

UCLA

UCLA Electronic Theses and Dissertations

Title

Hafnium-Zirconium Oxide-Based Synaptic Resistor Circuit for Neuromorphic Computing

Permalink

<https://escholarship.org/uc/item/7cg111kc>

Author

Lee, Jungmin

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Hafnium-Zirconium Oxide-Based Synaptic Resistor Circuit for Neuromorphic Computing

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Mechanical Engineering

by

Jungmin Lee

2024

© Copyright by

Jungmin Lee

2024

ABSTRACT OF THE DISSERTATION

Hafnium-Zirconium Oxide-Based Synaptic Resistor Circuit for Neuromorphic Computing

by

Jungmin Lee

Doctor of Philosophy in Mechanical Engineering

University of California, Los Angeles, 2024

Professor Yong Chen, Chair

Within its tight power budget of 20 W, the human brain effortlessly processes a huge volume of information, including high-definition images with ~60 frames per second, an acoustic signal ranging from 20 Hz to 15 kHz, pressure sensing information over the skin, and chemical compositions through the receptor cells in the taste bud. The empirical observations on the brain revealed its massively parallel synaptic connection among neurons and their local interactions, which is believed to be one of the origins of its exceptional ability to infer and learn. Inspired by the brain's unparalleled efficiency and adaptability, neuromorphic computing has been introduced to emulate the neurobiological circuits of the brain to develop advanced artificial

intelligence systems. To achieve this, it is crucial to develop novel electronic devices, particularly those capable of implementing synaptic plasticity with minimal power consumption. This dissertation introduces a synaptic resistor based on HfZrO ferroelectric material, offering stable multi-level, non-volatile memory. Utilizing its inherent learning rules, a circuit of HfZrO-based synaptic resistors demonstrated real-time learning capabilities by navigating a drone through tree obstacles and erratic wind conditions in a simulated environment. Furthermore, an associative memory circuit based on synaptic resistors was proposed and employed to classify the MNIST handwritten dataset, illustrating its general applicability to AI and machine learning tasks. These advancements highlight the potential of HfZrO-based synaptic resistors in creating efficient and adaptive neuromorphic computing systems.

The dissertation of Jungmin Lee is approved.

Pei-Yu Chiou

Yongjie Hu

Lei He

Yong Chen, Committee Chair

University of California, Los Angeles

2024

For the love of God

TABLE OF CONTENTS

LIST OF FIGURES	viii
ACKNOWLEDGEMENTS	xix
VITA	xx
PUBLICATIONS	xxi
1. Introduction	1
1.1. The Brain, a Neurobiological Computer	5
1.2. Analog Circuits for Energy-Efficient Computing	8
1.3. Emerging Neuromorphic Devices	12
1.4. Research Goals	14
2. Synaptic Resistor based on Ferroelectric HfZrO	15
2.1. Device Fabrication	19
2.2. Material Characterization	23
2.2.1. Ferroelectric HfZrO	23
2.2.2. Electron Microscopy and Spectroscopy Analysis	24
2.3. Device Characterization	30
2.3.1. Nonlinear Analog Conductance Tuning	30
2.3.2. Energy-band Structure of Synaptic Resistor	35
3. Neuromorphic Computing using Synaptic Resistor Circuits	40
3.1. Autonomous Drone Navigation	42
3.1.1. Experimental Setup	46
3.1.2. Results	49
3.1.3. Discussion	56

3.2.	Associative Memory Circuit (AMC)	59
3.2.1.	Hopfield Network and Associative Memory	60
3.2.2.	Self-supervised Learning of Associative Memory Circuit	68
3.3.3.	MNIST Hand-written Digits Classification	71
4.	Conclusions	78
	REFERENCES	80

LIST OF FIGURES

Figure 1: The trend survey ³ on the energy demand in the computing field and a comparison to the global energy production limit. Extrapolated expectations on energy demand growth exceed the global energy production limit before 2040.....	1
Figure 2: Conventional computing architecture demands information transfer between the CPU and memory unit, which causes the “Von-Neumann Bottleneck,” slowing down the processing speed and increasing the power consumption with repeated charge-discharge of the bus wires with a large parasitic capacitance between the adjacent layers. In contrast, neuromorphic computing architecture employs the brain-like construction of circuit elements to enable parallel computing, emulating the brain’s way of computing.	3
Figure 3: A schematic diagram of a synaptic resistor based on a Silicon channel. A synaptic resistor device communicates with presynaptic and postsynaptic neuron circuits to concurrently inference and learn from their signals, based on its own learning rule which will be further discussed in the following chapter.	4
Figure 4: A schematic of neural communication via synapse.	7
Figure 5: A droplet falling on the still water surface using the finite-element-method (FEM) fluid dynamics solver (left). A photo of a droplet in the water (right).....	9
Figure 6: The disc-type integrator for analog calculation of integration.	10
Figure 7: Crossbar circuit based on the synaptic resistor devices. Input vector $\mathbf{X}$ is applied to the synaptic resistor circuit with its conductance matrix $\mathbf{W}$ and the result of the matrix multiplication $\mathbf{XW}$ is given as a current response $\mathbf{I}$	11

Figure 8: A synaptic resistor (synaptic resistor) circuit for intelligent systems. (A) A schematic diagram showing a crossbar synaptic resistor circuit composed of $M \times N$ synaptic resistors connected with M input electrodes and N output electrodes. $\mathbf{x}$ denotes a vector with element x_m as the voltage pulses applied to the m^{th} input electrode, $\mathbf{z}$ denotes a vector with element z_n as the voltage pulses applied to the n^{th} output electrode, and $\mathbf{I}$ denotes a vector with element I_n as the current induced at the n^{th} output electrode. (B) Sensors detect deviations of the states of a system $\mathbf{s}$ from targeted states $\hat{\mathbf{s}}$, denoted as $\mathbf{s} - \hat{\mathbf{s}}$, and convert them to input voltage pulses $\mathbf{x}$ using an input interface circuit. The output currents, $\mathbf{I} = \mathbf{W}\mathbf{x}$, triggered by $\mathbf{x}$ from the synaptic resistor circuit flow into an output interface circuit, generating output voltage pulses ($\mathbf{y}$) applied on actuators to modify $\mathbf{s}$, as well as voltage pulses ($\mathbf{z}$) applied on the output electrodes of the synaptic resistor circuit to modify the synaptic resistor conductance matrix ($\mathbf{W}$) according to the learning rule $\dot{\mathbf{W}} = \alpha \mathbf{z} \otimes \mathbf{x}$ 18

Figure 9: The structure of a synaptic resistor circuit. (A) An optical top-view image reveals an 8×8 synaptic resistor crossbar arrangement consisting of 8 rows of Ti input electrodes, 8 columns of Ti output electrodes, and 8 columns of $\text{WO}_{2.8}$ reference electrodes. (B) Another top-view optical image (upper part) and a corresponding 3D schematic illustration (lower part) illustrate a synaptic resistor composed of a silicon (Si) channel underneath the thermal silicon dioxide (SiO_2) layer. The $\text{WO}_{2.8}$ reference electrode is grown on an $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, which is deposited on the SiO_2 layer over the Si channel. This Si channel is connected to Ti input/output electrodes via a $\text{TiSi}_{0.9}$ layer. The $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer comprises multiple ferroelectric domains, denoted by red and blue arrows, as well as dielectric domains, which are depicted without arrows. (C) A cross-sectional scanning transmission electron microscopy (STEM) image illustrates the $\text{WO}_{2.8}/\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2/\text{SiO}_2/\text{Si}$ heterojunction, highlighting tetragonal (T) and orthorhombic (O)

ferroelectric phases within the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer. Arrows indicate the migration of oxygen vacancies, shown as open red circles, from the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer into the $\text{WO}_{2.8}$ layer. (D) A cross-sectional STEM image (on the right) is complemented by energy-dispersive X-ray (EDX) Spectroscopy chemical maps of Ti and Si elements (on the left), showing the distribution across the Ti/TiSi _{0.9} /Si heterojunction..	19
Figure 10: Schematic diagram of HfZrO-based synaptic resistor fabrication process..	22
Figure 11: Electron microscopy analysis on synaptic resistor. (A) Fast Fourier transform (FFT) image of an atomic resolution TEM, demonstrating the ferroelectric orthorhombic phase by its additional lattice points (010) originates from the asymmetry in HfZrO unit cell, due to the shifted oxygen ion of orthorhombic phase relative to the dielectric tetragonal phase. (B) EDX analysis on Ti/TiSi _{0.9} /Si stack. (C) EELS analysis on $\text{WO}_{2.8}/\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2/\text{SiO}_2/\text{Si}$ memory stack..	25
Figure 12: Energy dispersive x-ray (EDS) spectroscopy plot over (top) Ti/TiSi _{0.9} /Si (bottom) junction..	26
Figure 13: Simulated atomic unit cell of tetragonal vs. orthorhombic phase of HfZrO.....	27
Figure 14: Fast Fourier transform (FFT) image (top left), simulated diffraction pattern (bottom left), and simulated 2D atomic lattice fitting (inset) on tetragonal HfZrO TEM image.....	27
Figure 15: Fast Fourier transform image (top left), simulated diffraction pattern (bottom left), and simulated 2D atomic lattice fitting (inset) on orthorhombic HfZrO TEM image. The added asymmetry in the orthorhombic phase relative to the tetragonal phase is evidenced by the additional lattice points in the FFT image.	28
Figure 16: Annular dark-field images of (top) $\text{WO}_{2.8}/\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2/\text{SiO}_2/\text{Si}$ (bottom) stack.	29
Figure 17: Bright-field images. The numbering corresponds to the annular dark-field images...	29

Figure 18: The electric properties of synaptic resistors. (A) The current, I , on the output electrode of a synaptic resistor is displayed versus voltage, x , applied on the input electrode of the synaptic resistor with respect to its grounded output and reference electrodes. (B) The percentage changes in the synaptic resistor conductance, $(\Delta w/w_0) \times 100\%$, are plotted versus various x and z pulses concurrently applied on the input and output electrodes under the conditions of $-3 V \leq x = z < 0$ (red triangles), $0 < x = z \leq 3 V$ (blue triangles), $-3 V \leq x \leq 3 V$ and $z = 0$ (green squares), and $-3 V \leq z \leq 3 V$ and $x = 0$ (brown circles). (C) The percentage changes in the synaptic resistor conductance, $(\Delta w/w_0) \times 100\%$, induced by various x and z pulses with $x = z = -3 V$ (red triangles), $x = z = 3 V$ (blue triangles), $x = 0$ and $z = 3 V$ (brown circles), $x = 3 V$ and $z = 0$ (green squares), $x = -3 V$ and $z = 0$ (grey circles), and $x = 0$ and $z = -3 V$ (brown diamonds) are plotted against the numbers of applied pulses. (D) The percentage changes in conductance w , $(\Delta w/w_0) \times 100\%$, of a synaptic resistor are shown against the widths of the applied x and z pulses on its input and output electrodes, respectively, after the synaptic resistor experienced pulses with a total duration of 500 ms and amplitudes $x = z = -4 V$ (red triangles), $x = z = 4 V$ (blue triangles), $x = 0$ and $z = 4 V$ (brown circles), $x = 4 V$ and $z = 0$ (green squares), $x = -4 V$ and $z = 0$ (grey circles), and $x = 0$ and $z = -4 V$ (brown diamonds). (E) The average analog conductance values, $\bar{w}$, of 64 synaptic resistors are plotted against 1000 targeted analog conductance values, w_t , evenly separated by $\Delta w_t = 59 pS$. The inset in the top-left corner shows the distribution of the synaptic resistor conductance values when $w_t = 14.938 nS$, $14.997 nS$, and $15.056 nS$. The inset in the bottom-right corner shows $\bar{w}$ with $\pm 3\sigma_w$ standard deviations against w_t at $44.940 nS$, $44.999 nS$, and $45.058 nS$. (F) A synaptic resistor was iteratively modified to its high and low analog conductance values, w_H and w_L , which are displayed as blue and red squares..... 31

Figure 19: Learning accuracy, rate, endurance, and nonvolatile memory tests of synaptic resistors. Learning accuracy: (A) The absolute values of the differences between the average analog conductance values $\bar{w}$ and the targeted analog conductance values of synaptic resistors w_t , $|\bar{w} - w_t|$, and (B) the standard deviations of w , σ_w , are plotted against their corresponding targeted analog conductance values of synaptic resistors, w_t . (C) The probability densities of conductance values w of the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ -based synaptic resistors (blu), $\text{Al}_2\text{O}_{2.4}$ -based synaptic resistors¹⁰ (green), and CNT- based synaptic resistors⁸ (red) are plotted after the synaptic resistors were tuned toward their targeted conductance values. 34

Figure 20: Nonvolatile memory tests of synaptic resistors and device conductance comparison. After synaptic resistors were modified to different initial analog conductance values, the conductance value of each synaptic resistor, $w(t)$, measured at (A) 40 °C and (B) 85 °C is shown by dots in different colors as a function of time t over 10^6 s..... 35

Figure 21: The energy-band structures of a synaptic resistor. (A) A 3D schematic illustration of a synaptic resistor composed of a grounded $\text{WO}_{2.8}$ reference electrode, $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, SiO_2 dielectric layer, and Si channel along the H axis, and a $\text{TiSi}_{0.9}$ input electrode, Si channel, and $\text{TiSi}_{0.9}$ output electrode along the L axis. The $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer is composed of multiple ferroelectric domains, indicated by red or blue arrows, and dielectric domains, depicted without arrows. The simulated energy-band structures along the L axis (top) and along the H axis (bottom) of a synaptic resistor under the conditions of (B) the potentials at the input electrode $x = 3$ V and the output electrode $z = 3$ V, (C) $x = z = -3$ V, and (D) $x = 3$ V and $z = 0$ (green lines), $x = -3$ V and $z = 0$ (grey lines), and $x = z = 0$ (yellow lines). The top scale (L) represents the lateral distance from the input electrode to the output electrode, and the bottom scale (H) represents the vertical distance from the Si channel to the $\text{WO}_{2.8}$ reference electrode. 37

Figure 22: The energy-band structures of a synaptic resistor with a channel length of 100 nm.

The simulated energy-band structures of the $\text{TiSi}_{0.9}$ input electrode, Si channel, and $\text{TiSi}_{0.9}$ output electrode along the L axis in fig. S5A, and of the grounded $\text{WO}_{2.8}$ reference electrode, a 10 nm-thick $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, an 1 nm-thick SiO_2 dielectric layer, and Si channel along the H axis in fig. S5A of a synaptic resistor with a channel length of 100 nm under the conditions of (A) the potential at the input electrode $x = 1.8 \text{ V}$, the potential at the output electrode $z = 1.8 \text{ V}$, (B) $x = z = -1.0 \text{ V}$, and (C) $x = z = 0 \text{ V}$ (yellow lines), $x = -1.0 \text{ V}$ and $z = 0$ (deep blue lines), and $x = 1.8 \text{ V}$ and $z = 0$ (green lines). When the electric field, $\mathbf{E}$, in the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer satisfies $\mathbf{E} > \mathbf{E}_t^+ > 0$ in **a**, or $\mathbf{E} < \mathbf{E}_t^- < 0$ in **b**, with $\mathbf{E}_t^+$ and $\mathbf{E}_t^-$ as the positive and negative electric threshold fields to switch the dipole, $\mathbf{P}$, of the ferroelectric domains in the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, $\mathbf{P}$ is modified to align along $\mathbf{E}$. When $\mathbf{E}_t^- < \mathbf{E} < \mathbf{E}_t^+$ in (C), thus $\Delta\mathbf{P} \approx 0$ 39

Figure 23: The synaptic resistor chip is connected to neurons at the PCB through the adapter to constitute a neuromorphic computing circuit. Input voltage pulses from input presynaptic neurons drive the current from M synaptic resistors, which are connected to N output neurons. Depending on the logic of the integrate and fire circuit, the feedback pulse from the output neuron is applied to synaptic resistors to modify their conductance. 41

Figure 24: Schematics of experimental systems. Experimental systems for (A) a synstor circuit with real-time learning, (B) humans with real-time learning, and (C) a computer-based ANN with off-line learning to drive a drone toward a target position in a simulated environment. (D) Obstacles are detected by a simulated camera on the drone, a moving target position ($\hat{s}$) is identified to avoid the obstacles, and the deviation of the drone position (s) from its target position, $s - \hat{s}$, is input to the synstor circuit, humans, or ANN via interface circuits. 44

Figure 25: Experiments to navigate a drone by a synaptic resistor circuit, humans, and ANN. (A) a synaptic resistor circuit (top), humans (middle), and a computer-based ANN (bottom) (B) drive a drone toward a target position by avoiding obstacles in a simulated windy environment. The starting point of the flights is marked by a drone image, and the final target position is marked by a green target symbol. The flying traces are shown for the drone driven by the synaptic resistor circuit (blue line), the human (yellow line), and the ANN (red line). The inset shows the details in the area near the target with tree obstacles. The red crash symbols indicate that the drone crashed into the tree.. 45

Figure 26: Experimental results of drone navigations. (Top row) The flying traces are shown for a drone driven by (A) a synaptic resistor circuit, (B) human operators, and (C) an ANN toward a target position by avoiding obstacles in a simulated windy environment. The starting point of the flights is marked by drone images. (Middle row) The zoom-in images show the details of the flying traces in the area near the target with tree obstacles. (Bottom row) objective functions, $F = \frac{1}{2}(s - \hat{s})^2$, with $s - \hat{s}$ as the deviations of the drone position (s) from its target position ($\hat{s}$) versus accumulative flight and learning time t . The $F - t$ curves are best-fitted by $F(t) = (F(0) - F_e) e^{-t/T_L} + F_e$ (dashed lines). The red crash symbols indicate that the drone crashed into the tree, and the final target position is marked by green target symbols. Comparison of average (D) learning time (T_L), (E) equilibrium objective function (F_e), (F) adaptability to the changing environment, and (G) computing power consumption of the synaptic resistor circuit, human operators, and ANN in multiple trials. The adaptability to changing environment is evaluated by the successful rate of learning how to drive the drone to its final target without collisions in multiple trials. The power consumption of the synaptic resistor circuit is for

concurrent inference and learning, while the power consumption of the ANN is the aggregate of the power consumptions for sequential inference and learning. 51

Figure 27: Data from the experiments with a drone driven by a synstor circuit, a human operator, and an ANN running on a computer. The wind speed V_W , the wind direction measured in degrees clockwise from due north ϕ_W , the deviations of the drone position from the local target measured along the left-right $s_1 - \hat{s}_1$, up-down $s_2 - \hat{s}_2$, forward-backward $s_3 - \hat{s}_3$, clockwise-counterclockwise yaw $s_4 - \hat{s}_4$, directions, and the actuation pulses, $y_1, y_2, y_3, y_4, y_5, y_6, y_7$, and y_8 to move the drone along left, right, up, down, forward, backward, clockwise, and counterclockwise yaw directions are displayed against the time t in the experiments with a drone driven by (A) a synstor circuit, (B) a human operator, and (C) an ANN running on a computer. The input voltage pulses $x_1, x_2, x_3, x_4, x_5, x_6, x_7$, and x_8 converted from the $\mathbf{s} - \hat{\mathbf{s}}$ signals along left, right, up, down, forward, backward, clockwise, and counterclockwise yaw directions, respectively, and the back-propagating voltage pulses $z_1, z_2, z_3, z_4, z_5, z_6, z_7$, and z_8 on the output electrodes associated with the actuation pulses, $y_1, y_2, y_3, y_4, y_5, y_6, y_7$, and y_8 , respectively, are also displayed against the time t 53

Figure 28: Experiments to drive a drone by a synstor circuit without learning. During the flight processes of a drone driven by a synstor circuit with its conductance matrix set to random values and no learning, the objective functions, $F = \frac{1}{2}(\mathbf{s} - \hat{\mathbf{s}})^2$, with $\mathbf{s} - \hat{\mathbf{s}}$ as the deviations of the drone position ($\mathbf{s}$) from its target position ($\hat{\mathbf{s}}$) are displayed versus flight time t . The red crash symbols indicate that the drone crashed into the tree. 55

Figure 29: Learning experiments to drive a drone by ANN with different structures and learning parameters. (A) a schematic shows the structure of an ANN with 8 neurons at its input layer, no hidden layer, and 9 neurons at its output layer. The local objective functions, $F_L = \frac{1}{2}\mathbf{s}_L^2$, with $\mathbf{s}_L$

as the deviations of the drone position from its local target position, are displayed versus accumulative flight and learning time t during the flight processes of the drone driven by the ANN shown in (A) with a learning rate of r_L equal to (B) 10^{-8} , (C) 2×10^{-8} , (D) 5×10^{-8} , and (E) 10^{-7} . (F) the duration of learning episode to execute learning algorithm based on the data from a flight episode is shown against the duration of flight episode. (G) A schematic shows the structure of an ANN with 8 neurons at its input layer, 128 neurons at its hidden layer, and 9 neurons at its output layer. The objective functions, $F = \frac{1}{2}(s - \hat{s})^2$, are displayed versus accumulative flight and learning time t during the flight processes of the drone driven by the ANN shown in (G) with a learning rate of r_L equal to (H) 10^{-8} , (I) 2×10^{-8} , (J) 5×10^{-8} , and (K) 10^{-7} . The $F - t$ curves are best-fitted by $F(t) = (F(0) - F_e) e^{-t/T_L} + F_e$ (green lines) to extrapolate the average learning times T_L . (L) T_L is displayed versus the leaning rate r_L in the ANN shown in (A) (with no hidden layer, green circles) and (G) (with 1 hidden layer, red circles)..... 57

Figure 30: A schematic diagram of associative memory circuit based on the synaptic resistors.

