

# Fully-Parallel 2-Terminal Update Scheme for Tensor Product in ECRAM Arrays

M. Porzani, L. Micheletti, P. Porta, S. Ricci, F. Carletti, M. Farronato, D. Ielmini

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano and IU.NET 20133 Milan, Italy

Email: matteo.porzani@polimi.it, daniele.ielmini@polimi.it

**Abstract**—We demonstrate fully-parallel tensor product in electrochemical random access memory (ECRAM) arrays enabled by multi-terminal operation to modulate ion injection into the metal-oxide channel. We show extensive characterization of the combined drain and gate pulsed response with a compact physics-based modeling to describe conductance update as a function of gate/drain voltages. We demonstrate fully-parallel product (AND) operation within the ECRAM array and accurate two-terminal weight update based on stochastic product, which paves the way for ECRAM-based hardware training accelerators. Finally, a realistic ECRAM model was calibrated on experimental data and used to simulate training of large convolutional neural networks (CNNs).

**Index Terms**—Electrochemical random-access memory (ECRAM), in-memory computing, tensor product, neural network training, artificial intelligence (AI)

## I. INTRODUCTION

Deep neural networks (DNNs) provide the backbone of several artificial intelligence (AI) functions, such as image recognition and natural language processing [1]. Training of modern DNNs requires massive computational resources and energy consumption [2]. To achieve energy-efficient AI training, novel circuit architectures, such as near-memory and in-memory computing (IMC) have recently been proposed [3]. In particular, IMC has been shown to accelerate AI training by enabling matrix vector multiplication (MVM) and weight update within memory arrays with extremely high parallelism and low power [4]. However, IMC-based training accelerators require precise and efficient devices capable of parallel weight update via tensor product (TP) [5].

The electrochemical random access memory (ECRAM) is a novel 3-terminal resistive memory device featuring a metal-oxide stack including a conductive channel, a solid-state electrolyte for ionic transport, and a reservoir layer (Fig. 1a) [6]. Information is stored in the channel conductance  $G$ , which can be controlled by voltage-driven ionic migration across the electrolyte layer [7]. ECRAM displays excellent properties, such as low-power operation, multilevel states and linear/symmetric response [8], which provide the basis for stochastic weight update [9] with advanced bias schemes [10]. 2-terminal (2-t) operation allows for tight control of ion injection and channel conductance [11]. However, a systematic characterization of ECRAM 2-t response is still missing.

This project has received funding from the European Research Council under grant no. 101054098

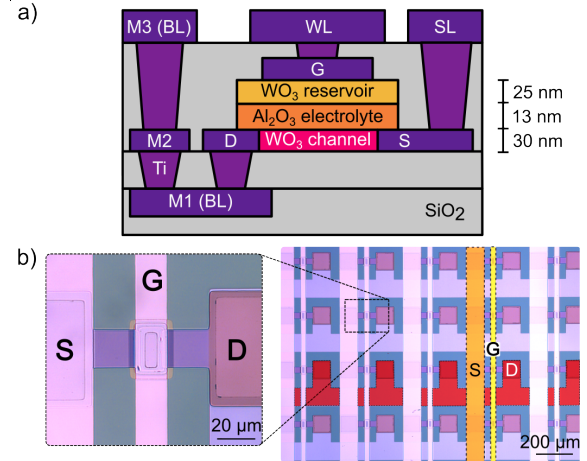


Fig. 1. ECRAM cell. (a) Schematic cross-section of ECRAM device based array, with a stack made of channel, electrolyte, and reservoir layers. (b) Optical microscope photography of an ECRAM array and detail on a single cell.

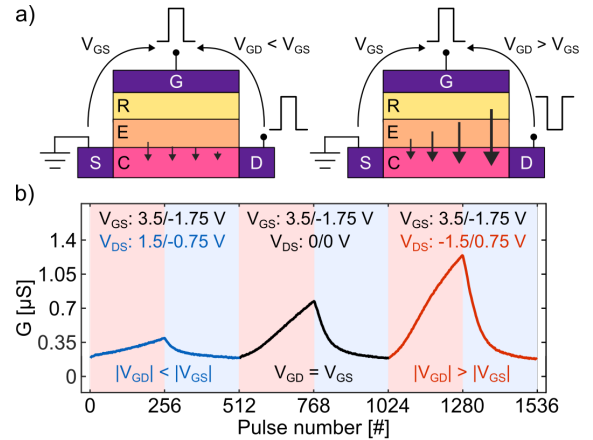


Fig. 2. 2-terminal (2-t) programming of ECRAM. (a) For a given  $V_{GS}$  pulse, a coincident  $V_{DS}$  pulse can inhibit or enhance ion migration in the oxide. (b) Switching characteristics of ECRAM device with 2-t scheme.

