IEEE Access* **8**, 228392–228408 (2020).
- <sup>263</sup>G. Krishnan, S. K. Mandal *et al.*, "Interconnect-aware area and energy optimization for in-memory acceleration of DNNs," *IEEE Des. Test* **37**(6), 79–87 (2020).
- <sup>264</sup>C. Villa, D. Mills *et al.*, "A 45nm 1Gb 1.8V phase-change memory," in *2010 IEEE International Solid-State Circuits Conference - (ISSCC)* (IEEE, San Francisco, CA, 2010), pp. 270–271.
- <sup>265</sup>T.-Y. Liu, T. H. Yan *et al.*, "A 130.7mm<sup>2</sup> 2-layer 32Gb ReRAM memory device in 24nm technology," *IEEE J. Solid-State Circuits* **49**(1), 140–153 (2014).
- <sup>266</sup>M. Guo, J. Jiang *et al.*, "Flexible robust and high-density FeRAM from array of organic ferroelectric nano-lamellae by self-assembly," *Adv. Sci.* **6**(6), 1801931 (2019).
- <sup>267</sup>S. Aggarwal, K. Nagel *et al.*, "Demonstration of a reliable 1 Gb standalone spin-transfer torque MRAM for industrial applications," in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2019), pp. 2.1.1–2.1.4.
- <sup>268</sup>B. Yuan, X. Liang *et al.*, "150 nm × 200 nm cross-point hexagonal boron nitride-based memristors," *Adv. Electron. Mater.* **6**(12), 1900115 (2020).
- <sup>269</sup>N. Verma, H. Jia *et al.*, "In-memory computing: Advances and prospects," *IEEE Solid-State Circuits Mag.* **11**(3), 43–55 (2019).
- <sup>270</sup>Y. Cassuto, S. Kvatinsky, and E. Yaakobi, "Sneak-path constraints in memristor crossbar arrays," in *2013 IEEE International Symposium on Information Theory* (IEEE, Istanbul, Turkey, 2013), pp. 156–160.
- <sup>271</sup>D. Kadetotad, Z. Xu *et al.*, "Parallel architecture with resistive crosspoint array for dictionary learning acceleration," *IEEE J. Emerging Sel. Top. Circuits Syst.* **5**(2), 194–204 (2015).
- <sup>272</sup>Z. Xuan and Y. Kang, "High-efficiency data conversion interface for reconfigurable function-in-memory computing," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **30**(9), 1193–1206 (2022).
- <sup>273</sup>I. Chakraborty, M. Ali *et al.*, "Resistive crossbars as approximate hardware building blocks for machine learning: Opportunities and challenges," *Proc. IEEE* **108**(12), 2276–2310 (2020).
- <sup>274</sup>Y. Zha and J. Li, "Reconfigurable in-memory computing with resistive memory crossbar," in *Proceedings of the 35th International Conference on Computer-Aided Design* (ACM, Austin, TX, 2016), pp. 1–8.
- <sup>275</sup>W. Jiang, Q. Lou *et al.*, "Device-circuit-architecture co-exploration for computing-in-memory neural accelerators," *IEEE Trans. Comput.* **70**(4), 595–605 (2021).
- <sup>276</sup>P.-Y. Chen, X. Peng, and S. Yu, "NeuroSim+: An integrated device-to-algorithm framework for benchmarking synaptic devices and array architectures," in *2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, CA, 2017), pp. 6.1.1–6.1.4.
- <sup>277</sup>S. Misailovic, M. Carbin *et al.*, "Chisel: Reliability- and accuracy-aware optimization of approximate computational kernels," *ACM SIGPLAN Not.* **49**(10), 309–328 (2014).
- <sup>278</sup>S. Achour, R. Sarpeshkar, and M. C. Rinard, "Configuration synthesis for programmable analog devices with Arco," *ACM SIGPLAN Not.* **51**(6), 177–193 (2016).
- <sup>279</sup>S. Achour and M. Rinard, "Noise-aware dynamical system compilation for analog devices with Legno," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems* (ACM, Lausanne, Switzerland, 2020), pp. 149–166.