(A) Associative memories (W) is modeled with a crossbar array of $M \times N$ devices to represent N memories with each memory have M components. A crossbar array is connected to M presynaptic neurons and N postsynaptic neurons, where the presynaptic neurons feed the input pulses and postsynaptic neurons integrate and fire the feedback signal for training..... 61

Figure 31: Characteristics of modern Hopfield network with different energy functions. (A) Four different memorized patterns (U, C, L, and A) are memorized to the networks. A total of 31 input patterns are selected to demonstrate the energy landscape for Hopfield networks. Spurious minima states were denoted as red dots, while blue dots represent the memorized pattern along the energy landscape. (B) A 2D energy landscape using multidimensional scaling (MDS)

technique. The blue/red dots represent the location of memorized patterns and spurious minima on an energy landscape for the corresponding energy functions..... 65

Figure 32: Binary operators based on AMC. (A) Input signals X_1 and X_2 are given to the AMC to select the output Y which fires first, to recall the result Z of a binary operation. (B) Binary operators implemented on AMC with their truth table memorized on the circuit..... 68

Figure 33: Hebbian learning, anti-Hebbian learning, and the global unlearning scheme of AMC.

(A) Input pulses for signed values with alternating voltage pulses. (B) Schematics for the input signals for concurrent inference/learning. For the columns where the post-synaptic neurons fire, (C) Hebbian learning is executed by applying feedback pulses for $W_{\mu_e}^+$, and $W_{\mu_e}^-$ respectively, and (E) anti-Hebbian learning could be implemented with the reversed order of feedback pulses with respect to Hebbian learning. (D) Global unlearning is to reduce the overall conductance of the memory when their current response (V) exceeds the threshold, with the feedback pulse width relative to the columnar response V_i 72

Figure 34: Evolution of associative memories assigned for digit ‘5’ after each training epoch. It begins with the network initialized to random noise and evolves under Hebbian learning, anti-Hebbian learning, and unlearning rules. After three epochs, the memories begin to form abstract representations of the digits, and they have evolved to create prototypes for each of the digits after the fifth ~ sixth epoch..... 74

Figure 35: Training AMC with different combinations of learning mechanisms. For the cases “Hebbian + anti-Hebbian” and “Hebbian + anti-Hebbian + unlearning”, only 10 memories are displayed. 76

Figure 36: Train and test accuracy for the network under different learning algorithms. Train/test accuracy at epoch 30, HL: 69%/69%, HL+UL: 76%/77%, HL+AHL: 95%/94%, HL+AHL+UL: 99%/96%.....	76
Figure 37: Total number of memories used for classification among 900 associative memories. HL: 30, HL+UL: 32, HL+AHL: 539, and HL+AHL+UL: 780. Using a combined learning algorithm, the AMC extracted 780 prototypical memories (images) for all the digitsTotal number of memories used for classification among 900 associative memories. HL: 30, HL+UL: 32, HL+AHL: 539, and HL+AHL+UL: 780. Using a combined learning algorithm, the AMC extracted 780 prototypical memories (images) for all the digits.	77
Figure 38: Extracted prototypical images of different digits using AMC network with 2700 memories (network size 784×2700). Out of 2700 memories, only the active memories that participate in classifying the digits are displayed..	79

ACKNOWLEDGEMENTS

I dedicate this thesis to my father, Kangwoo Lee, the most decent gentleman in the universe, and my mother, Sooyeon Kwan, who always made me realize what unconditional love means.

I want to express my deepest gratitude to Professor Yong Chen, my advisor, for his guidance throughout my doctoral studies. It would have been impossible for me to complete this journey without your generous support and kind yet life-steering advice.

To my brother, Shin Lee, my sister-in-law, Jieun Chung, and my beautiful nieces, Serin Lee, and Seyeon Lee, I cannot thank you all enough for your devotion to this family.

To my committee members, Professor Pei-Yu Chiou, Professor Yongjie Hu, and Professor Lei He, for their guidance and support.

To my past and present group members, Dr. Dhruva Nathan, Dr. Rahul Shenoy, Dawei Gao, Dr. Atharva Deo, Zixuan Rong, and Dr. Yong Hei, for all the stimulating discussions and laughter we had.

The research provided in this thesis is funded by the Air Force Office of Scientific Research (AFOSR) under the programs “Brain Inspired Networks for Multifunctional Intelligent Systems in Aerial Vehicles (FA9550-19-0213),” and “Multi-layer Self-Programming Neuromorphic Integrated Circuits for Deep Learning” (FA9550-20-1-0230).

The material in this thesis is either submitted or in preparation for publication.

VITA

2014	B.S. Mechanical Engineering Yonsei University
2016	M.S. Mechanical Engineering Yonsei University

PUBLICATIONS

Lee, J.*, Shenoy, R.* , Atharva, D., Yi, S., Gao, D., ..., and Chen, Y. HfZrO-based synaptic resistor circuit for super-Turing intelligent system. (*Equal contribution, Submitted)

Lee, J., Yi, S., Gao, D., and Chen, Y. Associative memory circuit based on synaptic resistors for concurrent inference and learning. (In preparation)

Deo, A., **Lee, J.**, ..., Chen, Y. An Intelligent self-driving car with energy-efficient synaptic resistor circuits for fast real-time learning based on HfZrO. (Submitted)

Deo, A., **Lee, J.**, ..., Chen, Y. An Intelligent non-Turing reusable spaceship system on HfZrO-based energy-efficient synaptic resistor circuits with fast real-time learning. (Submitted)

Deo, A., **Lee, J.**, ..., Chen, Y. Intelligent morphing wing based on synaptic resistor circuits with real-time learning. (Submitted)

Gao, D., Shenoy, R., Yi, S., **Lee, J.**, Xu, M., Rong, Z., Deo, A., Nathan, D., Zheng, J.G., Williams, R.S. and Chen, Y. Synaptic Resistor Circuits Based on Al Oxide and Ti Silicide for Concurrent Learning and Signal-Processing in Artificial Intelligence Systems. *Advanced Materials*, p.2210484 (2023)

Asapu, S., Pagaduan, J.N., Zhuo, Y., Moon, T., Midya, R., Gao, D., **Lee, J.**, ..., Yang, J.J. Large remnant polarization and great reliability characteristics in W/HFZRO/W ferroelectric capacitors. *Frontiers in Materials*, 9, p.969188 (2022)

Nathan, D., Deo, A., Haughn, K., Yi, S., **Lee, J.**, Gao, D., Shenoy, R., Xu, M., Tran, I.C., ..., Chen, Y. Si-based self-programming neuromorphic integrated circuits for intelligent morphing wings. *Journal of Composite Materials*, 56(30), pp.4561-4575 (2022)

1. Introduction

The artificial intelligence based on the large language model passed the Turing test^{1,2}. They have demonstrated remarkable capabilities, significantly enhancing productivity by seamlessly supporting and replacing human efforts. However, the widespread use of computers and machine learning accelerators is causing energy demand in computing to grow exponentially. If this trend continues, the energy consumption of computing could eventually reach the global energy production limit. (Figure 1)³. It is not far-fetched to state that energy consumption is one of the foremost tasks for artificial intelligence to continue its innovations and broaden its impact. The worst-case scenario suggests that if the energy demand in computing keeps increasing, the price of energy will rise in general, impacting people's daily lives, especially the populations in a lower financial bracket, completely nullifying the positive impact of artificial intelligence⁴⁻⁶.

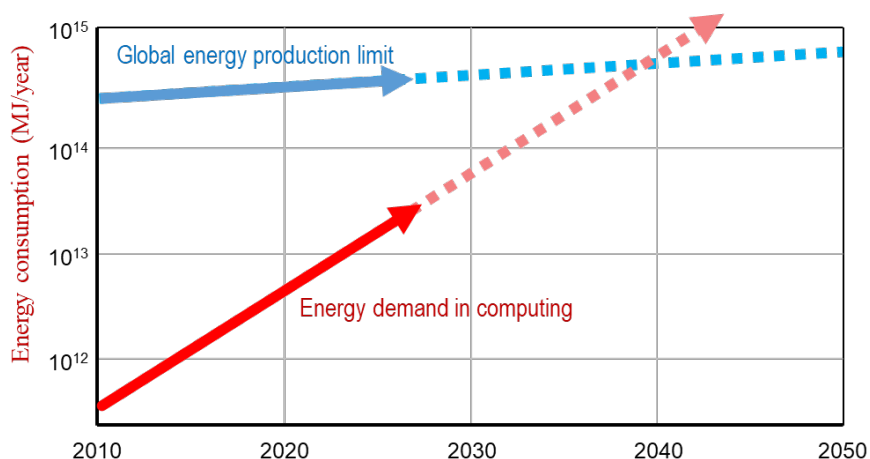


Figure 1. The trend survey³ on the energy demand in the computing field and a comparison to the global energy production limit. Extrapolated expectations on energy demand growth exceed the global energy production limit before 2040.

For the last decades, the advances in computing power and its energy efficiency have been closely related to the miniaturization of the transistors, with the trend following Moore's law, which states that the number of transistors on an integrated circuit is doubled every two years with a moderate amount of price adjustments over the unit chip area. However, this paradigm of the silicon industry has been forced to alter its direction of improvements since we have approached the fundamental limits on reducing the transistor feature sizes to the atomic level, e.g., about 25 Si atoms over a 5 nm trench. Other than mildly reducing the feature dimension, there are also various alternative solutions to continue Moore's law. A considerable amount of interest is drawn to the development of logic transistors with reduced power and additional scalability. For instance, complementary field-effect transistors (CFETs) are actively researched to overcome the aforementioned physical scaling issues by combining the PMOS and NMOS logic transistors in a vertical direction⁷. In addition, CFETs have a low supply voltage $V_{dd} \sim 0.6$ V, which can further lower the energy consumption. However, using low voltage amplitude can also slow down the performance of the logic circuitry because lower voltage levels typically result in reduced switching speeds and increased propagation delays within the circuit elements. As a result, while power consumption is minimized, the overall operational speed of the circuit is compromised, limiting further power reduction efforts without adversely affecting performance. Unarguably, for sustainable innovations in artificial intelligence, architecture-level innovations must be conducted in conjunction with device research to keep computing resources approachable. For these reasons, the concept of neuromorphic computing has particularly attracted more attention since it utilizes brain-inspired circuit elements to implement parallel computing architecture with a large potential to lower the energy consumption for artificial intelligent tasks. It averts the fundamental drawbacks of Von-Neumann architecture, where it has

a separate block for processors and memories with relentless information dispatch between the processors and a memory unit to perform sequential computing.

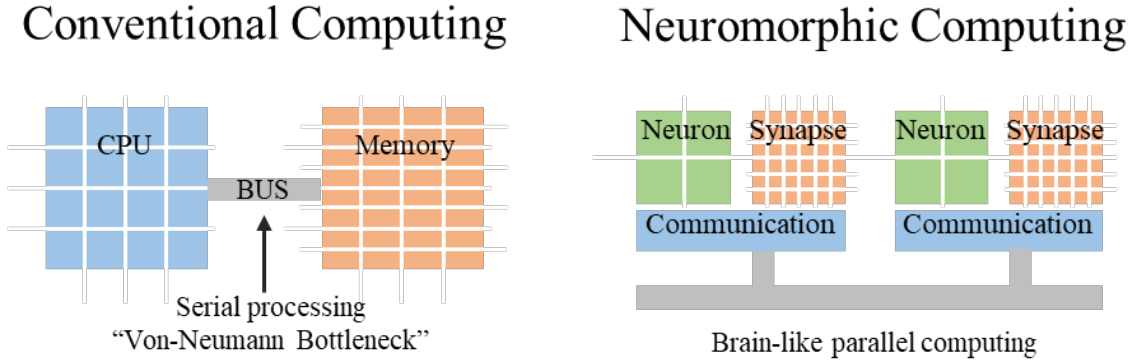


Figure 2. Conventional computing architecture demands information transfer between the CPU and memory unit, which causes the “Von-Neumann Bottleneck,” slowing down the processing speed and increasing the power consumption with repeated charge-discharge of the bus wires with a large parasitic capacitance between the adjacent layers. In contrast, neuromorphic computing architecture employs the brain-like construction of circuit elements to enable parallel computing, emulating the brain’s way of computing.

The synaptic resistor (synaptic resistor) is a new type of analog memory device with a large number of conductance states and stable endurance and retention characteristics that make it highly applicable in neuromorphic computing (Figure 3.). In addition, the synaptic resistor has its own learning rule that enables concurrent inference and learning, mimicking the synaptic activities in the human brain, which will be presented further in the following chapter. Prior works on synaptic resistors had successfully demonstrated analog memory actuation based on charge-trapping dynamics and its capability to perform real-time learning^{8–10}. However, these devices often suffer from non-ideal characteristics where the learning rule is not strictly followed. This issue partly arises from the distributed trap energy states between the electrode

and the memory oxide material where some trap states have a shallow energy gap between the Fermi level and the trapped charges, allowing the charges to escape easily. Consequently, this leads to partial shifts in the memory's conductance states with applied voltages where the conductance should not change.

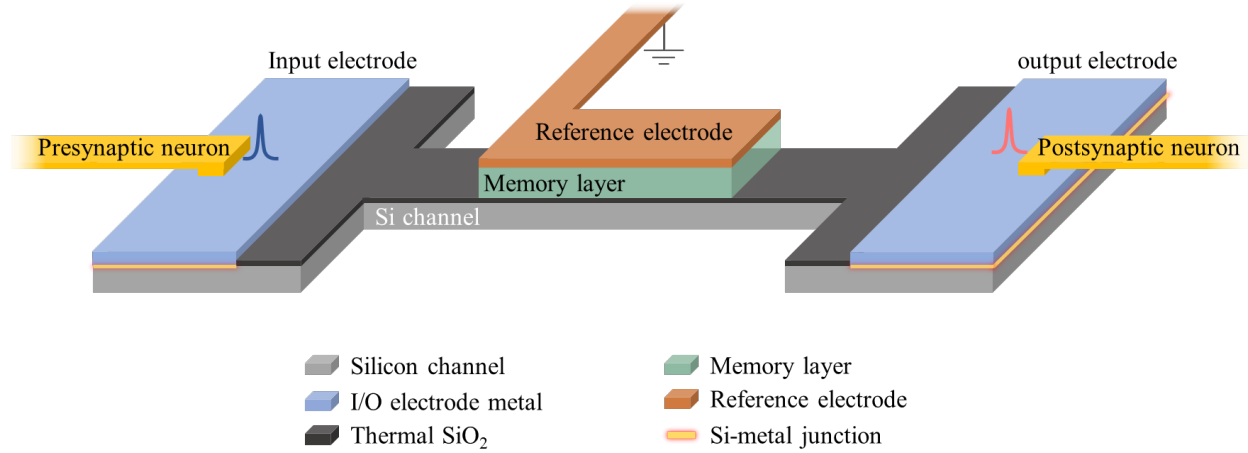


Figure 3. A schematic diagram of a synaptic resistor based on a Silicon channel. A synaptic resistor device communicates with presynaptic and postsynaptic neuron circuits to concurrently inference and learn from their signals, based on its own learning rule which will be further discussed in the following chapter.

This dissertation introduces the ferroelectric HfZrO-based synaptic resistors with their non-ideal learning characteristics that have been greatly improved relative to the charge-trap-based synaptic resistors. This new type of synaptic resistor utilizes the permanent polarizations of the ferroelectric unit cell, where its polarizations cannot be altered with the electric field strength below the coercive field strength, to demonstrate the superior non-linear characteristics, thus improving and expanding the synaptic resistor's implications, to artificial intelligence and neuromorphic computing. As a real-time learning application, autonomous driving of drones in a simulated environment is demonstrated, where the synaptic resistor circuit outperforms the

humans and artificial neural network. Lastly, the associative memory circuit based on synaptic resistors is proposed. Based on the model of the modern Hopfield network, the hardware-friendly update rule is derived and implemented on the circuit to classify the MNIST hand-written digit classification.

In the remaining portion of the introduction, a simple overview of the brain will be provided, with a focus on its implementations with hardware circuits. Firstly, empirical observations on the brain are presented to give the intuition of how the brain processes information and learns from it. Next, an introduction to emerging devices for neuromorphic computing and the efforts to emulate brain-like computing are addressed.

1.1. The Brain, a Neurobiological Computer

The human brain has 10^{11} neurons and 10^{14} synapses, with each neuron connected to ~ 1000 adjacent neurons via synapse, on average. With this biological computing unit, humans can process massive volumes of data in a fraction of a second¹¹, including high-resolution image processing through eyes with more than 60 frames per second, acoustic information ranging from 20 Hz to 12000 kHz with impeccable resolutions, pressure sensing information through skins, analyze the chemical compositions through the receptor cells in the taste buds, and the other less-common yet necessary information such as the sense related to balancing. Since the beginning of the century, brilliant researchers have devoted their time to deciphering the secret of the brain, but no one was able to fathom nor take one step closer to its design. However, there have been many empirical endeavors regarding their anatomy and signal processing in the basic building blocks of the brain. The findings have inspired many computer scientists and hinted them to develop brain-inspired computing algorithms to emulate how the brain works^{12–14}. One of the most important discoveries is about how neurons communicate. In Figure 4, a simplified neural

communication is described with a simple schematic diagram. When a membrane potential exceeds the threshold in its cell body, a presynaptic neuron sends an electric signal (action potential) toward another neuron. An action potential is a rapid electrical impulse signal from the neuron's cell body, propagating through the axon toward the synaptic junction between the presynaptic and postsynaptic neurons, triggering the release of neurotransmitters at the synapse, thereby communicating information between neurons. Synapse is the junction between the presynaptic and postsynaptic neurons where the transmission of signals occurs. On an axon terminal, neurotransmitters are released in synapses to activate the postsynaptic neurotransmitter receptors, inducing the electrical responses in a dendritic branch of a postsynaptic neuron. Dendrites are the branched projections of a neuron that receive signals from other neurons and convey this information to the cell body. This neural communication between adjacent neurons normally happens in a time window of hundreds of microseconds to a few milliseconds scale.

Based on this observation, it seems unlikely that the brain can process the aforementioned information with a huge volume in such a fraction of a second, because the single information propagation from a neuron to an adjacent neuron takes about a millisecond. An average human visual perception of 60 Hz implies that a brain is effortlessly processing the high-resolution image within a 17 ms scale, which is a narrow time window in which only a small number of information propagation through neurons is possible. Even more impressive is that the brain can easily process such a continuous flow of images at 60 Hz for more than 10 hours. As stated

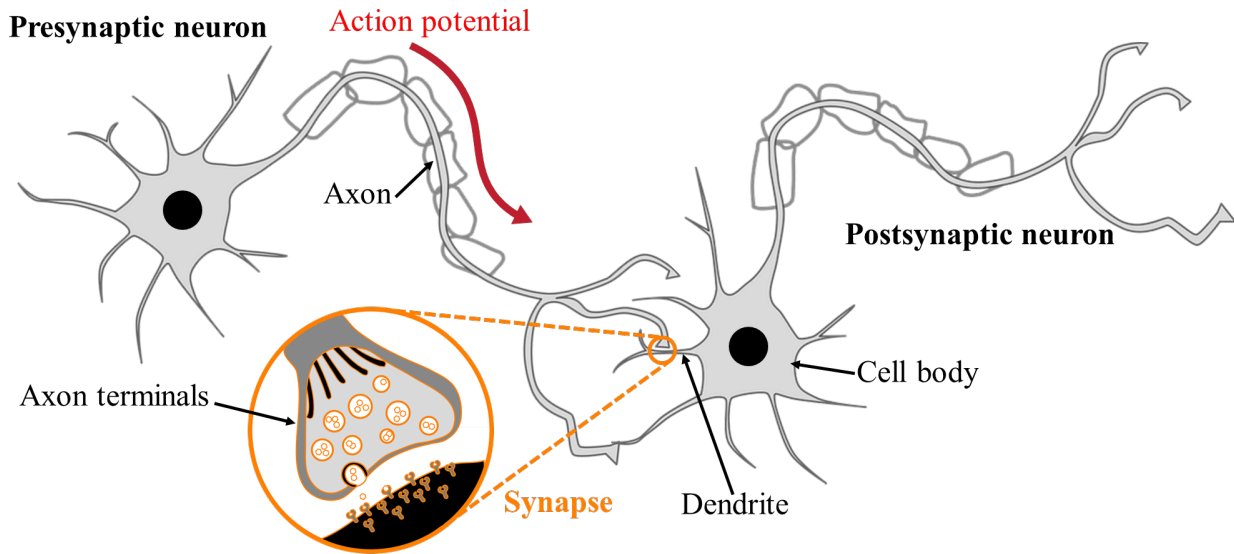


Figure 4. A schematic of neural communication via synapse.

above, no one has a complete understanding of the brain. However, despite the challenges, based on empirical observations, most contemporary researchers agree that the secret is mainly related to its massive parallelism and interconnected networks with immense redundancy^{15–19}, where the signal is mostly event-driven in a temporal space, with each neuron having a self-limiting mechanism with membrane potential repolarization after ‘firing.’ Furthermore, it has been observed that synaptic weights (conductance) can be modified during concurrent inference and learning processes by following a learning rule known as spike-time-dependent plasticity (STDP). The ability to simultaneously infer and learn sets the brain apart from computers, leading to significantly shorter learning latency and greater adaptability to dynamically changing environments compared to computers²⁰.