This study shows fully-parallel in-memory TP and stochastic product in metal-oxide-based ECRAM arrays (Fig. 1b), based on extensive characterization of the 2-t pulse response of ECRAM. The role of the drain and gate terminals is studied by a compact physics-based model, allowing to identify the most suitable bias voltages for selective weight update. Finally,

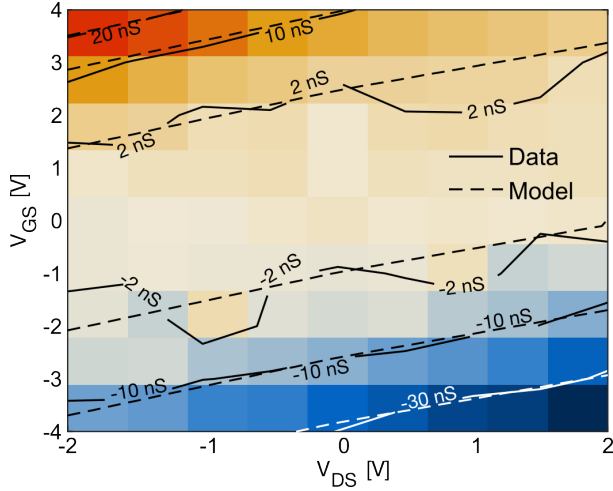


Fig. 3. Measured  $\Delta G$  as a function of  $V_{GS}$  and  $V_{DS}$  for pulsewidth  $t_p = 50$  ms and with baseline  $G_0 = 1 \mu\text{S}$ . Calculations from Eq. (1) are also shown, indicating good correlation.

experimental characteristics are used to simulate the training of a LeNet5 convolutional neural network (CNN) with ECRAM resistive processing unit (RPU).

## II. ECRAM CHARACTERIZATION AND MODELING

Metal-oxide ECRAM devices were fabricated with  $\text{WO}_3$  channel,  $\text{Al}_2\text{O}_3$  electrolyte layer and  $\text{WO}_3$  reservoir. Fig. 2a presents the 2-t ECRAM programming working principle based on coincident pulses. A positive gate voltage pulse induces the migration of oxygen vacancies (VOs) into the channel, thus resulting in the potentiation of ECRAM conductance  $G$ . The simultaneous application of a negative drain voltage pulse can locally enhance ion injection in the channel, thus causing higher potentiation. Conversely, a positive drain voltage pulse can locally decrease ion injection in the channel, thus causing lower potentiation. Similar 2-t operation holds for the case of depression pulses, under a negative  $V_{GS}$ .

Fig. 2b shows the programming characteristics of the ECRAM device under a sequence of pulses of pulsewidth  $t_p = 25$  ms for various  $V_{GS}$  and  $V_{DS}$ . The results indicate control of the potentiation/depression depending on the combination of voltage amplitudes at the gate and drain terminals, which is the key for parallel array-level operations [10]. This is further demonstrated by Fig. 3, showing the measured conductance change  $\Delta G$  as a function of  $V_{GS}$  and  $V_{DS}$  applied to an ECRAM cell from an initial conductance  $G_0 = 1 \mu\text{S}$ .

To describe the evolution of  $\Delta G$ , we developed a compact physics-based model where the conductance change  $\Delta G$  is given as a function of the nonlinear drift-diffusion of VOs across the electrolyte layer, namely:

$$\Delta G = AG_0 t_p \sinh\left(\frac{V_{GS} - \gamma V_{DS}}{V_0}\right) \quad (1)$$

where  $A$  and  $V_0$  are constants and  $\gamma$  is a scaling factor that accounts for the electrostatic control of the drain over the electrolyte. Fig. 4a summarizes the main equations describing

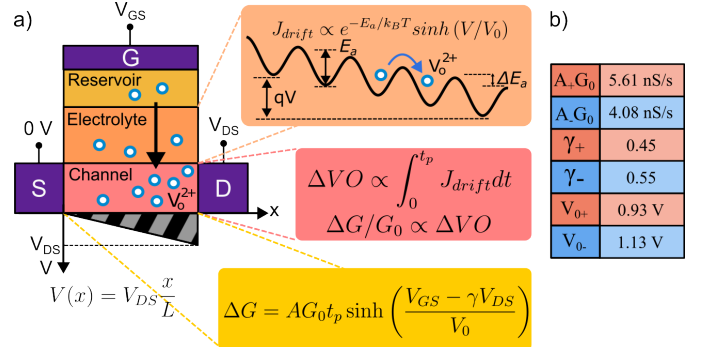


Fig. 4. 2D compact modeling of ECRAM devices. (a) Voltage-driven oxygen vacancy (VO) flux resulting in a modulation of VO concentration in the channel, which leads to a  $\Delta G$  update. (b) Model parameters for positive and negative  $\Delta G$ .