A simple comparison between the brain and cutting-edge artificial intelligence can provide a more vivid contrast between digital computing and neurobiological computing. It is not

an “apples-to-apples” comparison because it is impossible to find a brain with a computing performance like artificial intelligence models, but let’s assume the ability of the artificial intelligence to answer the question to be as good as one of the genius minds of all time, such as Sir. Isaac Newton. GPT-4 is a large language model developed by Open-AI, which has proven its general task-solving capabilities and is publicly used by many users²¹. It is said to have 1.76×10^{12} parameters, consumes ~ 100 MW of power during the training based on its GPUs. The training is mostly about carefully optimizing the vectorized tokens and their connections to come up with a natural language based on the context. The brain, in contrast, has $\sim 10^{14}$ synapses yet consumes 20 W, more than six orders less than GPT-4, for concurrent inference and training. This quick comparison suggests that the complexity of some contemporary neural networks has almost reached the level of biological computers, but the energy consumption efficiency of such networks is still far behind the superiority of their biological counterparts.

1.2. Analog Circuits for Energy-Efficient Computing

However, the question regarding whether computing using the biological brain is the most efficient method remains debatable. Further discussion should be made on the lower limit of power dissipation for different computing operations ranging from generalized to specific tasks. Intuitively, the trend is as follows; the more specific the task is, the less the power consumption is. On the contrary, the more general the task becomes, the more power consumption there will be. For instance, a finite-element-method fluid dynamics solver can simulate the behavior of a droplet on the still water surface (Figure 5. left), and it is a general computing platform with a large freedom on many parameters. For example, it can also simulate a liquid with a different viscosity and a different volume with geometrical constraints and boundary conditions. But it consumes a large amount of energy to calculate, even for a few

seconds scale. On the other hand, in real life, the same physical phenomena can be reproduced with minimal energy consumption by falling a droplet from the sky into the water's surface. Therefore, in case a higher-order equation (an equation analogous to a Navier-Stokes equation), which will consume a huge computation power to solve, could be modeled into real-world phenomena (like droplet behavior in water), an answer to the equation is obtained with minimal energy consumption, just by observing the real-time behavior of the system.

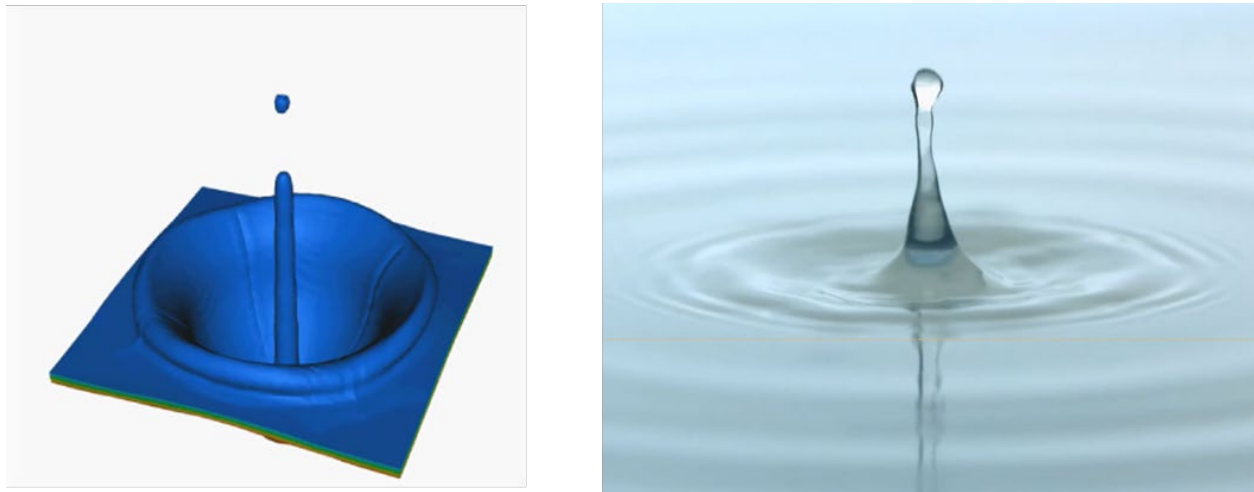


Figure 5. A droplet falling on the still water surface using the finite-element-method (FEM) fluid dynamics solver (left). A photo of a droplet in the water (right).

This method is not practical in many cases because of nature's fundamental noisiness due to thermal vibration and the difficulty of modeling the equations into the governing equations of natural phenomena, except for the case where the target task is directly related to the governing equation, e.g., a vortex experiment in a wind tunnel governed by the Navier-Stokes equation. Therefore, a practical approach to achieving generality in computing using physical systems is to simplify the complexity of the tasks being performed. For instance, the analog integrator can come up with a calculation result with a reasonable amount of precision, using the mechanical

moving parts modeled to be analogous to the integration (Figure 6). Unlike a digital computer, these mechanical parts can only perform a limited number of tasks with limited precision. However, their energy consumption can potentially be negligible compared to digital computers.

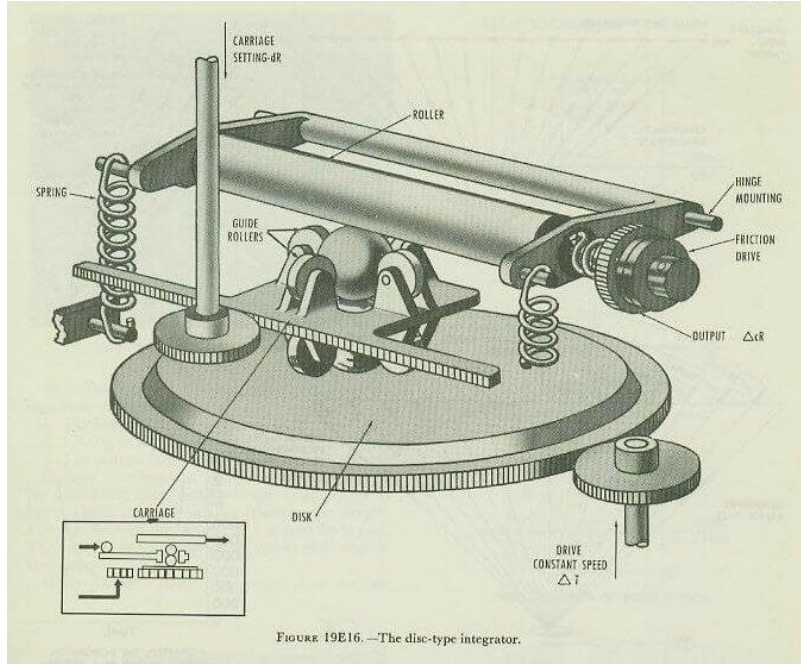


Figure 6. The disc-type integrator for analog calculation of integration.

The same principle applies to electronic circuits. Like mechanical mass-spring-damper systems, circuits consisting of resistors, capacitors, and inductors can easily model a wide range of second-order phenomena and be used for analog computing. The most widely used concept of such analog computing for machine learning is the crossbar array²². The crossbar circuit is composed of resistive circuit elements, which became the typical circuit layout for neuromorphic computing circuit chips with its ability to perform the multiply-and-accumulate (MAC) operations, a backbone of most machine learning algorithms. The circuit layout models the MAC operation based on Ohm's law and Kirchoff's law (Figure 7). First, the input voltage X_i is multiplied by the conductance W_{ij} of the resistive device based on Ohm's law $X_i \times W_{ij} = I_{ij}$,

and then the current response of the resistive devices in the same column in a crossbar is merged $\sum_i I_{ij} = I_j$ based on Kirchoff's law. The current response $\mathbf{I}$ for a whole column is equivalent to a result of a multiply-and-accumulate result $\mathbf{XW}$.

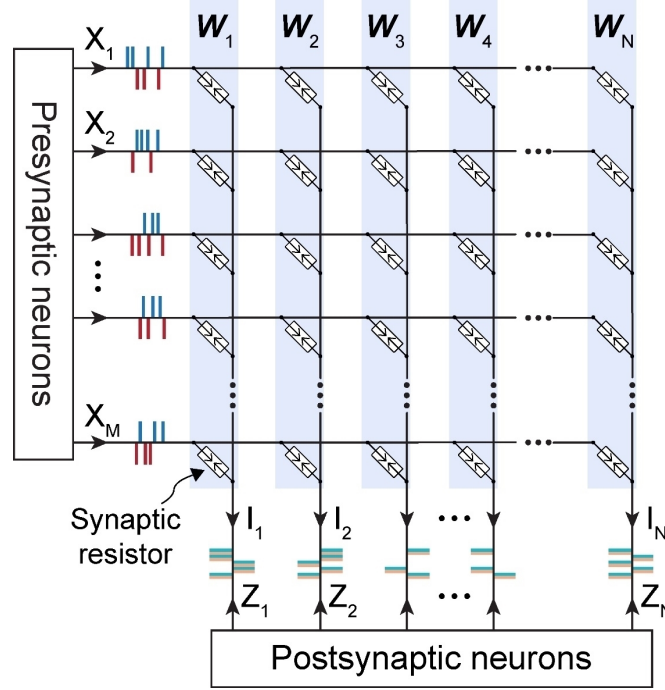


Figure 7. Crossbar circuit based on the synaptic resistor devices. Input vector $\mathbf{X}$ is applied to the synaptic resistor circuit with its conductance matrix $\mathbf{W}$ and the result of the matrix multiplication $\mathbf{XW}$ is given as a current response $\mathbf{I}$.

Other than the synaptic resistor in Figure 7, any non-volatile memory device can be used to configure a crossbar circuit to implement MAC operation, including two terminal devices, such as resistive random-access memory (RRAM)^{23,24}, conductive bridge random-access memory (CBRAM)²⁵, phase-change memory (PCM)²⁶, ferroelectric tunnel junctions (FTJs)^{27,28}, and magnetoresistive random-access memory (MRAM)²⁹, and 3-terminal transistors such as

floating gate field-effect transistors³⁰, ferroelectric field-effect transistors (FeFETs)³¹, organic electrochemical transistors (OECTs)³², and so on.

1.3. Emerging Neuromorphic Devices

In his seminal 1971 paper, Leon Chua introduced the memristor, a theoretical fourth fundamental passive circuit element, alongside the resistor, capacitor, and inductor³³. Unlike traditional elements, the resistance of a memristor changes based on the history of the current that has passed through it, providing non-volatile memory properties that are highly applicable to intelligent systems. It became a mainstream device for neuromorphic systems after its first physical realization by HP research team²³. Depending on the physical origin of memory, different types of memristors are proposed. Compared to traditional memory technologies such as floating-gate field-effect transistors and DRAM, memristors have proven to offer several advantages, including faster switching speeds, low energy consumption, and greater scalability.

Resistive random-access memory (RRAM)^{23,24,34–38} has emerged as a leading candidate for next-generation non-volatile memory technology. RRAM devices operate by modulating the resistance of a dielectric material, typically a transition metal oxide, through the application of an external voltage. With the external voltage applied, defect states and oxygen vacancies within the material form and drift, which results in its conductance change.

Conductive bridge random-access memory (CBRAM)²⁵ operates by creating and breaking conductive bridges within a solid electrolyte material, typically a silver or copper-based compound, through the application of an external voltage. This process of forming and dissolving metallic filaments allows CBRAM to switch between different resistance states in an analog manner.

Phase-change memory (PCM)^{26,39} operates by switching the phase of the chalcogenide glass material between its amorphous (high resistance) and crystalline (low resistance) states through the application of heat generated by a Joule-heating, induced by an electric current. However, such a working principle of PCM makes it typically more power-hungry than RRAM or MRAM.

Ferroelectric tunnel junctions (FTJs)^{27,28} consist of an ultrathin ferroelectric layer sandwiched between two metallic electrodes, forming a tunnel barrier. By applying an electric field, the polarization of the ferroelectric layer can be switched, which in turn modulates the tunneling electroresistance (TER) of the device.

Magnetic Random-Access Memory (MRAM)²⁹ operates by modulating the orientation of a magnetic polarization of a ferromagnetic layer relative to a permanent magnet reference layer, where the relative alignment determines the tunnel magnetoresistance. This magnetic orientation can be switched using spin-transfer torque (STT) or spin-orbit torque (SOT) mechanisms, enabling fast and efficient data writing.

Floating gate transistors^{40,41} are a well-established three-terminal non-volatile memory widely used in Flash memory devices. A floating gate is a conductive layer isolated by an insulating layer on top of the channel, with various material compositions ranging from polysilicon and silicon nitride to hafnium oxide with a huge density of trap states for charge carrier injection. By trapping hot electrons or removing trapped charges in the floating gate, the threshold voltage of the transistor can be modulated by their electric field effect. It is a mature technology in the semiconductor industry, which makes it an attractive non-volatile memory device for neuromorphic applications³⁰.

Ferroelectric field-effect transistors (FeFETs)³¹ are an emerging non-volatile memory technology that integrates ferroelectric materials into the gate stack of traditional transistor structures. FeFETs utilize a ferroelectric layer as their memory, allowing an external electric field to switch the polarization state of the ferroelectric material. This polarization switching modulates the threshold voltage of the transistor, thereby changing the conductance of the channel.

Organic electrochemical transistors (OECTs)³² Technology modulates a channel's conductivity through ionic movements within an electrolyte, which are controlled by an applied gate voltage. This ion migration alters the channel's conductivity, allowing the device to switch between different resistance states depending on the density of ions in the channel. It offers great potential for bioelectronics, where it can bridge electronics and biological matters as a biocompatible device.

1.4. Research Goals

To emulate the brain, a neuromorphic computing platform could be developed based on the following intuitions derived from the experiments and observations on the brain. First, the neurobiological circuit learns from the local information. Synaptic connection is altered based on the signals from its presynaptic and postsynaptic neurons, without global information. In other words, there is no central processor in the brain. Second, each neuron has synaptic connections to more than a thousand adjacent neurons, which enables the massively parallel processing of information. Third, a neurobiological circuit learns in real time, which means the signals for inference and learning are not distinct from each other. Fourth, the brain exhibits a high degree of fault tolerance and robustness, capable of functioning despite the loss of individual neurons or synaptic connections. Lastly, the brain is an event-driven circuit, meaning that neurons fire and

communicate based on discrete events, such as spikes in neuronal activity, rather than continuous signals. This event-driven nature allows the brain to process information efficiently, responding dynamically to changes in the environment and conserving energy by only using resources when necessary.

In this work, a novel synaptic device with a local learning rule and concurrent inference and learning capability is developed, and its application to neuromorphic tasks is demonstrated. Among the above features, massive parallelism, fault tolerance, and event-driven nature are mostly related to the circuit layout and peripheral circuit design, and they are outside the scope of this dissertation and left for future endeavors.

2. Synaptic Resistor based on Ferroelectric HfZrO

Analogous to a neurobiological network interconnected by synapses, a circuit consisting of M input electrodes and N output electrodes, interconnected by $M \times N$ synaptic resistors, is illustrated in Figure 8A. Voltage pulses $\mathbf{x}$ applied at the input electrodes generate currents $\mathbf{I}$ through the synaptic resistors at the output electrodes, thereby executing an inference algorithm expressed as:

$$\mathbf{I} = \mathbf{W}(t)\mathbf{x} \quad (1)$$

In this equation, $\mathbf{W}(t)$ represents the conductance matrix of the synaptic resistors. The excitatory (or inhibitory) currents $\mathbf{I}$ subsequently trigger (or inhibit) the voltage pulses $\mathbf{y}$ applied on actuators via neuron and interface circuits, which modify the states of a system $\mathbf{s}$. The deviations of $\mathbf{s}$ from the targeted states $\hat{\mathbf{s}}$ represented as $\mathbf{s} - \hat{\mathbf{s}}$, are detected by sensors and converted to $\mathbf{x}$

by interface circuits (Figure 8B). The goal is to minimize the objective function $F =$

$\frac{1}{2}(\mathbf{s} - \hat{\mathbf{s}})^2$, thus tuning $\mathbf{s}$ toward $\hat{\mathbf{s}}$.

Unlike computer-based artificial neural networks (ANNs), where $\mathbf{W}$ remains constant during the execution of their inference algorithms, the $\mathbf{W}(t)$ in the synaptic resistor circuit, similar to a neurobiological network, may not remain static throughout the inference process. Instead, it can be modified during a concurrent learning process by following a correlative learning rule⁴², described as:

$$\frac{d\mathbf{W}}{dt} = \alpha \mathbf{z}(t) \otimes \mathbf{x}(t) \quad (2)$$

Here, α denotes a learning coefficient, $\mathbf{z}$ represents voltage pulses triggered at the output electrodes of the circuit, and $\mathbf{z}(t) \otimes \mathbf{x}(t)$ represents the outer product between $\mathbf{z}(t)$ and $\mathbf{x}(t)$.

When $\overline{\left(\frac{dF}{dt}\right)} < 0$, $\mathbf{W}$ deviates from $\widehat{\mathbf{W}}$ due to environmental, system, or circuit changes, and $\mathbf{W}$ is adjusted toward $\widehat{\mathbf{W}}$ to decrease $\bar{F}$ during concurrent learning and inference processes.

To execute the correlative learning rule $\dot{\mathbf{W}} = \alpha \mathbf{z} \otimes \mathbf{x}$ by applying various $\mathbf{z}$ pulses, different learning algorithms in all three major machine learning paradigms (unsupervised, supervised, and reinforcement learning) can be implemented in multi-layer synaptic resistor circuits.

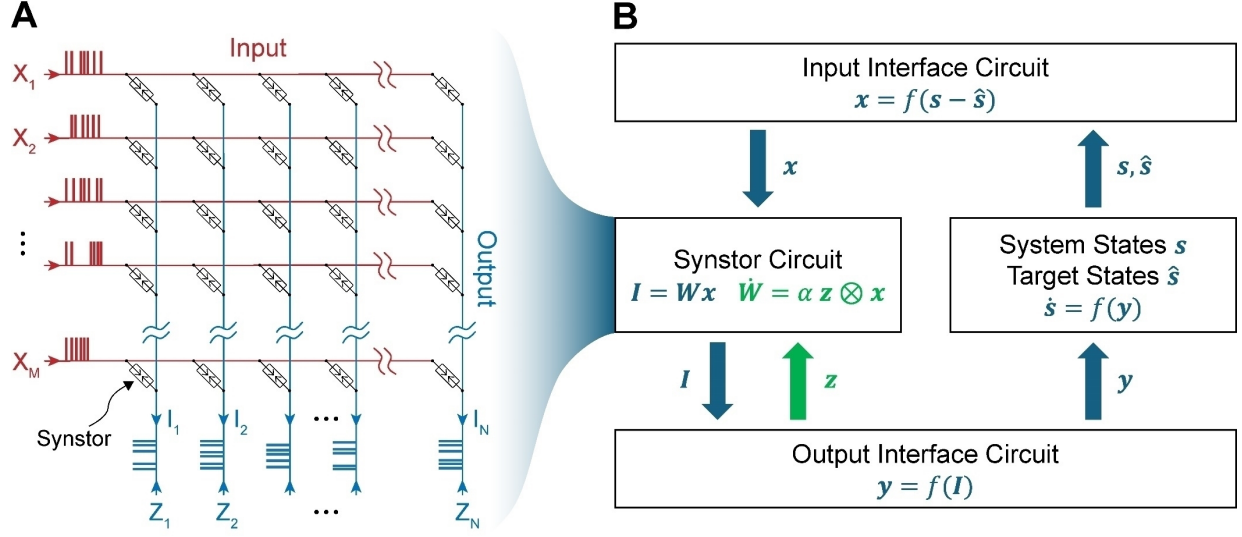


Figure 8. A synaptic resistor (synaptic resistor) circuit for intelligent systems. (A) A schematic diagram showing a crossbar synaptic resistor circuit composed of $M \times N$ synaptic resistors connected with M input electrodes and N output electrodes. $\mathbf{x}$ denotes a vector with element x_m as the voltage pulses applied to the m^{th} input electrode, $\mathbf{z}$ denotes a vector with element z_n as the voltage pulses applied to the n^{th} output electrode, and $\mathbf{I}$ denotes a vector with element I_n as the current induced at the n^{th} output electrode. (B) Sensors detect deviations of the states of a system $\mathbf{s}$ from targeted states $\hat{\mathbf{s}}$, denoted as $\mathbf{s} - \hat{\mathbf{s}}$, and convert them to input voltage pulses $\mathbf{x}$ using an input interface circuit. The output currents, $\mathbf{I} = \mathbf{W}\mathbf{x}$, triggered by $\mathbf{x}$ from the synaptic resistor circuit flow into an output interface circuit, generating output voltage pulses ($\mathbf{y}$) applied on actuators to modify $\mathbf{s}$, as well as voltage pulses ($\mathbf{z}$) applied on the output electrodes of the synaptic resistor circuit to modify the synaptic resistor conductance matrix ($\mathbf{W}$) according to the learning rule $\dot{\mathbf{W}} = \alpha \mathbf{z} \otimes \mathbf{x}$.

The fabrication of the synaptic resistor circuit presents significant challenges, necessitating the development of materials and devices capable of executing both inference and

learning simultaneously in an analog mode with high precision, reliability, speed, and repeatability. Synaptic resistors are fabricated by integrating a $\text{WO}_{2.8}/\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ heterojunction with a silicon (Si) channel within each device. As depicted in Figure 9A, the crossbar synaptic resistor circuit was fabricated following the detailed process described in the next section and illustrated in Figure 10. The structural and material properties of the synaptic resistors were thoroughly characterized using scanning transmission electron microscopy (STEM), energy-dispersive X-ray (EDX) spectroscopy, and electron energy loss spectroscopy (EELS), as shown in Figures 9B, 9C, 9D, Figure 11.

The synaptic resistor device has a 40 μm -long, 220 nm-thick silicon channels with a boron doping concentration of $10^{15}/\text{cm}^3$, derived from a single-crystal silicon layer on a 3 μm -thick silicon dioxide (SiO_2) box layer. A 14 μm -long, 80 nm-thick $\text{WO}_{2.8}$ reference electrode was fabricated on top of a 12.6 nm-thick $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, which was itself deposited on a 3.5 nm-thick SiO_2 thermal oxide layer over the Si channel. The $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer was composed of different phases of ferroelectric orthorhombic, dielectric tetragonal, and monoclinic phases. Additionally, a 12 nm-thick metallic $\text{TiSi}_{0.9}$ layer was interposed between the titanium (Ti) input/output electrodes and the Si channel, establishing a Schottky contact with the Si channel.

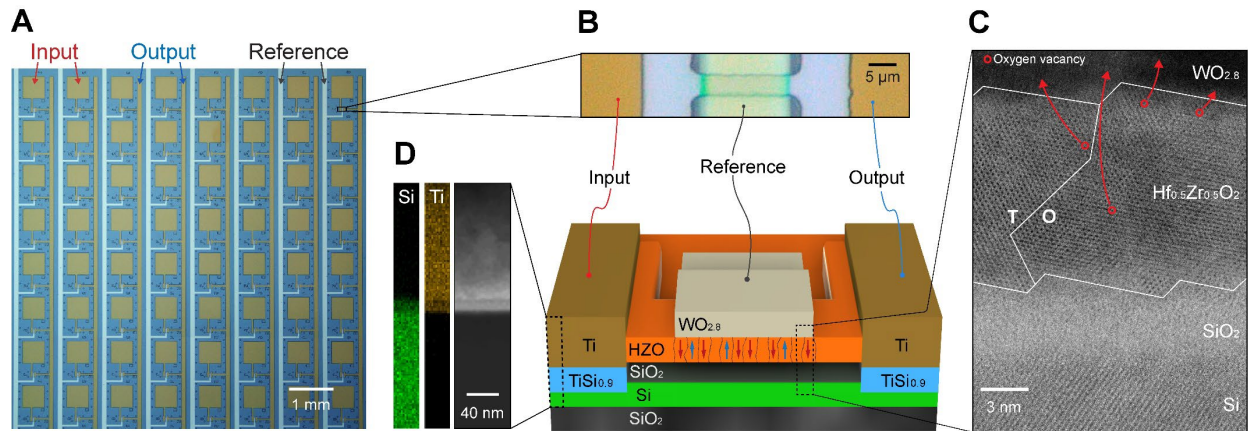


Figure 9. The structure of a synaptic resistor circuit. (A) An optical top-view image reveals an 8×8 synaptic resistor crossbar arrangement consisting of 8 rows of Ti input electrodes, 8 columns of Ti output electrodes, and 8 columns of $\text{WO}_{2.8}$ reference electrodes. (B) Another top-view optical image (upper part) and a corresponding 3D schematic illustration (lower part) illustrate a synaptic resistor composed of a silicon (Si) channel underneath the thermal silicon dioxide (SiO_2) layer. The $\text{WO}_{2.8}$ reference electrode is grown on an $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, which is deposited on the SiO_2 layer over the Si channel. This Si channel is connected to Ti input/output electrodes via a $\text{TiSi}_{0.9}$ layer. The $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer comprises multiple ferroelectric domains, denoted by red and blue arrows, as well as dielectric domains, which are depicted without arrows. (C) A cross-sectional scanning transmission electron microscopy (STEM) image illustrates the $\text{WO}_{2.8}/\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2/\text{SiO}_2/\text{Si}$ heterojunction, highlighting tetragonal (T) and orthorhombic (O) ferroelectric phases within the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer. Arrows indicate the migration of oxygen vacancies, shown as open red circles, from the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer into the $\text{WO}_{2.8}$ layer. (D) A cross-sectional STEM image (on the right) is complemented by energy-dispersive X-ray (EDX) Spectroscopy chemical maps of Ti and Si elements (on the left), showing the distribution across the Ti/ $\text{TiSi}_{0.9}$ /Si heterojunction.

2.1. Device Fabrication

The fabrication process of a synaptic transistor (synaptic resistor) circuit is detailed in Figure 10. This synaptic resistor circuit was developed on a silicon-on-insulator (SOI) chip. The chip comprised a 220 nm-thick p-type silicon (Si) layer with a boron doping concentration of $10^{15} / \text{cm}^3$, set on a 3 μm -thick buried silicon oxide (SiO_2) layer on a 725 μm -thick silicon wafer. The process began with patterning a photoresist layer on the silicon surface, which served as an

etching mask created through ultraviolet photolithography. As illustrated in Figure 10A, a 5 μm -wide silicon channel was formed using reactive ion etching (RIE) with an Oxford Plasmalab 80 Plus system. Subsequently, the photoresist was removed using a combination of acetone, isopropanol, and deionized water, followed by a thorough cleaning of the wafer using the standard Radio Corporation of America (RCA) cleaning procedure.

Next, as depicted in Figure 10B, the surface of the silicon channel underwent oxidation in a thermal oxidation furnace at 900 $^{\circ}\text{C}$ in an oxygen atmosphere for 10 seconds, resulting in the growth of a 3.5 nm-thick SiO_2 layer. Another layer of photoresist was patterned by UV photolithography on top of this SiO_2 layer. These photoresist patterns acted as an etching mask for RIE, which was used to etch the SiO_2 layer in the contact areas designated for the input/output electrodes, as shown in Figure 10C.

A 300 nm-thick titanium (Ti) layer was then deposited onto the chip using electron beam evaporation (CHA Industries Mark 40). The titanium input/output electrodes were created in the contact areas by lifting off a pre-patterned photoresist layer, as depicted in Figure 10D. Following this, a 40 nm-thick aluminum oxide (Al_2O_3) sacrificial layer was deposited onto the chip via atomic layer deposition (ALD) using a Fiji Ultratech ALD system. The chip was subsequently annealed in forming gas (a mixture of 5% hydrogen in nitrogen) at 460 $^{\circ}\text{C}$ for 30 minutes, which resulted in the formation of a titanium silicide layer sandwiched between the silicon channel and the titanium input/output electrodes (Figure 10E). During this annealing process, the Al_2O_3 sacrificial layer served to protect the materials from unwanted oxidation.

After annealing, the Al_2O_3 sacrificial layer was selectively etched away using a 3% tetramethylammonium hydroxide aqueous solution. Subsequently, a 12.6 nm-thick layer of hafnium zirconium oxide ($\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$) was deposited onto the chip using ALD at 200 $^{\circ}\text{C}$,

employing tetrakis(dimethylamino)hafnium(IV) and tetrakis(dimethylamino)zirconium(IV) as precursors (Figure 10F). Following this, an 80 nm-thick tungsten oxide (W oxide) layer was deposited onto the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer through magnetron sputtering (Denton Discovery). The tungsten oxide reference electrodes were then patterned by lifting off a photoresist layer (Figure 10G).

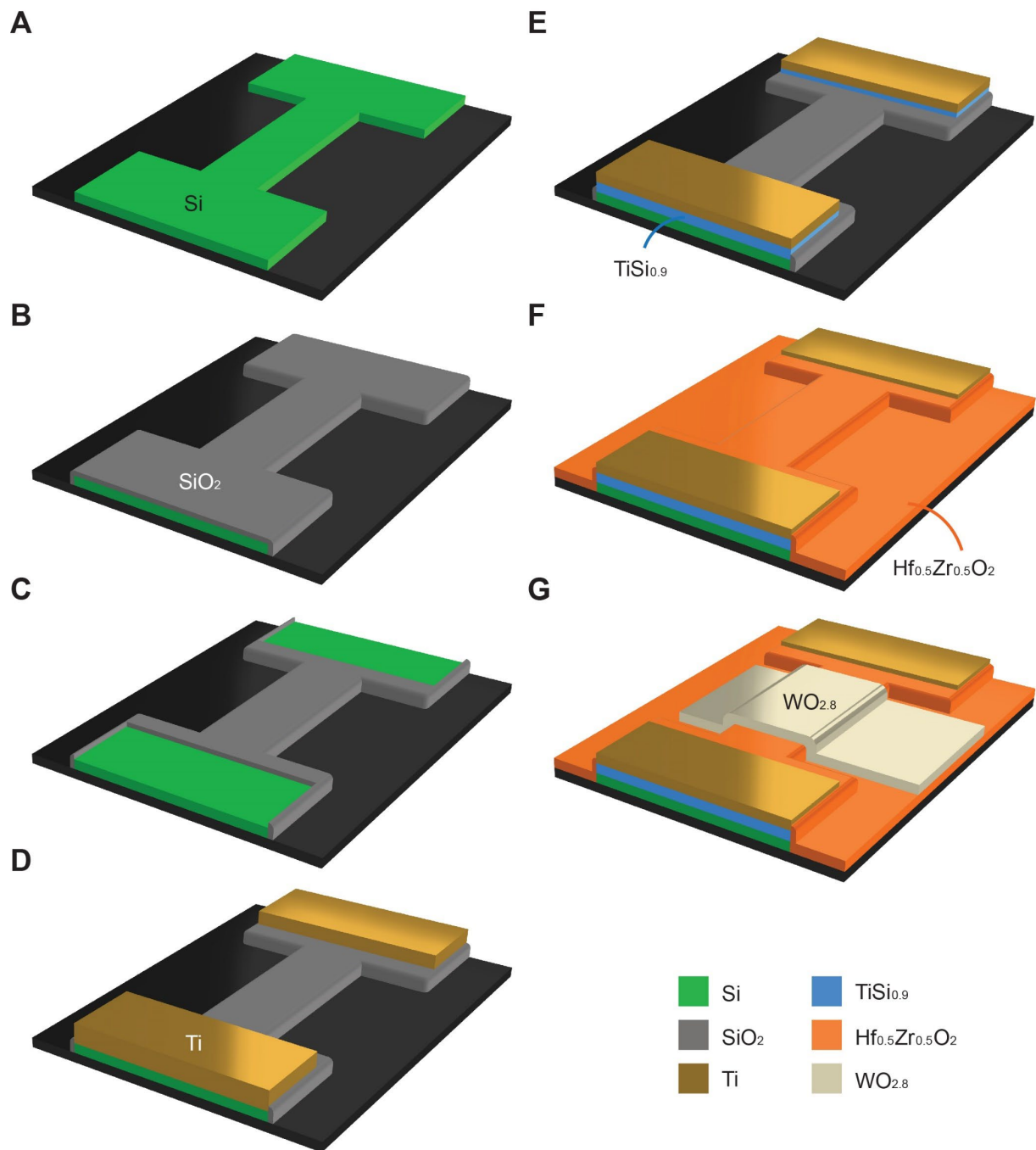


Figure 10. Schematic diagram of HfZrO-based synaptic resistor fabrication process.

The chip then underwent rapid thermal annealing in a Modular Process Technology RTP-600xp system at 500 °C for 1 minute to crystallize the materials and induce a heterojunction between the ferroelectric $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer and the $\text{WO}_{2.8}$ reference electrode. The final step involved etching away the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer on the contact pads of the input/output electrodes using photolithography and RIE (Oxford Plasmalab 80 Plus RIE). This thorough and intricate fabrication process ensured the creation of a functional and efficient synaptic resistor circuit, as illustrated in Figure 10.

2.2. Material Characterization

2.2.1. Ferroelectric HfZrO

Ferroelectric hafnium zirconium oxide ($\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$, or HfZrO) has emerged as a highly promising material in the field of advanced electronics, particularly in the development of non-volatile memory devices, neuromorphic computing, and other applications requiring stable, energy-efficient ferroelectric properties^{31,43–45}. HfZrO is a solid solution of hafnium dioxide (HfO_2) and zirconium dioxide (ZrO_2), both of which are renowned for their high dielectric constants, thermal stability, and compatibility with silicon technology. What sets HfZrO apart is its ability to exhibit robust ferroelectricity at nanoscale dimensions, a property that traditional ferroelectrics like lead zirconate titanate (PZT) struggle to maintain. This makes HfZrO particularly attractive for integration into modern semiconductor processes, where miniaturization is a critical concern. The ferroelectric phase of HfZrO, typically stabilized by specific processing conditions such as doping concentrations, annealing temperatures, and deposition methods, allows for the reversible polarization switching essential for memory and logic devices. Additionally, HfZrO's compatibility with standard complementary metal-oxide-semiconductor (CMOS) fabrication processes reduces the complexity and cost of device

manufacturing. This material's unique combination of high remnant polarization, low coercive field, and endurance against fatigue and retention loss makes it an ideal candidate for applications in ferroelectric random-access memory (FeRAM), ferroelectric field-effect transistors (FeFETs), and other emerging technologies^{28,46}. Moreover, the environmental and health concerns associated with lead-based ferroelectrics are mitigated with HfZrO, as it is a lead-free alternative, further enhancing its appeal in the push toward greener electronic solutions. Researchers continue to explore the fundamental properties of HfZrO, optimizing its ferroelectric performance through various doping strategies and deposition techniques, ensuring it remains at the forefront of next-generation electronic materials.

2.2.2. Electron Microscopy and Spectroscopy Analysis

The ferroelectric phase of HfZrO is a metastable state, achieved by inducing in-plane tensile stress with post-deposition annealing. The dielectric tetragonal phase is the most stable state of HfZrO at room temperature, but the mismatch of thermal expansion coefficient between the metal capping layer and HfZrO imposes a strain inside HfZrO, modulating the energy landscape, in turn slightly off-centers the central oxygen anion to reduce the system's energy, which is an origin of the ferroelectricity in HfZrO (Figure 13). The ferroelectric phase of the HfZrO material has been analyzed using a fast Fourier transform (FFT) technique on an atomic-resolution TEM image (Figure 14, 15), which clearly reveals the added asymmetry in the orthorhombic phase relative to the tetragonal phase, evidenced by the additional lattice points in Figure 11A.

EDX analysis was performed using JEOL JEM-2800 TEM operated at 200 kV (Figure 12). Atomic resolution STEM and EELS analysis was performed with a JEOL Grand ARM TEM operated at 300 kV with a spherical aberration corrector (Figure 16, 17).

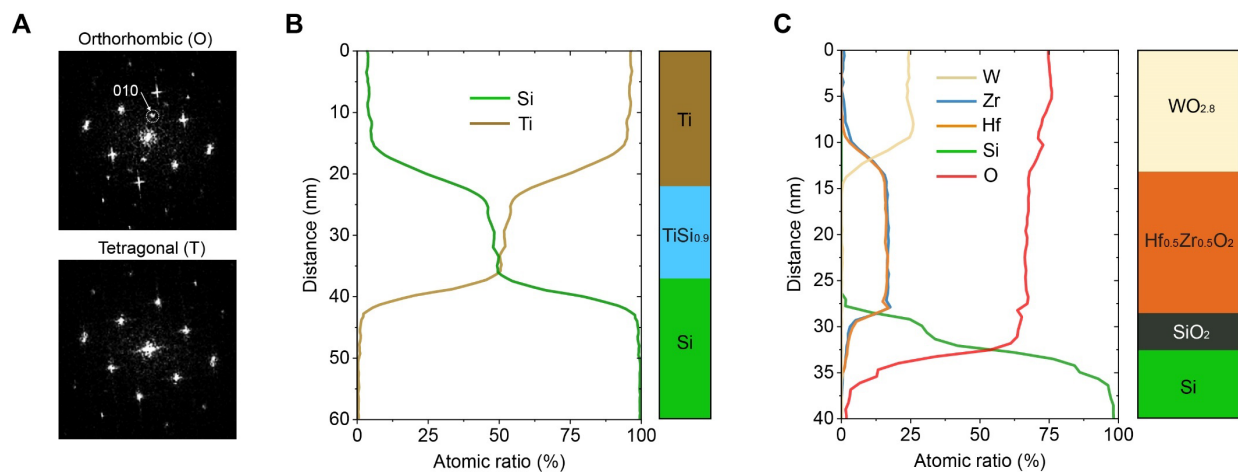


Figure 11. Electron microscopy analysis on synaptic resistor. (A) Fast Fourier transform (FFT) image of an atomic resolution TEM, demonstrating the ferroelectric orthorhombic phase by its additional lattice points (010) originates from the asymmetry in HfZrO unit cell, due to the shifted oxygen ion of orthorhombic phase relative to the dielectric tetragonal phase. (B) EDX analysis on Ti/TiSi_{0.9}/Si stack. (C) EELS analysis on WO_{2.8}/Hf_{0.5}Zr_{0.5}O₂/SiO₂/Si memory stack.

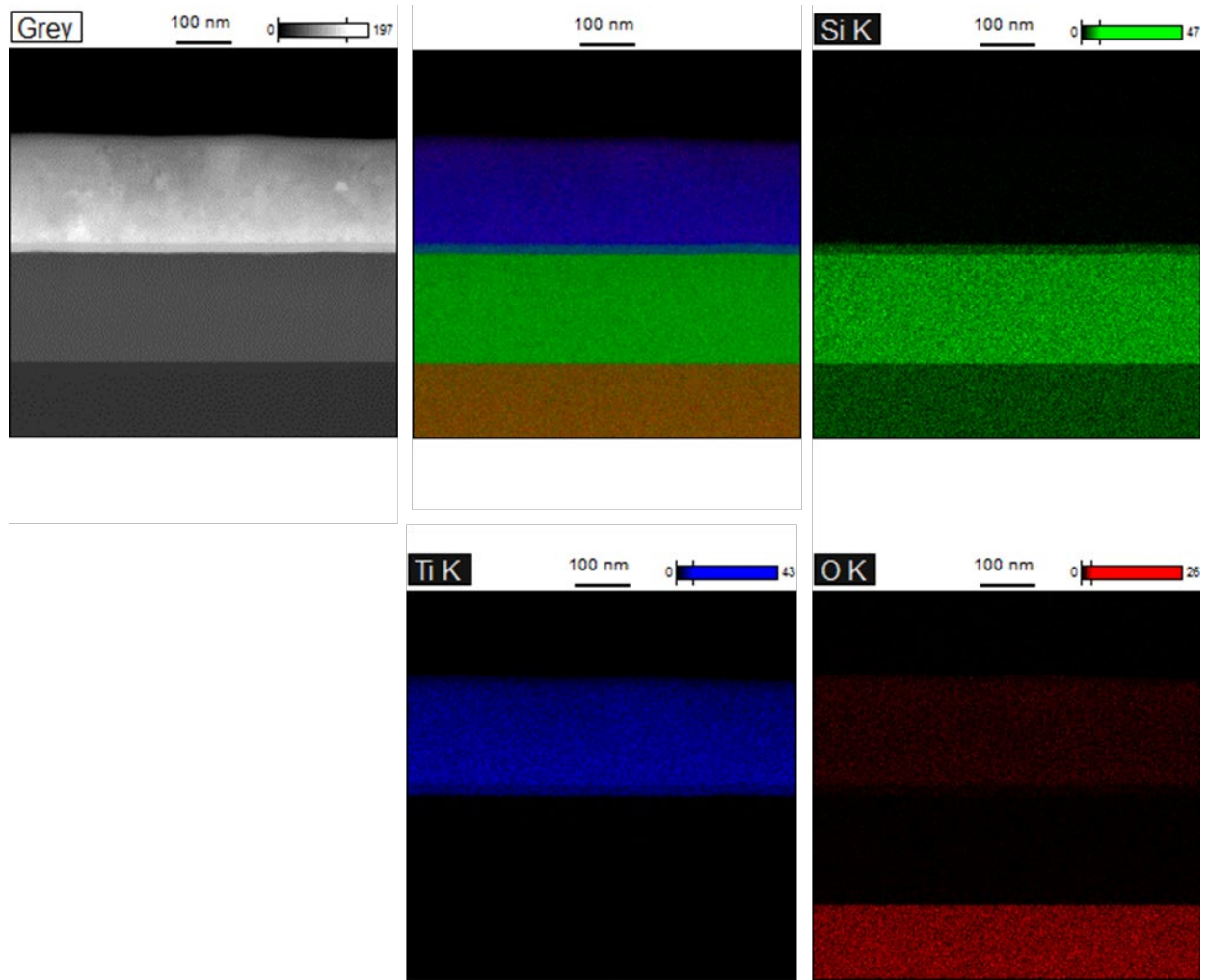


Figure 12. Energy dispersive x-ray (EDS) spectroscopy plot over (top) Ti/TiSi_{0.9}/Si (bottom) junction.