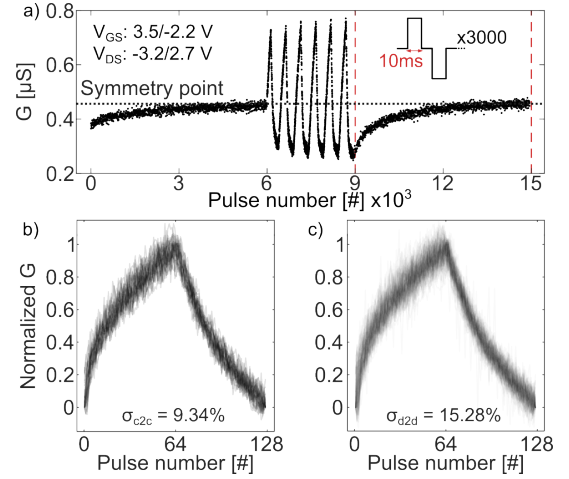


Fig. 5. Conductance switching characteristics of ECRAM with 2-t programming. (a) Set of 256 up/256 down pulses in between alternating up/down pulses. Convergence to symmetry point is shown. (b) Cycle-to-cycle and (c) device-to-device normalized responses.

ECRAM programming [12], namely a transport equation for VOs based on ion injection into the channel layer and linearized  $\Delta G$  update. Fig. 4b illustrates the model parameters extracted for positive and negative  $\Delta G$ . The measured  $\Delta G$  in Fig. 3 is well described by Eq. (1), thus supporting the accuracy of the compact modeling of 2-t operated ECRAM.

For accurate TP, it is necessary to first initialize the ECRAM cells in a reference conductance  $G_0$ . This is chosen as the *symmetry point* (SP), which is reached by a suitable sequence of potentiation/depression pulses [13]. Fig. 5a shows the ECRAM conductance characteristic obtained with the 2-t programming scheme, featuring (i) a first phase of 3000 alternating pulse pairs converging to the SP, (ii) a second phase of 6 trains of 256 pairs of potentiation/depression pulses, and (iii) a third phase to re-establish the SP state. Fig. 5b shows 48 consecutive cycles of 64 up/64 down pulses, evidencing good linearity, large memory window and relatively low cycle-to-cycle variability with a relative standard deviation of 9.34%. Fig. 5c shows the same measurement for a batch of several

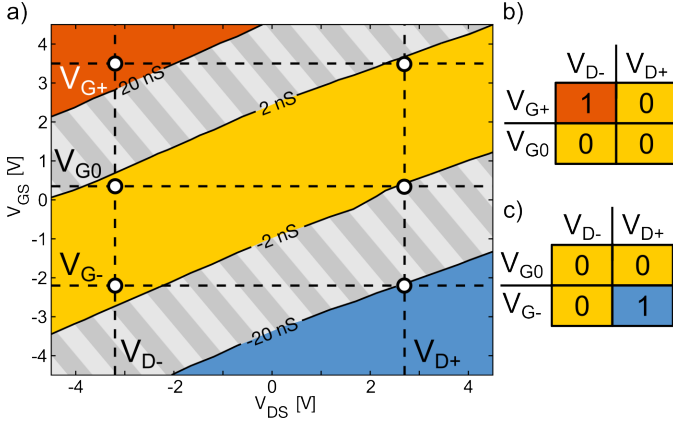


Fig. 6. Summary of 2-t ECRAM operation. (a) Calculated  $\Delta G$  as a function of  $V_{GS}$  and  $V_{DS}$ . The non-linear Eq. (1) allows to select two  $V_{DS}$  values and three  $V_{GS}$  values, so that their combined effect results in significant or negligible potentiation/depression. (b) Table of the truth for AND operation in the potentiation regime. (c) Same as (b), but for the depression regime.

devices in different arrays, evidencing a relatively low device-to-device variability with a relative standard deviation of 15.28%.