$P4_2/nmc$ Tetragonal phase

$Pca2_1$ Orthorhombic phase (ferroelectric)

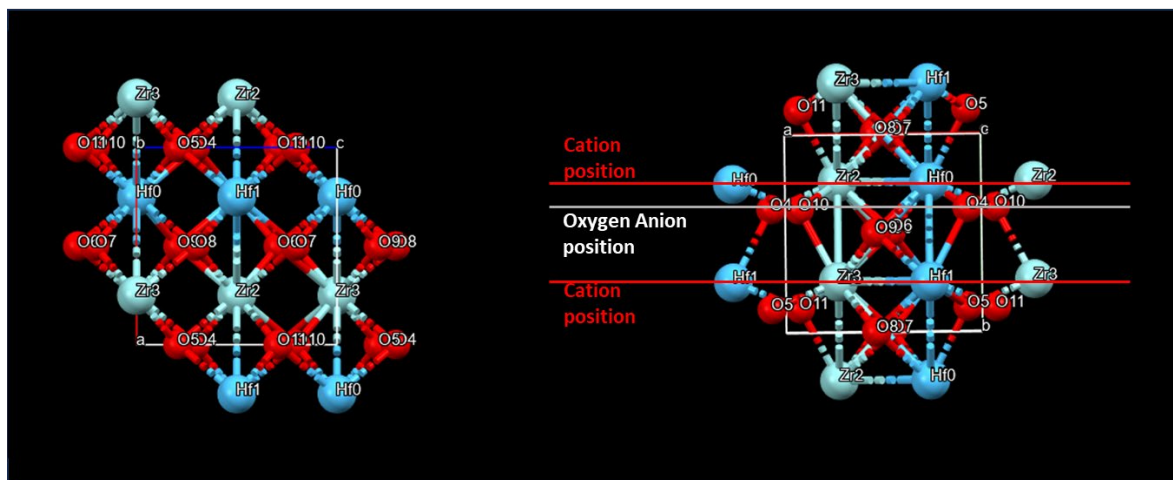


Figure 13. Simulated atomic unit cell of tetragonal vs. orthorhombic phase of HfZrO.

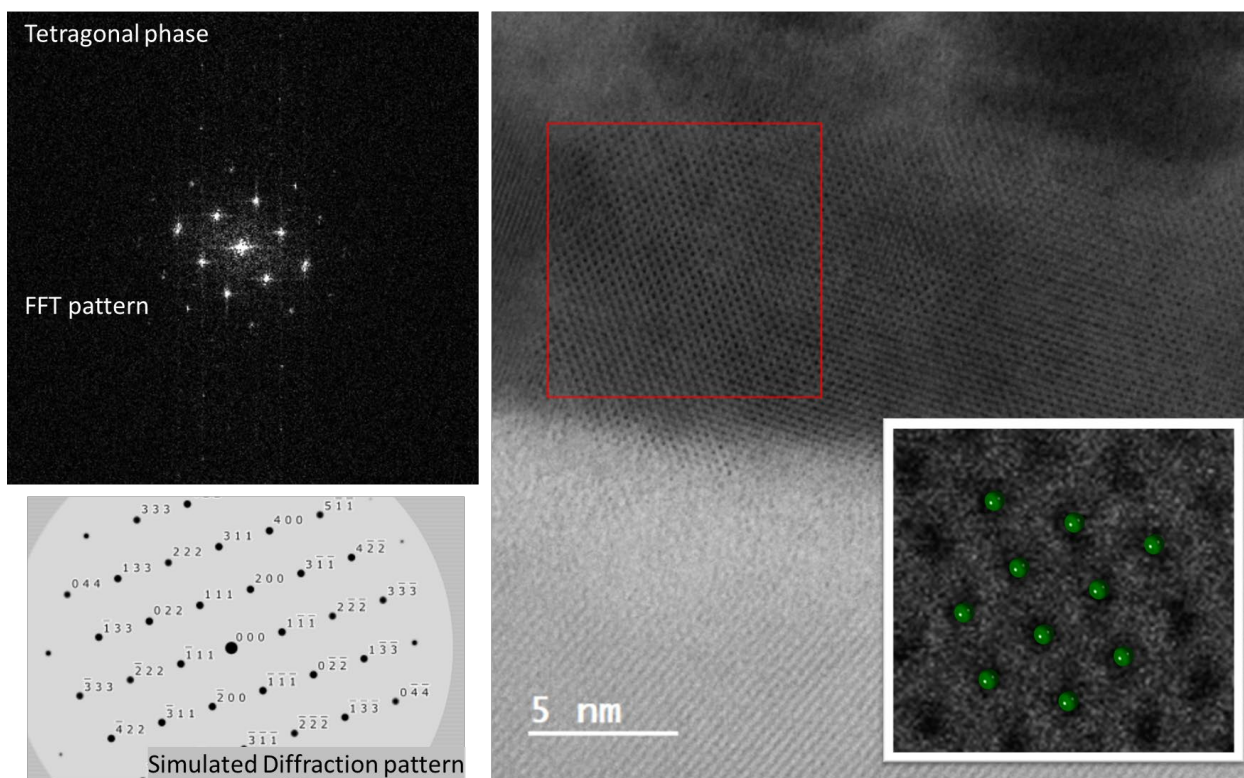


Figure 14. Fast Fourier transform (FFT) image (top left), simulated diffraction pattern (bottom left), and simulated 2D atomic lattice fitting (inset) on tetragonal HfZrO TEM image.

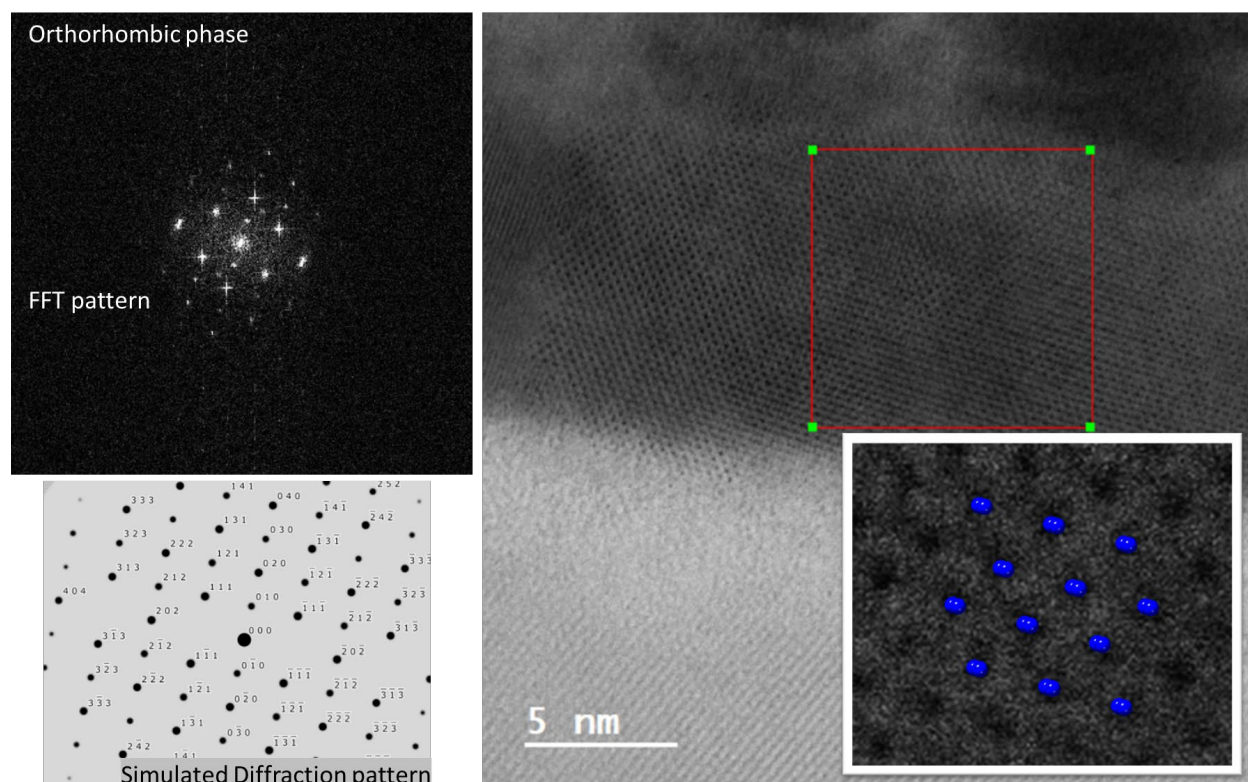


Figure 15. Fast Fourier transform image (top left), simulated diffraction pattern (bottom left), and simulated 2D atomic lattice fitting (inset) on orthorhombic HfZrO TEM image. The added asymmetry in the orthorhombic phase relative to the tetragonal phase is evidenced by the additional lattice points in the FFT image.

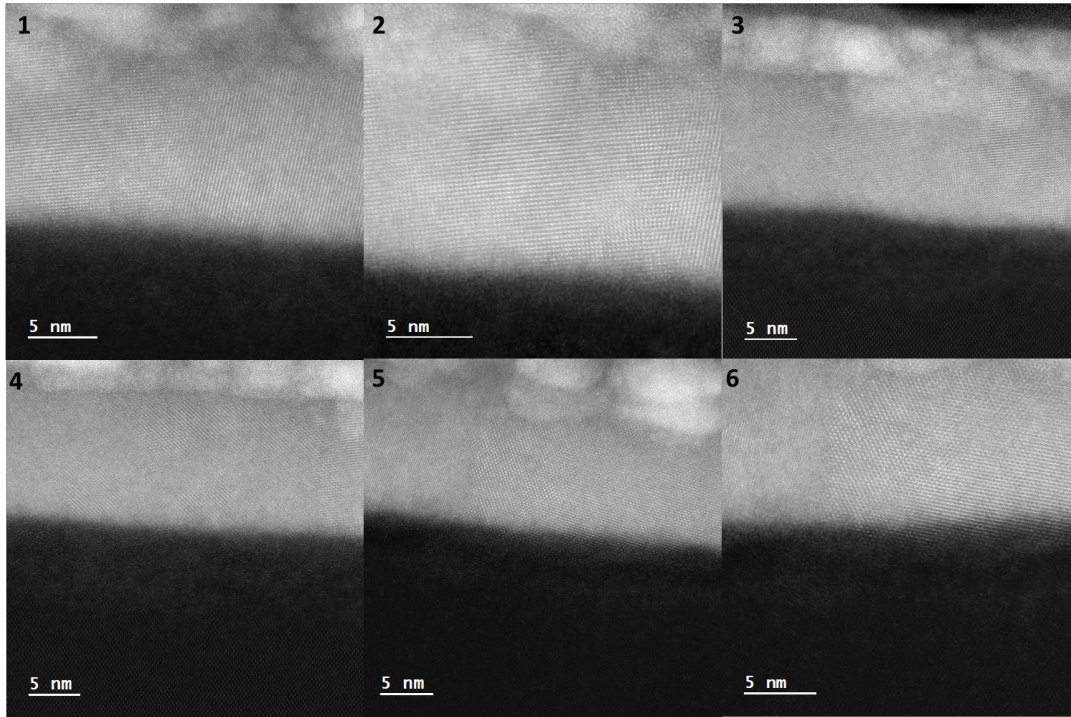


Figure 16. Annular dark-field images of (top) $\text{WO}_{2.8}/\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2/\text{SiO}_2/\text{Si}$ (bottom) stack.

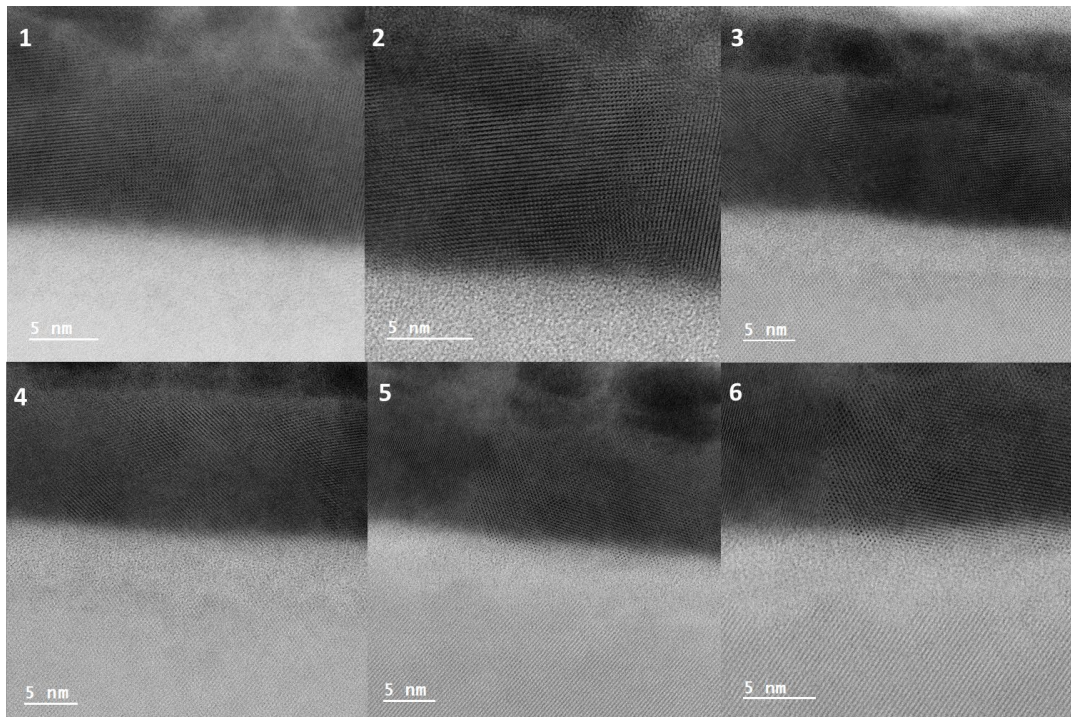


Figure 17. Bright-field images. The numbering corresponds to the annular dark-field images.

2.3. Device Characterization

2.3.1. Nonlinear Analog Conductance Tuning

The analog tuning functionality of synaptic resistors was tested by the application of input pulse x and feedback z pulses with various amplitudes, numbers, and widths on their input and output electrodes, respectively. After 690 paired x and z pulses with amplitudes $x = z$ and a width of 2 ms, the percentage changes in w , $(\Delta w/w_0) \times 100\%$, increased from 0 to 102% when the voltage amplitudes increased from 1.2 V to 3 V; $\Delta w/w_0$ decreased from 0 to -95% when x and z decreased from -1.2 V to -3 V; when the voltage amplitudes $-1.2 V \leq x = z \leq 1.2 V$, $|\Delta w/w_0| \approx 0$ (Figure 18B). The synaptic resistor conductance was gradually tuned in analog mode by applying individual paired pulses with amplitudes $x = z = \pm 3 V$ and a width of 2 ms (Figure 18C). After the synaptic resistor experienced x or z pulses with $-3 V \leq x \leq 3 V$ and $z = 0 V$, or $x = 0 V$ and $-3 V \leq z \leq 3 V$, $|\Delta w/w_0| \approx 0$ (Figure 18B and 18C). Synaptic resistors in the circuit executed the learning algorithm, $\dot{w}_{nm} = \alpha z_n x_m$ with the modification coefficient $\alpha \geq 0$ when $3 V \geq x_m = z_n \geq 1.2 V$, and $\alpha \leq 0$ when $-3 V \leq x_m = z_n \leq -1.2 V$. When $z_n = 0$, and $3 V \geq x_m \geq 1.2 V$, or $-3 V \leq x_m \leq -1.2 V$, the circuit executed the learning algorithm $\dot{w}_{nm} = z_n x_m = 0$, and the inference algorithm $I_{nm} = w_{nm} x_m$ concurrently.

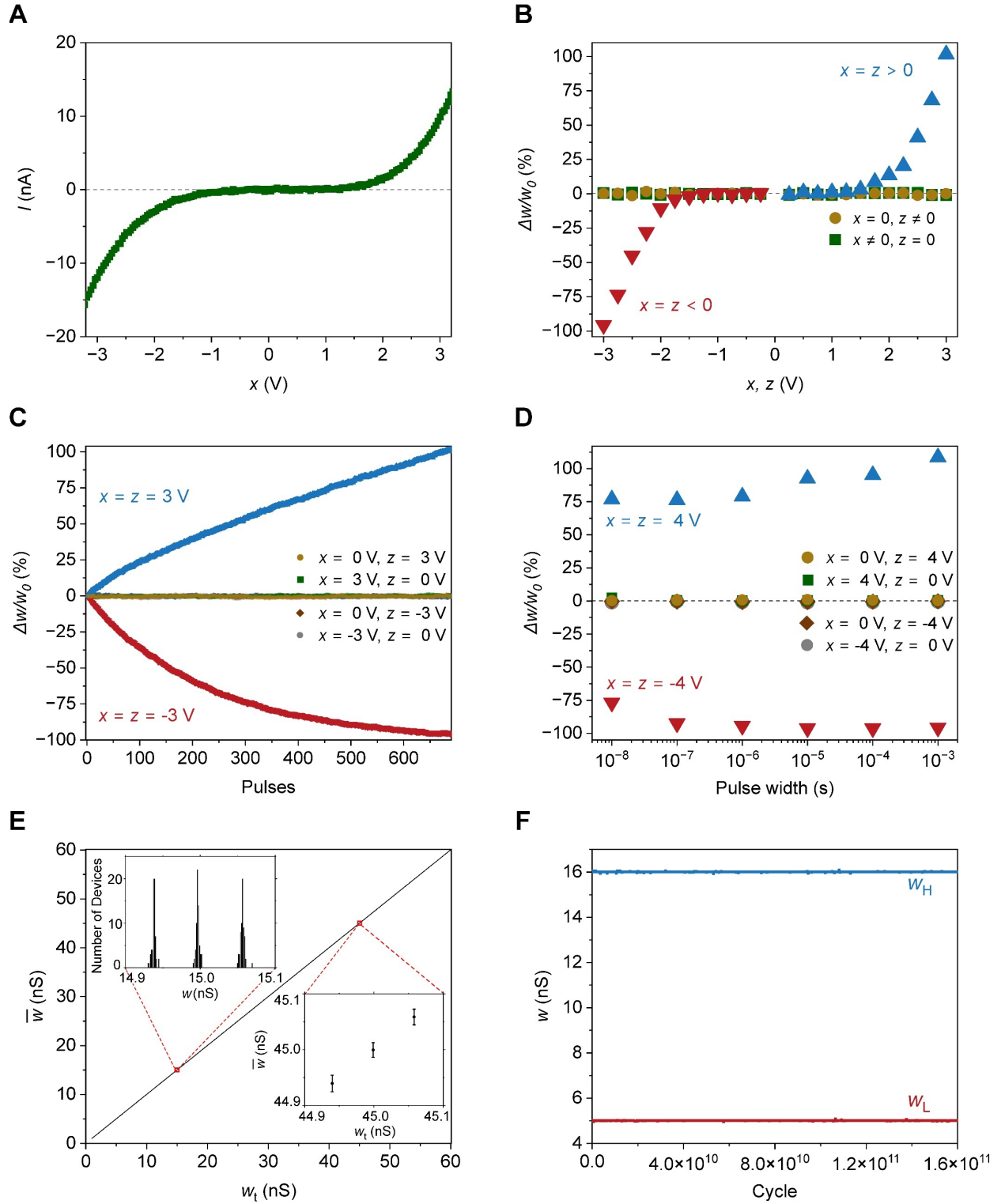


Figure 18. The electric properties of synaptic resistors. (A) The current, I , on the output electrode of a synaptic resistor is displayed versus voltage, x , applied on the input electrode of the synaptic

resistor with respect to its grounded output and reference electrodes. (B) The percentage changes in the synaptic resistor conductance, $(\Delta w/w_0) \times 100\%$, are plotted versus various x and z pulses concurrently applied on the input and output electrodes under the conditions of $-3 V \leq x = z < 0$ (red triangles), $0 < x = z \leq 3 V$ (blue triangles), $-3 V \leq x \leq 3 V$ and $z = 0$ (green squares), and $-3 V \leq z \leq 3 V$ and $x = 0$ (brown circles). (C) The percentage changes in the synaptic resistor conductance, $(\Delta w/w_0) \times 100\%$, induced by various x and z pulses with $x = z = -3 V$ (red triangles), $x = z = 3 V$ (blue triangles), $x = 0$ and $z = 3 V$ (brown circles), $x = 3 V$ and $z = 0$ (green squares), $x = -3 V$ and $z = 0$ (grey circles), and $x = 0$ and $z = -3 V$ (brown diamonds) are plotted against the numbers of applied pulses. (D) The percentage changes in conductance w , $(\Delta w/w_0) \times 100\%$, of a synaptic resistor are shown against the widths of the applied x and z pulses on its input and output electrodes, respectively, after the synaptic resistor experienced pulses with a total duration of 500 ms and amplitudes $x = z = -4 V$ (red triangles), $x = z = 4 V$ (blue triangles), $x = 0$ and $z = 4 V$ (brown circles), $x = 4 V$ and $z = 0$ (green squares), $x = -4 V$ and $z = 0$ (grey circles), and $x = 0$ and $z = -4 V$ (brown diamonds). (E) The average analog conductance values, $\bar{w}$, of 64 synaptic resistors are plotted against 1000 targeted analog conductance values, w_t , evenly separated by $\Delta w_t = 59 pS$. The inset in the top-left corner shows the distribution of the synaptic resistor conductance values when $w_t = 14.938 nS$, $14.997 nS$, and $15.056 nS$. The inset in the bottom-right corner shows $\bar{w}$ with $\pm 3\sigma_w$ standard deviations against w_t at $44.940 nS$, $44.999 nS$, and $45.058 nS$. (F) A synaptic resistor was iteratively modified to its high and low analog conductance values, w_H and w_L , which are displayed as blue and red squares.

The synaptic resistor conductance was also effectively modified by applying paired pulses with amplitudes $x = z = \pm 4 \text{ V}$ and widths ranging from 10 ns to 1 ms (Figure 18D), indicating that the device conductance can be modified within 10 ns, which has also been observed in HfZrO-based transistors. The analog learning accuracy was assessed through the fine-tuning of 64 synaptic resistors in the circuit precisely toward 1000 distinct values, w_t , ranged between 1 nS and 60 nS and separated evenly by $\Delta w_t = 59 \text{ pS}$ by applying a train of paired $x_m = z_n$ pulses. As shown in Figure 18E and Figure 19A and 19B, the disparities between the average conductance values of the 64 synaptic resistors and their respective w_t , $|\bar{w} - w_t|$, are smaller than 1.5 pS or 2.5 % of Δw_t , and the standard divisions σ_w at different $\bar{w}$ are below 6 pS or 10 % of Δw_t , which are significantly superior to those of CNT-based and Al oxide-based synaptic resistors (Figure 19C). The 1000 dynamically tunable conductance levels and the precise conductance modification accuracy of 36 pS ($6\sigma_w$) could be attributed to the progressive switching of the multiple individual nanoscale ferroelectric domains in the Hf_{0.5}Zr_{0.5}O₂ layer (Figure 11C). To assess their learning endurance, synaptic resistors were iteratively adjusted to their high and low analog conductance values. No noticeable deterioration in device conductance was observed over 1.6×10^{11} switching cycles (Figure 18F). HZO-based devices typically degraded after $\sim 10^6$ switching cycles, mainly induced by oxygen vacancies in the Hf_{0.5}Zr_{0.5}O₂ layer. In the WO_{2.8}/Hf_{0.5}Zr_{0.5}O₂ heterojunction of the synaptic resistor, the oxygen vacancies with higher defect energy levels in the Hf_{0.5}Zr_{0.5}O₂ layer tended to migrate toward the W oxide layer with lower defect energy levels (Figure 11C), significantly enhancing the switching cycles of the Hf_{0.5}Zr_{0.5}O₂ layer and converting the W oxide layer to metallic reference electrode⁴⁷. The conductance values of synaptic resistors were measured at 40 °C and 85 °C versus time after the synaptic resistors were modified by paired pulses (Figure 20A and

20B). Like the dynamic changes of synaptic weights in neurobiological networks, the conductance values of the synaptic resistors changed at the initial testing stage, possibly due to incomplete charge compensation in the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer. However, the ferroelectric dipoles and device conductance values remained stable for long-term memory⁴⁸.

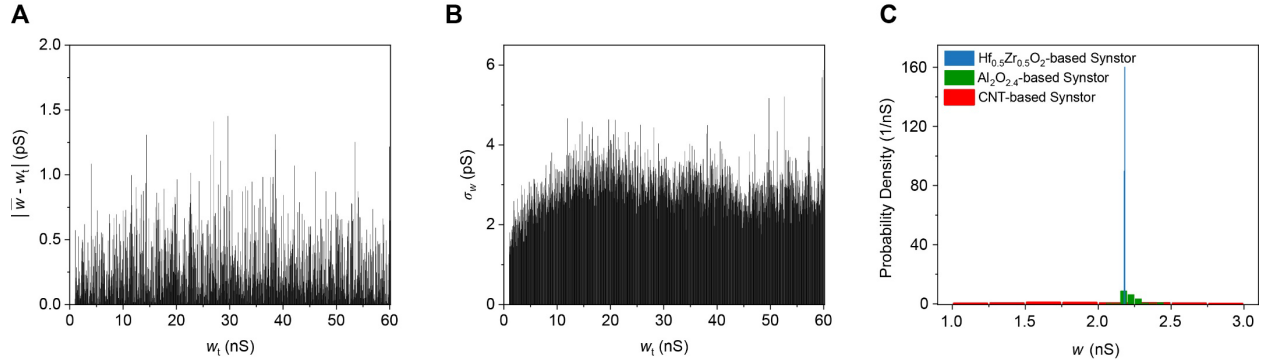


Figure 19. Learning accuracy, rate, endurance, and nonvolatile memory tests of synaptic resistors. Learning accuracy: (A) The absolute values of the differences between the average analog conductance values $\bar{w}$ and the targeted analog conductance values of synaptic resistors w_t , $|\bar{w} - w_t|$, and (B) the standard deviations of w , σ_w , are plotted against their corresponding targeted analog conductance values of synaptic resistors, w_t . (C) The probability densities of conductance values w of the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ -based synaptic resistors (blue), $\text{Al}_2\text{O}_{2.4}$ -based synaptic resistors¹⁰ (green), and CNT- based synaptic resistors⁸ (red) are plotted after the synaptic resistors were tuned toward their targeted conductance values.

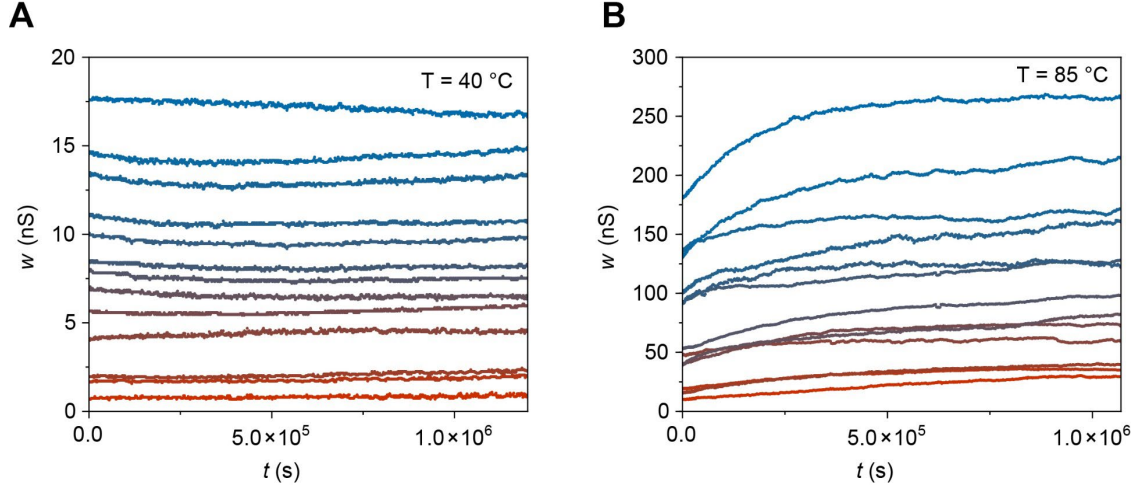


Figure 20. Nonvolatile memory tests of synaptic resistors and device conductance comparison. After synaptic resistors were modified to different initial analog conductance values, the conductance value of each synaptic resistor, $w(t)$, measured at (A) $40\text{ }^{\circ}\text{C}$ and (B) $85\text{ }^{\circ}\text{C}$ is shown by dots in different colors as a function of time t over 10^6 s.

2.3.2. Energy-band Structure of Synaptic Resistor

The ability to dynamically adjust device conductance in an analog mode is attributed to the gradual switching of multiple individual nanoscale ferroelectric domains in the ferroelectric $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer. Based on the properties of synaptic resistors and their circuits, we have designed and simulated synaptic resistors and their circuits simulator (Spectre, Cadence) by a Technology Computer-Aided Design (TCAD) simulator (Sentaurus Device, Synopsys). The simulator performed numerical calculations of the device physics by solving Poisson's equation describing the electrostatics and drift-diffusion carrier transport under a set of boundary conditions defined by the device structure. Quasi-stationary simulations were conducted under various voltage biases on the input/output electrodes of the synaptic resistors with respect to the grounded reference electrodes. The band diagrams of the synaptic resistors were extracted from

the simulations, and the electronic properties of the synaptic resistors were analyzed by the simulations.

As shown in the simulated band structures of a synaptic resistor (Figure 21), a pair of positive or negative x and z voltage pulses with $x = z = 3.0\text{ V}$ or -3.0 V generates an electric field in the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, $\mathbf{E} \approx 790\text{ kV/cm}$ or -790 kV/cm (Figs. 4A and 4B), which satisfies the conditions $\mathbf{E} > \mathbf{E}_t^+$ or $\mathbf{E} < \mathbf{E}_t^-$, where $\mathbf{E}_t^+$ and $\mathbf{E}_t^-$ represent the positive and negative threshold electric fields required to switch individual nanoscale ferroelectric domains in the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer. Under these conditions, the paired voltage pulses progressively switch the individual ferroelectric domains within the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, modify their ferroelectric dipoles with $\Delta\mathbf{P} > 0$ or $\Delta\mathbf{P} < 0$, and attract or repel the holes in the p-type Si channel, resulting in the increase or decrease of the synaptic resistor conductance in analog mode (Figure 18C). When a synaptic resistor experienced single x and z voltage pulses with $x = 3\text{ V}$ and $z = 0$, or $x = -3\text{ V}$ and $z = 0$, or no pulses ($x = z = 0$), the electronic potential mainly drops across the Schottky junction between the Si channel and $\text{TiSi}_{0.9}$ layer, and the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer locates beyond the Schottky junction has an electric field, $\mathbf{E} = -4, -153, \text{ or } -3\text{ kV/cm}$, respectively (Figure 21D). Under these conditions, $\mathbf{E}$ satisfies $\mathbf{E}_t^- < \mathbf{E} < \mathbf{E}_t^+$, thus cannot modify the ferroelectric dipoles ($\Delta\mathbf{P} \approx 0$), or the synaptic resistor conductance (Figure 18B and 18C).

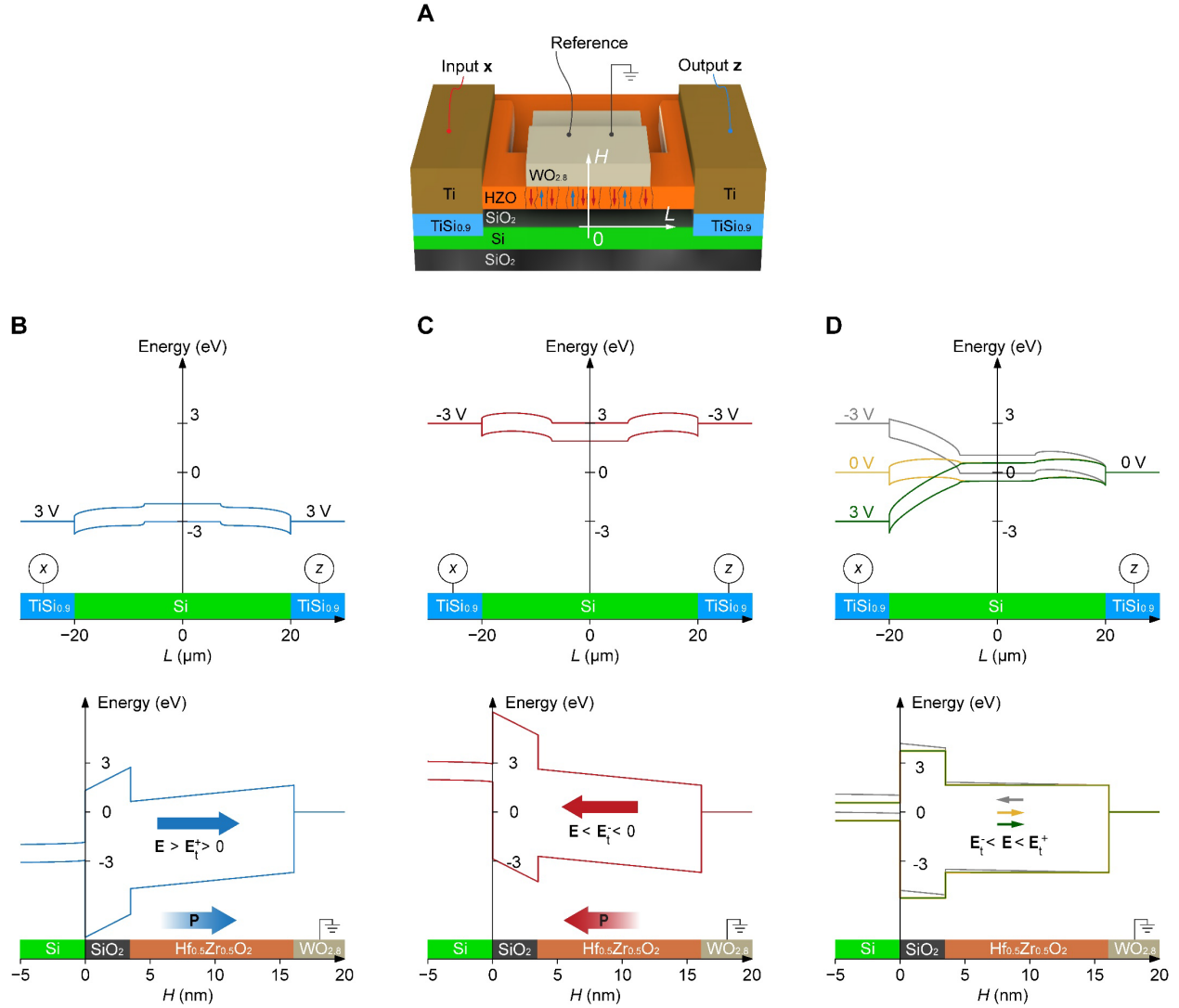


Figure 21. The energy-band structures of a synaptic resistor. (A) A 3D schematic illustration of a synaptic resistor composed of a grounded $\text{WO}_{2.8}$ reference electrode, $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, SiO_2 dielectric layer, and Si channel along the H axis, and a $\text{TiSi}_{0.9}$ input electrode, Si channel, and $\text{TiSi}_{0.9}$ output electrode along the L axis. The $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer is composed of multiple ferroelectric domains, indicated by red or blue arrows, and dielectric domains, depicted without arrows. The simulated energy-band structures along the L axis (top) and along the H axis (bottom) of a synaptic resistor under the conditions of (B) the potentials at the input electrode $x = 3 \text{ V}$ and the output electrode $z = 3 \text{ V}$, (C) $x = z = -3 \text{ V}$, and (D) $x = 3 \text{ V}$ and $z = 0$ (green

lines), $x = -3 \text{ V}$ and $z = 0$ (grey lines), and $x = z = 0$ (yellow lines). The top scale (L) represents the lateral distance from the input electrode to the output electrode, and the bottom scale (H) represents the vertical distance from the Si channel to the $\text{WO}_{2.8}$ reference electrode.

According to the TCAD simulation, when the synaptic resistor channel length and the reference electrode length are scaled down to 100 nm and 40 nm, respectively, the synaptic resistor still functions properly for inference and learning (Figure 22). A synaptic resistor circuit must satisfy the following conditions to execute inference and learning concurrently: (i) To avoid the phase shifts of input pulses, $\mathbf{x}$, and backpropagating pulses, $\mathbf{z}$, on the input and output electrodes, $M^2 R_e C_e, N^2 R_e C_e \ll t_d$, where M represents the number of input electrodes, N represents the number of output electrodes, t_d denotes the pulse duration, R_e denotes the resistance of the electrode per synaptic resistor unit, and C_e denotes the capacitance of the electrode per synaptic resistor unit. In the synaptic resistor circuit reported in this work, $C_e \approx 10^{-19} \text{ F}$, $R_e \approx 0.27 \text{ } \Omega$, $t_d = 10 \text{ ns}$; thus, $M, N \ll 10^6$. In a circuit of synaptic resistors with a 40 nm reference electrode, $C_e \approx 10^{-17} \text{ F}$, $R_e \approx 0.54 \text{ } \Omega$; thus, $M, N \ll 10^5$. (ii) To avoid significant voltage variations on the input and output electrodes induced by the sneak currents passing through the synaptic resistors, the worst-case voltage variation on an input or output electrode induced by the maximal sneak currents, $\Delta V_{Max} < k_B T / q \approx 26 \text{ mV}$, where q is the charge of an electron, k_B is the Boltzmann constant, and T is room temperature. We have designed and simulated a synaptic resistor circuit based on a TSMC 40 nm process, which has an average synaptic resistor conductance $\langle w \rangle \approx 0.2 \text{ nS}$, $R_e \approx 0.54 \text{ } \Omega$; thus, $M, N \lesssim 10^4$, i.e. a single crossbar synaptic resistor circuit can be scaled up to $M, N = 10^4$ with $M \cdot N = 10^8$ synaptic resistors.

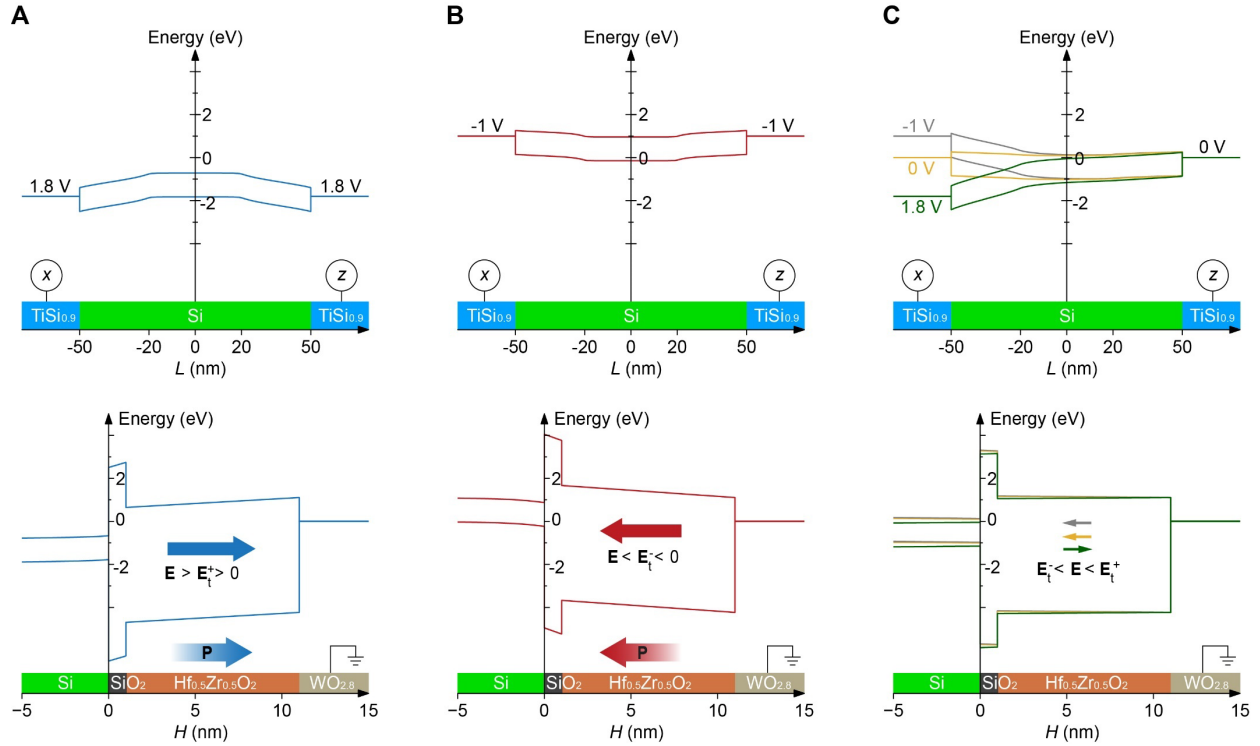


Figure 22. The energy-band structures of a synaptic resistor with a channel length of 100 nm.

The simulated energy-band structures of the $\text{TiSi}_{0.9}$ input electrode, Si channel, and $\text{TiSi}_{0.9}$ output electrode along the L axis in fig. S5A, and of the grounded $\text{WO}_{2.8}$ reference electrode, a 10 nm-thick $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, an 1 nm-thick SiO_2 dielectric layer, and Si channel along the H axis in fig. S5A of a synaptic resistor with a channel length of 100 nm under the conditions of (A) the potential at the input electrode $x = 1.8 \text{ V}$, the potential at the output electrode $z = 1.8 \text{ V}$, (B) $x = z = -1.0 \text{ V}$, and (C) $x = z = 0 \text{ V}$ (yellow lines), $x = -1.0 \text{ V}$ and $z = 0$ (deep blue lines), and $x = 1.8 \text{ V}$ and $z = 0$ (green lines). When the electric field, E , in the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer satisfies $E > E_t^+ > 0$ in a, or $E < E_t^- < 0$ in b, with E_t^+ and E_t^- as the positive and negative electric threshold fields to switch the dipole, P , of the ferroelectric domains in the $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ layer, P is modified to align along E . When $E_t^- < E < E_t^+$ in (C), thus $\Delta P \approx 0$.

3. Neuromorphic Computing using Synaptic Resistor Circuits

The synaptic resistor circuits are integrated into a neuromorphic computing platform built on a printed circuit board (PCB) using a customized adapter (Ironwood Electronics). The chip layout groups all output and reference electrodes within a row, forming a crossbar with all input electrodes when inserted into the adapter. The adapter interfaces with a custom PCB, which connects to a LabVIEW compactRIO controller (NI-9063) with FPGA, which controls 32 analog output channels, 32 analog input channels, and 32 digital input/output channels. Voltage pulses are applied using the analog output channels, and device currents are converted to voltages and measured by the analog input channels through an operational amplifier in an inverting configuration.