### III. DEMONSTRATION OF TP AT ARRAY LEVEL

The exponential voltage dependence in Eq. (1) allows to implement a 2-t pulse-coincidence update scheme based on suitable gate/drain pulses. This is illustrated in Fig. 6a, showing the calculated  $\Delta G$  as a function of  $V_{GS}$  and  $V_{DS}$ . From this plot, it is possible to identify 3 values of  $V_{GS}$ , namely,  $V_{G+}$ ,  $V_{G0}$  and  $V_{G-}$ , and two values of  $V_{DS}$ , namely  $V_{D+}$  and  $V_{D-}$ , for which the combination of gate/drain pulses results in almost negligible  $\Delta G$ , except for two corners, namely the combination of  $V_{G+}$  and  $V_{D-}$ , and the combination of  $V_{G-}$  and  $V_{D+}$ , for which the resulting  $\Delta G$  is a relatively large positive or negative value, respectively. This scheme allows to perform a parallel AND operation of binary input vectors applied at the WLs and BLs of an ECRAM array. To implement AND operation, a logic '1' is represented by  $V_{G+}$  and  $V_{D-}$  applied at the WL and BL, respectively, while a '0' is encoded by  $V_{G0}$  and  $V_{D+}$ , respectively (Fig. 6b). Similarly, a depression AND operation is achieved by encoding a logic '1' by  $V_{G-}$  and  $V_{D+}$  applied at the WL and BL, respectively, while a '0' is encoded by  $V_{G0}$  and  $V_{D-}$ , respectively (Fig. 6c). The resulting truth tables in Fig. 6bc show that the AND output is a relatively large positive/negative  $\Delta G$  to represent a logic '1', or a negligible  $\Delta G$  to represent a logic '0'.

The AND operation in Fig. 6 allows to implement analog, parallel TP in the ECRAM array. This is shown in Fig. 7a, illustrating the fully-parallel TP operation for a  $2 \times 2$  array, where simultaneous pulses are applied at the WL and BL terminals. Fig. 7b shows the measured  $\Delta G$  for all logic combinations of the TP scheme, with 4 trains of potentiation/depression pulses with  $t_p = 50$  ms. Fig. 7c shows  $\Delta G$  distribution for all operations. Results indicate that only the device performing a  $1 \times 1$  multiplication evidences a sig-

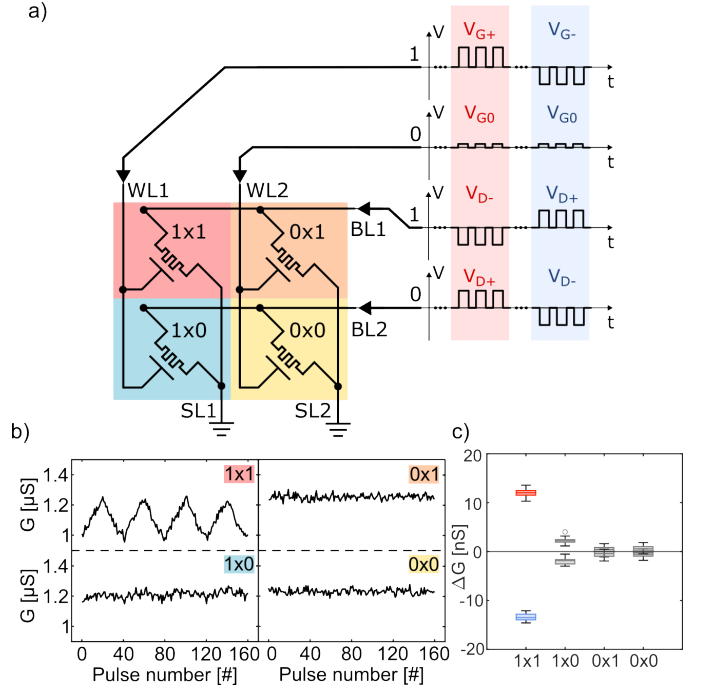


Fig. 7. Schematic of tensor product (TP) operation on a  $2 \times 2$  ECRAM array for all combinations of the applied voltages. (a) Selective update scheme, implementing a parallel logic AND for either potentiation or depression. (b) Device response for 4 sets of 20 up/down, 50 ms update pulses for all operations. (c) Measured  $\Delta G$  as a function of the logic input, showing good selectivity and symmetry under both polarities.