The integrate and fire neuron in the PCB were based on the basic functions of biological neurons in the Hodgkin-Huxley model⁴⁹. Current from synaptic resistors flows into a capacitor, increasing its potential. On the other hand, a leakage current flows through a resistor, decreasing the potential over time. This potential is proportional to the integration of the current minus the leakage current over time. When the potential reaches a threshold, a Schmitt trigger composed of transistors switches to generate an output pulse, which resets the potential back to zero, allowing the capacitor to restart the integration of the current. When $I < I_{th} \approx 30 \text{ nA}$, the threshold current of the neuron circuit, no actuation pulse was triggered. When $I_{th} \leq I \leq I_s \approx 60 \text{ nA}$, the saturation current of the neuron circuit, the average firing rate of the actuation pulses increased monotonically with increasing I . When $I > I_s$, the average firing rate of the actuation pulses was saturated at $\sim 14 \text{ Hz}$. The output pulses from the neuron circuits were converted to actuation signals, y , by the interface circuit (FPGA, Xilinx, Kintex-7). When a y_n pulse was triggered at moment $t = t_n$ from the n^{th} output neuron of the synaptic resistor circuit, a train of negative (or

positive) z_n pulses were triggered at the n^{th} (or n^{th} complimentary) output electrodes within the time window $t_n < t < t_n + t_w$, and a train of positive (or negative) z_n pulses were triggered at the n^{th} (or n^{th} complimentary) output electrodes within the time window $t_n + \tau_+ < t < t_n + \tau_+ + t_w$, where $\tau_+ = 1180 \text{ ms}$ and $t_w = 40 \text{ ms}$ represents the duration of the time window of the pulse trains.

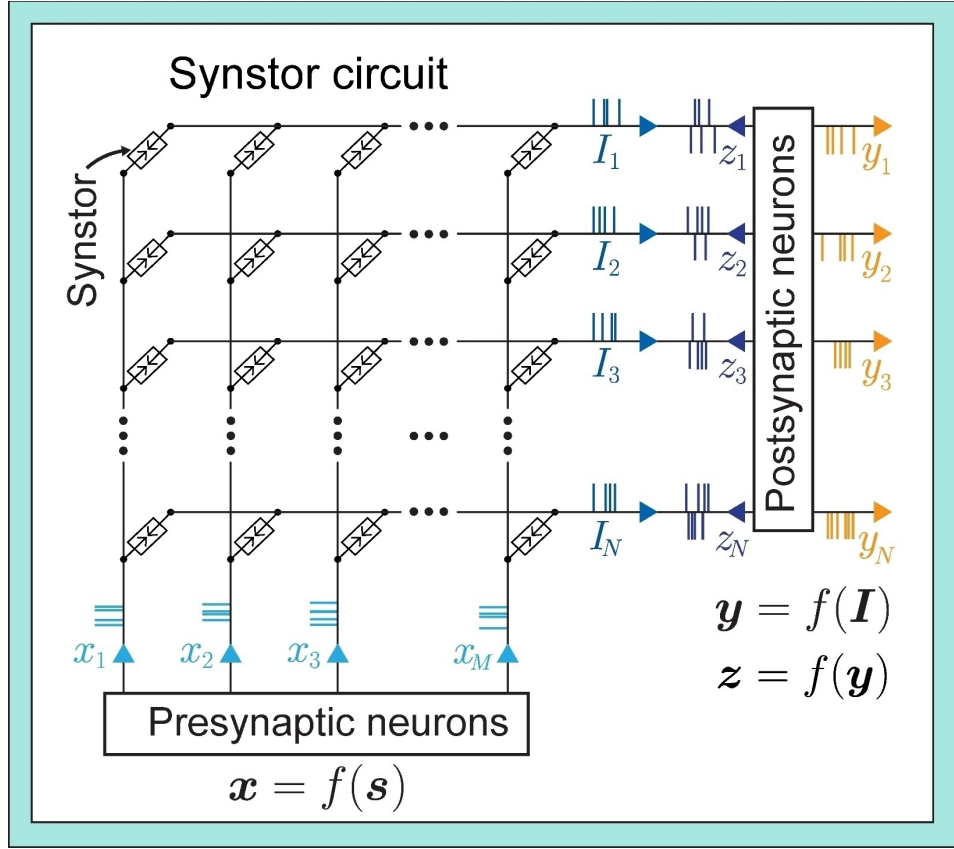


Figure 23. The synaptic resistor chip is connected to neurons at the PCB through the adapter to constitute a neuromorphic computing circuit. Input voltage pulses from input presynaptic neurons drive the current from M synaptic resistors, which are connected to N output neurons. Depending on the logic of the integrate and fire circuit, the feedback pulse from the output neuron is applied to synaptic resistors to modify their conductance.

3.1. Autonomous Drone Navigation

Synaptic resistor circuit navigated a drone toward a target position by avoiding obstacles in a simulated environment with time-varying strong wind (Figure 25). Obstacles were detected by a simulated camera on the drone, and a moving target position ($\hat{\mathbf{s}}$) was identified to avoid the obstacles and minimize an objective function $F = \frac{1}{2}(\mathbf{s} - \hat{\mathbf{s}})^2$, with $\mathbf{s} - \hat{\mathbf{s}}$ as the deviations of the drone position ($\mathbf{s}$) from its target positions (Figure 24D). In the windy environment, the speeds and directions of winds changed randomly during the flight (Figure 24A). During the synaptic resistor experiments, $\mathbf{s} - \hat{\mathbf{s}}$ was converted to input voltage pulses, $\mathbf{x}$, by an interface circuit to generate currents via the synaptic resistor circuit according to its inference algorithm, $\mathbf{I} = \mathbf{W} \mathbf{x}$ (Equation 1). The currents triggered actuation signals, $\mathbf{y}$, via a neuron and an interface circuit to drive the drone. The conductance matrixes ($\mathbf{W}$) of the synaptic resistor circuit were initialized to random values before each experiments started (Figure 28) to ensure that the circuit had no prior learning experience or predefined algorithm. Concurrently, voltage pulses, $\mathbf{z}$, meeting the conditions $\bar{\mathbf{z}} = 0$ and $\overline{z_n y_n} \leq 0$ were triggered at the output electrodes of the synaptic resistor circuit (Methods) to modify $\mathbf{W}$ according to the learning rule $\dot{\mathbf{W}} = \alpha \mathbf{z} \otimes \mathbf{x}$ (Equation 2).

Human operators also navigated a drone toward the target position by avoiding obstacles in the same simulated windy environment (Figure 25, and Figure 24B). The human operators had no prior experience with the drone control system, ensuring that they learned while navigating the drone. During the human experiments, the operators visually saw the analog numbers, $\mathbf{s} - \hat{\mathbf{s}}$, displayed on a monitor, converted $\mathbf{s} - \hat{\mathbf{s}}$ to voltage pulses, $\mathbf{x}$, and induced currents $\mathbf{I}$ via synapses in their neurobiological circuits, triggering actuation signal, $\mathbf{y}$, on a keyboard to drive the drone. Concurrently, the synaptic weight matrixes $\mathbf{W}$ could be modified by following synaptic learning rule $\dot{\mathbf{W}} = \alpha \mathbf{z} \otimes \mathbf{x}$ during the real-time learning process in a “Super-Turing” computing mode.

In the control experiments for computer-based ANN to navigate a drone toward the target position by avoiding obstacles in the simulated windy environment, various ANN structures and learning parameters were tested, and the optimal ANN with the shortest learning time was then deployed to navigate the drone (Figure 25, Figure 24C, and Figure 28). The time required to execute learning algorithm was significantly longer than that to execute the inference algorithm (Figure 28F), thus it was infeasible to execute the real-time learning on the computer, and the inference and learning had to be executed sequentially. The $\mathbf{s} - \hat{\mathbf{s}}$ signals were converted to input signals, $\mathbf{x}$, and induced output currents $\mathbf{I} = \mathbf{W}(n) \mathbf{x}$, triggering actuation signals, $\mathbf{y}$, from the ANN to drive the drone in its n^{th} flight episode. The saved experimental data, $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{W}(n)$ from the n^{th} episode were sent back to the computer to execute a reinforcement learning algorithm (Methods) to modify $\mathbf{W}(n)$ to $\mathbf{W}(n + 1)$, which reset the ANN to drive the drone at the $(n + 1)^{th}$ flight episode iteratively until the drone crashed into an obstacle or reached the final target (Figure 25, and Figure 24C).

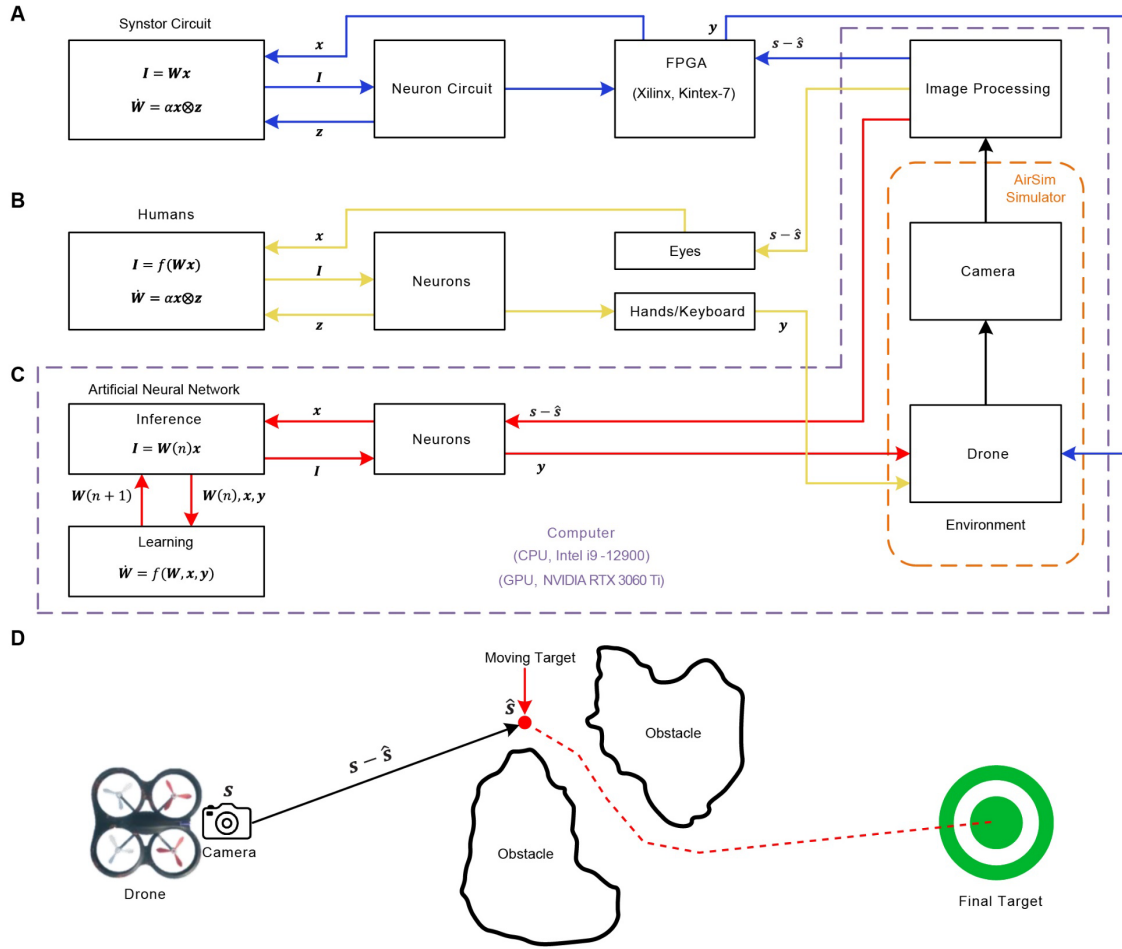


Figure 24. Schematics of experimental systems. Experimental systems for (A) a synstor circuit with real-time learning, (B) humans with real-time learning, and (C) a computer-based ANN with off-line learning to drive a drone toward a target position in a simulated environment. (D) Obstacles are detected by a simulated camera on the drone, a moving target position ($\hat{s}$) is identified to avoid the obstacles, and the deviation of the drone position (s) from its target position, $s - \hat{s}$, is input to the synstor circuit, humans, or ANN via interface circuits.

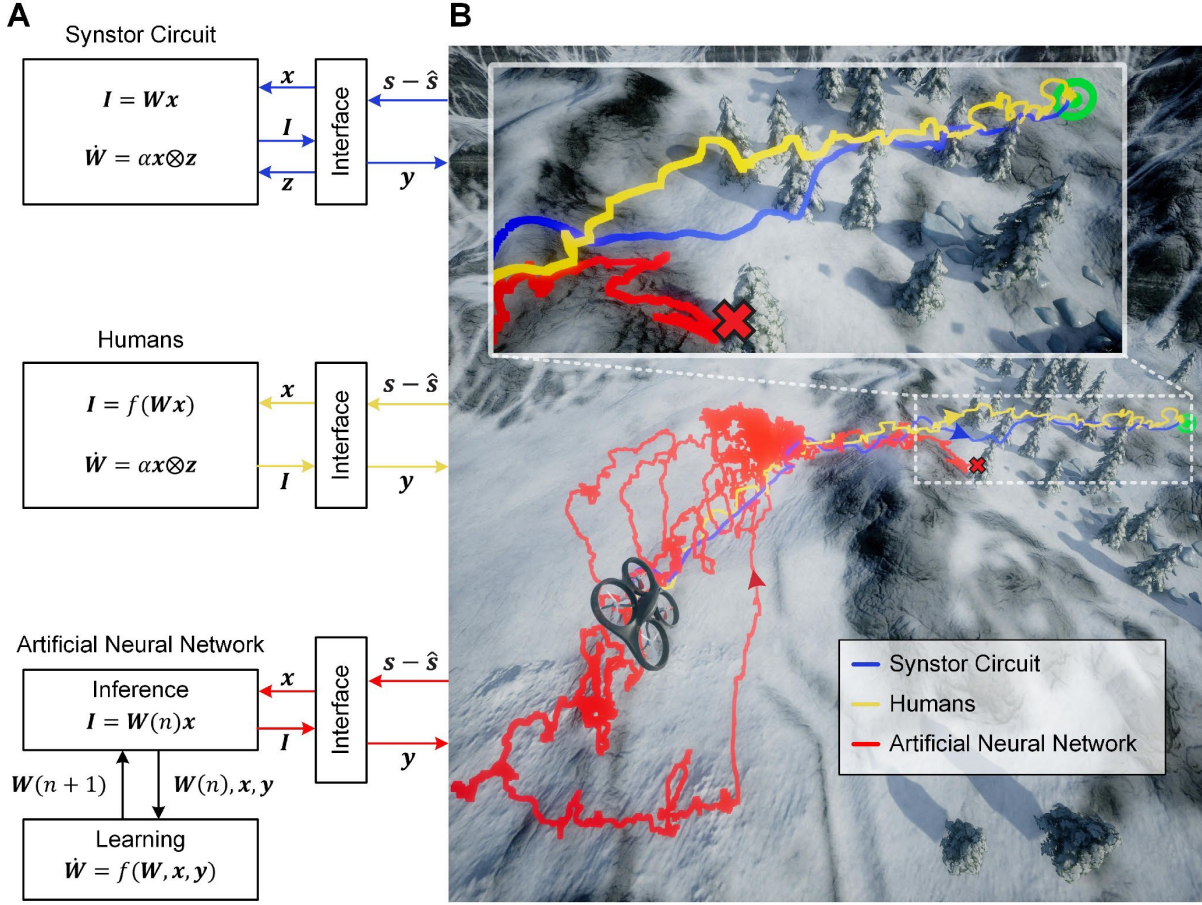


Figure 25. Experiments to navigate a drone by a synaptic resistor circuit, humans, and ANN. (A) a synaptic resistor circuit (top), humans (middle), and a computer-based ANN (bottom) (B) drive a drone toward a target position by avoiding obstacles in a simulated windy environment. The starting point of the flights is marked by a drone image, and the final target position is marked by a green target symbol. The flying traces are shown for the drone driven by the synaptic resistor circuit (blue line), the human (yellow line), and the ANN (red line). The inset shows the details in the area near the target with tree obstacles. The red crash symbols indicate that the drone crashed into the tree.

3.1.1. Experimental Setup

The experiments were performed by flying a drone in an environment (Figure 25) simulated by running an AirSim program (Microsoft Co.) in a Dell computer with an Intel i9-12900 CPU and NVIDIA RTX 3060 Ti GPU. During the flight, wind randomly changed its direction and speed from 0 to 17.5 m/s (Figure 24). The positions of the drone and target were determined by a simulated global positioning system. To avoid collisions with obstacles such as mountains and trees in the environment, the depth profiles of the obstacles were detected by a simulated camera mounted on the drone and analyzed to identify local flight target positions with obstacle clearance and minimal objective function $F = \frac{1}{2}(\mathbf{s} - \hat{\mathbf{s}})^2$, where $\mathbf{s} - \hat{\mathbf{s}}$ as the deviations of the drone position ($\mathbf{s}$) from its target position ($\hat{\mathbf{s}}$) along upward, downward, forward, backward, left, right, clockwise, and anticlockwise (yaw) dimensions (Figure 24D).

The conductance matrix of the synstor circuit, $\mathbf{W}$, were set to random values before each learning experiment started (Figure 27). As shown in Figure 25, and Figure 24A, $\mathbf{s} - \hat{\mathbf{s}}$ was converted to voltage pulses, $\mathbf{x}$, with an amplitude of -3 V or 4.2 V and a duration of 10 ns , which were input to the synstor circuit by the interface circuit (FPGA, Xilinx, Kintex-7). The firing rate of $\mathbf{x}$ pulses increased monotonically with increasing $\mathbf{s} - \hat{\mathbf{s}}$. $\mathbf{x}$ generated currents, $\mathbf{I} = \mathbf{W} \mathbf{x}$ (Equation 1), via the synstor circuit. The currents, $\mathbf{I}$, flowed through the synstors to interface circuits and triggered the actuation pulses, $\mathbf{y}$, to drive the drone, and the voltage pulses, $\mathbf{z}$, applied on the output electrodes of the synstor circuit to modify the synstor conductance matrix, $\mathbf{W}$, by following the learning rule, $\dot{\mathbf{W}} = \alpha \mathbf{z} \otimes \mathbf{x}$.

In the experiments with the drone driven by humans under the same environment of the synstor circuit experiments (Figure 25, and Figure 24B, 14 human operators without any prior knowledge of the drone and its control system visually received $\mathbf{s}_L$ signals displayed on a

computer monitor, and were instructed to drive the drone toward its local target position and avoid the collision with obstacles by minimizing the $\mathbf{s} - \hat{\mathbf{s}}$ values. The $\mathbf{s} - \hat{\mathbf{s}}$ signals were processed by the neurobiological circuits in the human brains, triggering actuation pulses $\mathbf{y}$ to drive the drone by pressing eight keys in a keyboard (Figure 24B). The firing rates of $\mathbf{y}$ pulses were proportional to the keystroke times.

The experiments with the drone driven by a computer-based ANN were performed under the same environment as the synstor circuit and human experiments (Figure 25, and Figure 24C). Before the learning experiment started, the digital synaptic weight matrix, $\mathbf{W}$, in the ANN was also set to random values, the same as for the trials of the synstor circuit. The time required to execute learning algorithm was significantly longer than that to execute the inference algorithm during the flight (Figure 28F), thus it is infeasible to execute the real-time learning on the computer, and the inference and learning are executed sequentially. The ANN running on a Dell computer (Intel i9-12900 CPU and NVIDIA RTX 3060 Ti GPU) received the $\mathbf{s} - \hat{\mathbf{s}}$ signals, converted $\mathbf{s} - \hat{\mathbf{s}}$ to input signals, $\mathbf{x}$, induced output current $\mathbf{I} = \mathbf{W}(n) \mathbf{x}$, triggering actuation pulses, $\mathbf{y}$, to modify $\mathbf{s}_L$ at n^{th} episode to drive the drone for ~ 30 s (Figure 25C). After the n^{th} flight episode ended, the saved data, $\mathbf{s} - \hat{\mathbf{s}}$, $\mathbf{y}$, and $\mathbf{W}(n)$ from the n^{th} episode were sent back to the computer to execute a reinforcement learning algorithm (Supporting Materials) for ~ 550 s to modify $\mathbf{W}(n)$ to $\mathbf{W}(n + 1)$, which reset the ANN to drive the drone at the $(n + 1)^{th}$ flight episode iteratively until the drone crashed into an obstacle or reached the final target. ANNs with 1) 8 neurons at its input layer, no hidden layer, and 9 neurons at its output layer interconnected by 72 synapses; and 2) 8 neurons at its input layer, 128 neurons at its hidden layer, and 9 neurons at its output layer interconnected by 2176 synapses were tested to drive the drone toward a local target position in the learning experiments with a learning rate of 10^{-8} , 2×10^{-8} , 5×10^{-8} , and

10^{-7} , respectively (Figure 28). The optimal ANN with no hidden layer and a learning rate of 5×10^{-8} had the shortest learning time (Figure 28I), and were then used in the learning experiments to drive the drone toward the final target.

In the learning processes of synstor circuits, humans, and ANN, the change rate of the objective function, $\dot{F}$, is a nonlinear function of F , but can be best-fitted by a linear dynamic model $\dot{F} = -(F - F_e)/T_L$ and its solution over the learning processes (Figure 26), $F(t) = (F(0) - F_e) e^{-t/T_L} + F_e$, where the fitting parameter T_L represents the average initial learning time, and F_e represents the quasi-equilibrium objective function when $t \gg T_L$ and $d\bar{F}/dt \approx 0$.

When voltage pulses were applied on a 8×8 crossbar synstor circuit, the average power consumption of the circuit (excluding the power consumptions of interface circuits), $P = \mathbf{I} \otimes \mathbf{x} = (\mathbf{w}\mathbf{x}) \otimes \mathbf{x} \approx w_T V_a^2 D_p$, where w_T denotes the total conductance of the synstors in the circuit, V_a denotes the magnitude of pulses, D_p denotes the average duty-cycle of the pulses. During the concurrent signal-processing and learning in synstor circuit for the drone, $w_T \approx 10 \text{ nS}$, $V_a = 4.2 \text{ V}$, and $D_p = 0.014$, thus $P \approx 158 \text{ nW}$. Based on the circuit simulation, the total power consumption of the 8 neuron circuits was $\sim 0.78 \text{ nW}$.