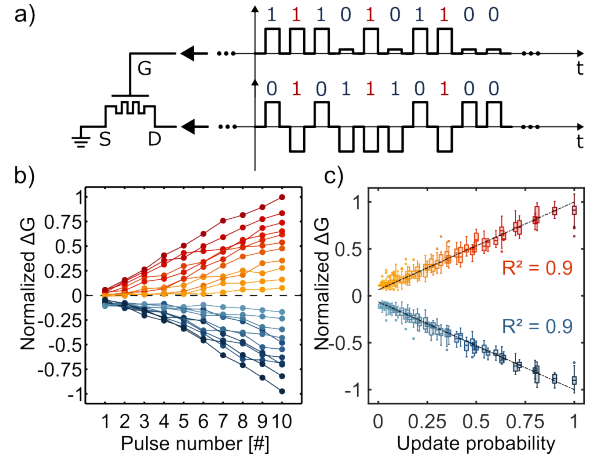


Fig. 8. 2-t stochastic product. (a) Random bit streams at drain and gate.  $\Delta G$  is proportional to the probability of 1x1 pulse coincidence. (b) Measured  $\Delta G$  for 10 streams of random bits, with probabilities ranging from 0 to 1 for both potentiation and depression. (c) Correlation between the final  $\Delta G$  and the ideal product value, evidencing the excellent accuracy of the stochastic TP.

nificant potentiation/depression, thus highlighting the correct execution of the parallel TP within the array.

AND operation also serves as the basis for stochastic update scheme [14], where input values are encoded as the probability of a bit being equal to '1' within a stream of random bits that are applied at the device terminals (see Fig. 8a). This operation can be extended at the array level by applying bit streams at

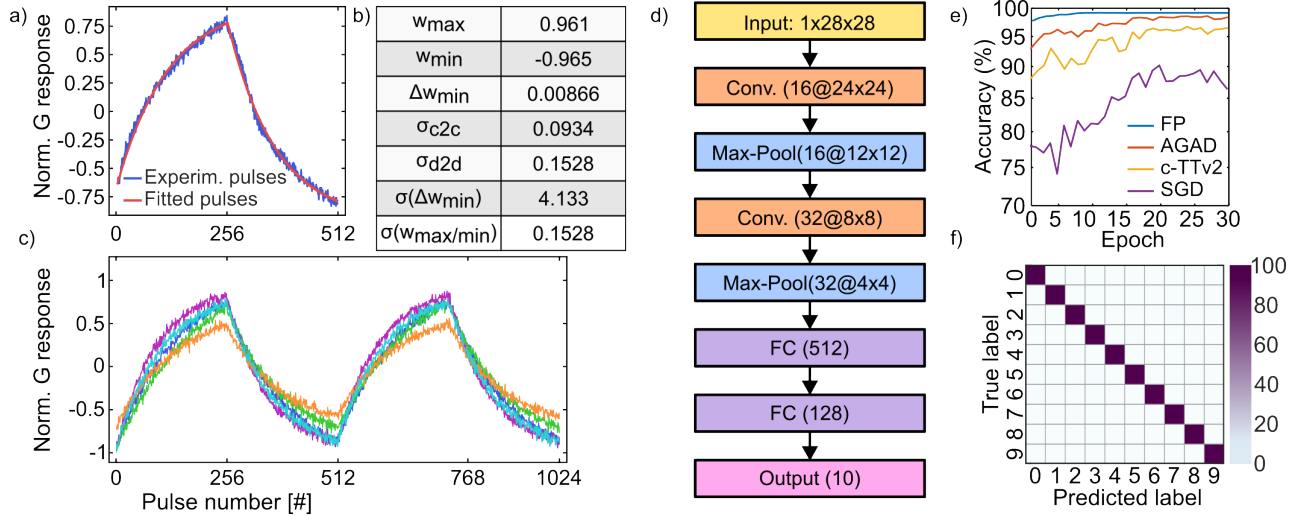


Fig. 9. Weight update modeling in AIHWKIT. (a) Normalized  $G$  response of ECRAM with 256 potentiation and depression pulses. (b) Experimental parameters for ECRAM model. (c) Simulated pulse responses including noise and device-to-device variation. (d) Structure of the LeNet5 CNN for image classification. (e) Accuracy during simulated training with ECRAM for various training algorithms, compared to floating point (FP) baseline. (f) Confusion matrix of trained CNN with AGAD algorithm for MNIST image recognition.

WLs and BLs, resulting in a  $\Delta G$  being proportional to the joint probabilities of WL and BL being equal to '1', thus performing parallel analog multiplication in each cell [4]. Fig. 8b shows the measured  $\Delta G$  for a pair of streams with 10 pulses, achieving both precise potentiation and depression. Fig. 8c shows the correlation plot of the measured  $\Delta G$  for several randomized bit streams applied in parallel on an ECRAM array, showing excellent update precision on all devices, with a high  $R^2 = 0.9$ .

#### IV. CNN TRAINING SIMULATION

Fully-parallel, stochastic TP allows for the acceleration of neural network training within ECRAM arrays [15]. Here, each ECRAM device operates as a synapse that can store analog weights [16] and execute (i) forward propagation by MVM and (ii) weight update by TP [17]. To support CNN training within ECRAM arrays, Fig. 9a shows the measured  $\Delta G$  for 256 potentiation and depression pulses obtained with the 2-t pulse-coincidence weight update scheme. A compact model for the measured  $\Delta G$  is obtained by using the 'SoftBounds' parametric device model in the IBM artificial intelligence hardware kit (AIHWKIT) [18] [19]. Device non-idealities such as noise, cycle-to-cycle variation and device-to-device variation were then included for a realistic device model. Fig. 9b shows the device parameters, while Fig. 9c shows the simulated ECRAM characteristics.

Based on the ECRAM model, the training of a CNN with LeNet5-like architecture (Fig. 9d) was simulated [20]. Fig. 9e shows the training accuracy after 30 epochs for various training algorithms, such as stochastic gradient descent (SGD) (accuracy 88.1%), and alternative training techniques, namely c-TTv2 (96.5%) and AGAD (98.7%), which were observed to be more resilient to non-ideal symmetry and linearity of the weight update [20]. Fig. 9f shows the confusion matrix

on the test data set for AGAD training. The results show that ECRAM device, thanks to the parallel TP algorithm for weight update enabled by the 2-t operation of ECRAM, is a promising device technology for accurate AI training within energy-efficient IMC circuits.

#### V. CONCLUSIONS

We presented a 2-t bias scheme supported by extensive characterization of the 2-t pulsed programming switching characteristics of ECRAM and physics-based compact modeling, enabling fully-parallel, precise and selective TP in 2x2 ECRAM arrays based on the stochastic pulsed technique. Based on these algorithms, CNN training with realistic ECRAM device characteristics is demonstrated with near-FP accuracy.

#### REFERENCES

- [1] Y. LeCun *et al.*, *Nature* 521, 436 (2015).
- [2] E. Strubell *et al.*, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, 3645 (2019).
- [3] P. Mannocci *et al.*, *APL Machine Learning* 1, 010902 (2023).
- [4] T. Gokmen *et al.*, *Front. Neurosci.* 10 (2016).
- [5] S. Agarwal *et al.*, *2016 International Joint Conference on Neural Networks (IJCNN)*. Vancouver, BC, Canada: IEEE, 929 (2016).
- [6] J. Cui *et al.*, *Current Opinion in Solid State and Materials Science* 32, 101187 (2024).
- [7] M. Porzani *et al.*, *IEEE Trans. Electron Devices* 71, 3240 (2024).
- [8] H. Kwak *et al.*, *Nano Convergence* 11, 9 (2024).
- [9] S. Kim *et al.*, *2019 IEEE International Electron Devices Meeting (IEDM)*. San Francisco, CA, USA: IEEE, 35.7.1 (2019).
- [10] S. Kim *et al.*, *Adv Elect Materials* 9, 2300476 (2023).
- [11] Y. Lu *et al.*, *Advanced Intelligent Systems*, 2401068 (2025).
- [12] M. Porzani *et al.*, *IEEE Trans. Electron Devices* 71, 3246 (2024).
- [13] K. Noh *et al.*, *Sci. Adv.* 10, ead13350 (2024).
- [14] A. Alaghi *et al.*, *ACM Trans. Embed. Comput. Syst.* 12, 1 (2013).
- [15] T. Gokmen *et al.*, *Front. Neurosci.* 11, 538 (2017).
- [16] E. R. W. Van Doremale *et al.*, *Sci. Adv.* 10, eado8999 (2024).
- [17] Y. Li *et al.*, *Front. Neurosci.* 15, 636127 (2021).
- [18] M. J. Rasch *et al.*, *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 1 (2021).
- [19] M. Le Gallo *et al.*, *APL Machine Learning* 1, 041102 (2023).
- [20] M. J. Rasch *et al.*, *Nat Commun* 15, 7133 (2024).