The speeds for sequentially executing the inference and learning programs on the computer were estimated to be 3.9 KFLOPS and 2.0 GFLOPS , according to analysis from Python toolkits "keras_flops" and "pyperf", respectively. With a computing energy efficiency of 3.1 GFLOPS/W , the power consumption for sequentially executing the inference and learning programs on the computer was $12 \text{ }\mu\text{W}$ and 6.3 W , respectively. The aggregate power consumption of the ANN for sequential inference and learning is 6.3 W (Figure 26).

3.1.2. Results

The synaptic resistor circuit, human operators, and ANN successfully learned to drive the drone toward its target position across the mountain area without trees or strong wind while minimizing the objective function $F = \frac{1}{2}(\mathbf{s} - \hat{\mathbf{s}})^2$. The $F - t$ curves can be best-fitted by $F(t) = (F(0) - F_e) e^{-t/T_L} + F_e$ to extrapolate the average learning time T_L and the equilibrium objective function F_e when $t \gg T_L$ and $dF/dt \approx 0$, which represents the performance of the drone when $\mathbf{W}$ has been modified to $\widehat{\mathbf{W}}$. The average T_L (4.4 s) of the synaptic resistor circuit in multiple trials is shorter than the average T_L (6.6 s) of the human operators, and significantly superior to the average T_L (127,568 s) of the ANN in multiple trials (Figure 26D). The drone performance, represented by F_e , of the drone driven by the synaptic resistor circuit (0.10 a.u.) and humans (0.25 a.u.) is significantly superior to that of the drone driven by the ANN (0.44 a.u.) in their multiple trials (Figure 26E). When the drone entered the environment with trees and strong wind, only the synaptic resistor circuit and the majority of human operators were able to navigate the drone to its final target position, demonstrating their ability to adapt to the environment with the gale and avoid collisions with the trees; in contrast, the ANN consistently failed to adapt to the complex environment with trees and gale, resulting in collisions in multiple trials (Figure 6). The adaptability to the changing environment is evaluated by the success rate of learning to drive the drone to its final target without collisions in multiple trials. The adaptability (100%) of the synaptic resistor circuit is better than that (64%) of the human operators and significantly superior to that (0%) of the ANN (Figure 26F). The power consumption of the synaptic resistor circuit (158 nW) for the concurrent execution of the inference and learning algorithms was also seven orders of magnitudes lower than the aggregate power consumption of the computer (6.3 W) executing the learning and inference algorithms sequentially.

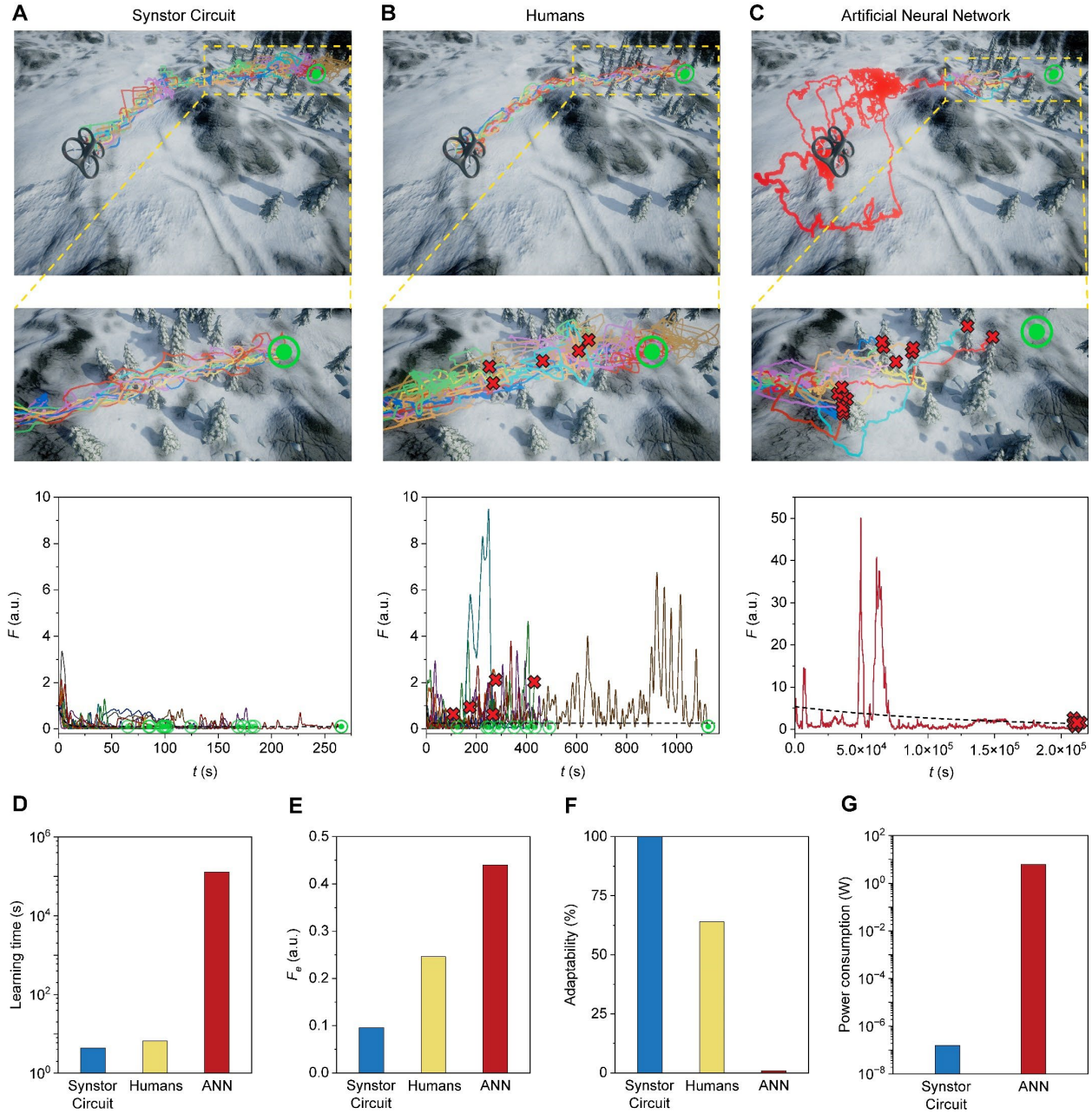


Figure 26. Experimental results of drone navigations. (Top row) The flying traces are shown for a drone driven by (A) a synaptic resistor circuit, (B) human operators, and (C) an ANN toward a target position by avoiding obstacles in a simulated windy environment. The starting point of the flights is marked by drone images. (Middle row) The zoom-in images show the details of the flying traces in the area near the target with tree obstacles. (Bottom row) objective functions,

$F = \frac{1}{2}(s - \hat{s})^2$, with $s - \hat{s}$ as the deviations of the drone position (s) from its target position ($\hat{s}$) versus accumulative flight and learning time t . The $F - t$ curves are best-fitted by $F(t) = (F(0) - F_e) e^{-t/T_L} + F_e$ (dashed lines). The red crash symbols indicate that the drone crashed into the tree, and the final target position is marked by green target symbols. Comparison of average (D) learning time (T_L), (E) equilibrium objective function (F_e), (F) adaptability to the changing environment, and (G) computing power consumption of the synaptic resistor circuit, human operators, and ANN in multiple trials. The adaptability to changing environment is evaluated by the successful rate of learning how to drive the drone to its final target without collisions in multiple trials. The power consumption of the synaptic resistor circuit is for concurrent inference and learning, while the power consumption of the ANN is the aggregate of the power consumptions for sequential inference and learning.

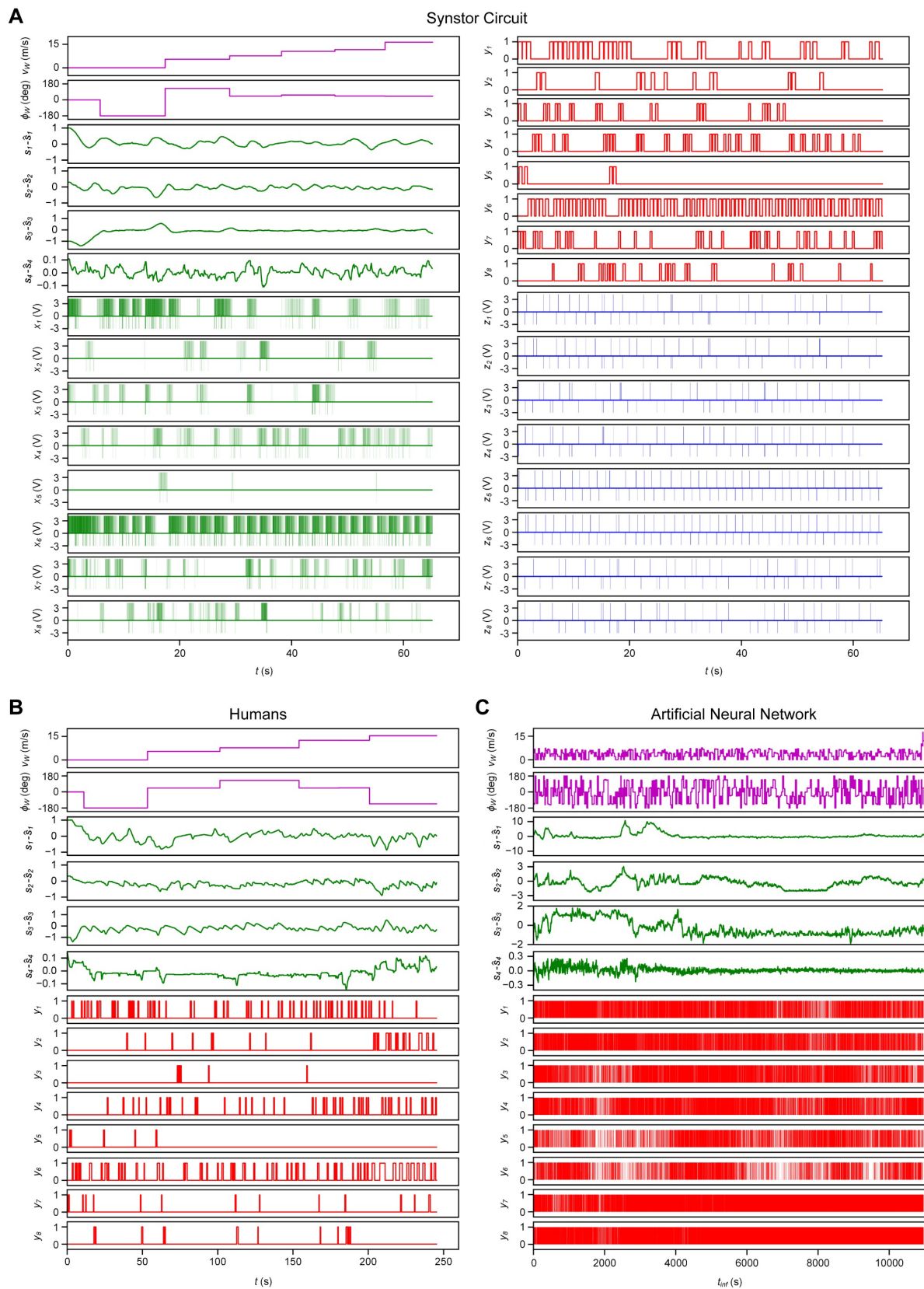


Figure 27. Data from the experiments with a drone driven by a synstor circuit, a human operator, and an ANN running on a computer. The wind speed V_W , the wind direction measured in degrees clockwise from due north ϕ_W , the deviations of the drone position from the local target measured along the left-right $s_1 - \hat{s}_1$, up-down $s_2 - \hat{s}_2$, forward-backward $s_3 - \hat{s}_3$, clockwise-counterclockwise yaw $s_4 - \hat{s}_4$, directions, and the actuation pulses, $y_1, y_2, y_3, y_4, y_5, y_6, y_7$, and y_8 to move the drone along left, right, up, down, forward, backward, clockwise, and counterclockwise yaw directions are displayed against the time t in the experiments with a drone driven by (A) a synstor circuit, (B) a human operator, and (C) an ANN running on a computer. The input voltage pulses $x_1, x_2, x_3, x_4, x_5, x_6, x_7$, and x_8 converted from the $s - \hat{s}$ signals along left, right, up, down, forward, backward, clockwise, and counterclockwise yaw directions, respectively, and the back-propagating voltage pulses $z_1, z_2, z_3, z_4, z_5, z_6, z_7$, and z_8 on the output electrodes associated with the actuation pulses, $y_1, y_2, y_3, y_4, y_5, y_6, y_7$, and y_8 , respectively, are also displayed against the time t .

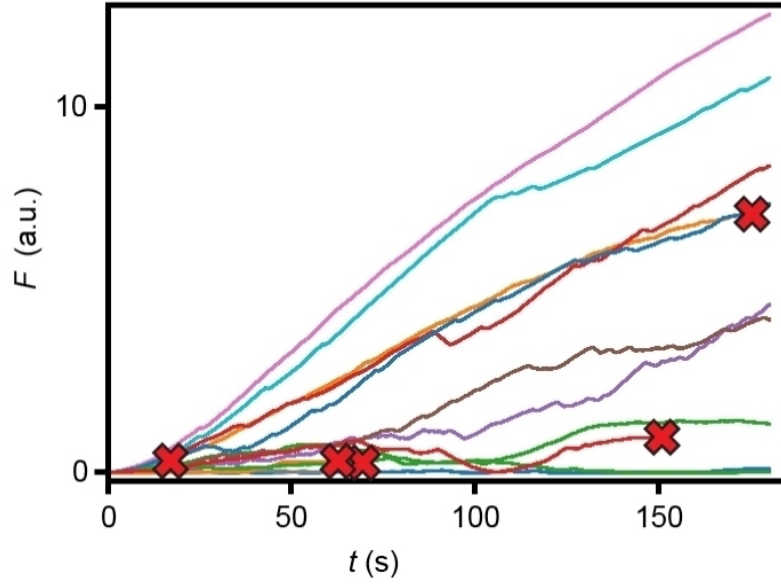


Figure 28. Experiments to drive a drone by a synstor circuit without learning. During the flight processes of a drone driven by a synstor circuit with its conductance matrix set to random values and no learning, the objective functions, $F = \frac{1}{2}(\mathbf{s} - \hat{\mathbf{s}})^2$, with $\mathbf{s} - \hat{\mathbf{s}}$ as the deviations of the drone position ($\mathbf{s}$) from its target position ($\hat{\mathbf{s}}$) are displayed versus flight time t . The red crash symbols indicate that the drone crashed into the tree.

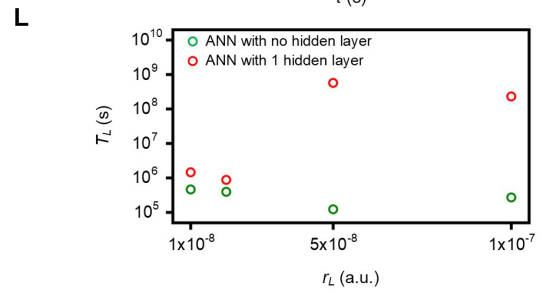
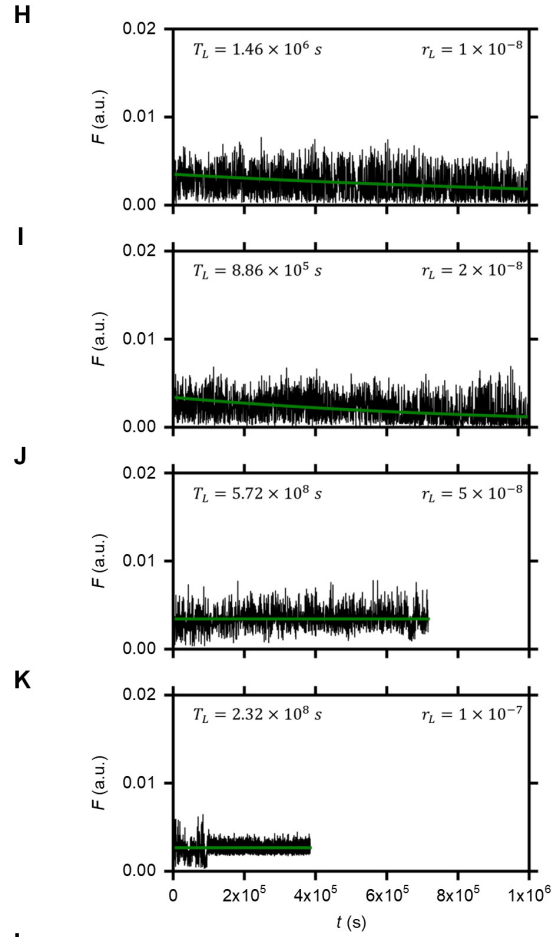
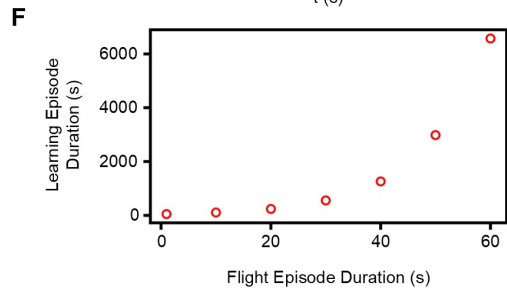
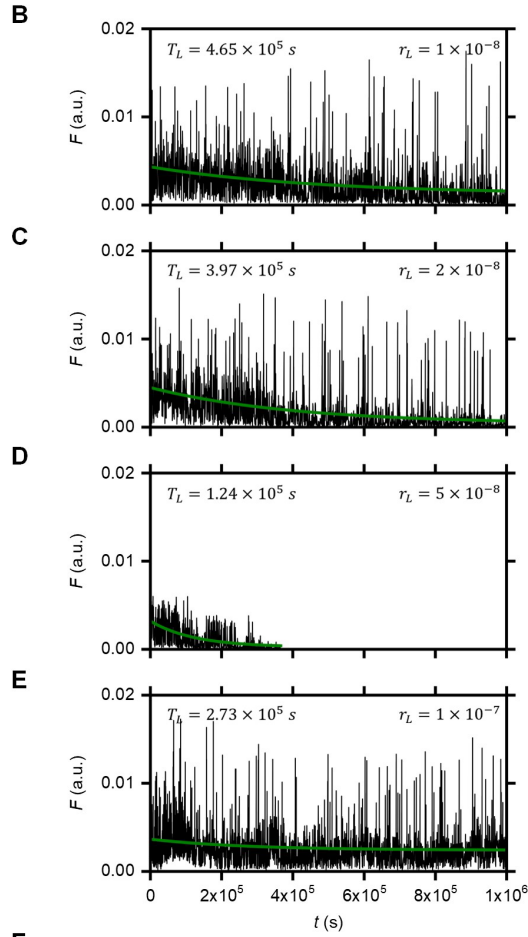
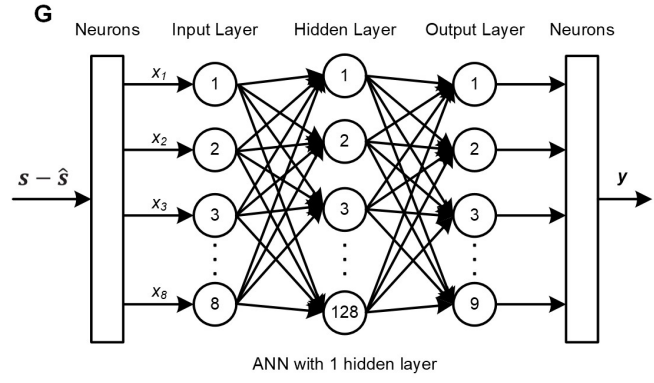
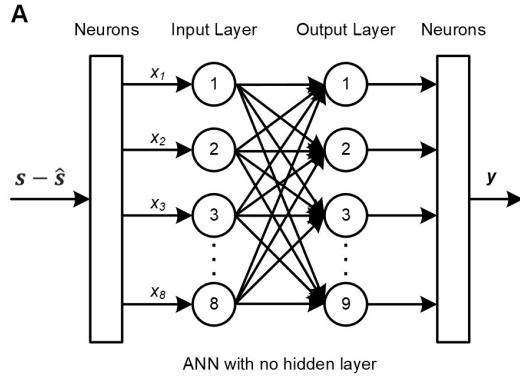


Figure 29. Learning experiments to drive a drone by ANN with different structures and learning parameters. (A) a schematic shows the structure of an ANN with 8 neurons at its input layer, no hidden layer, and 9 neurons at its output layer. The local objective functions, $F_L = \frac{1}{2} \mathbf{s}_L^2$, with $\mathbf{s}_L$ as the deviations of the drone position from its local target position, are displayed versus accumulative flight and learning time t during the flight processes of the drone driven by the ANN shown in (A) with a learning rate of r_L equal to (B) 10^{-8} , (C) 2×10^{-8} , (D) 5×10^{-8} , and (E) 10^{-7} . (F) the duration of learning episode to execute learning algorithm based on the data from a flight episode is shown against the duration of flight episode. (G) A schematic shows the structure of an ANN with 8 neurons at its input layer, 128 neurons at its hidden layer, and 9 neurons at its output layer. The objective functions, $F = \frac{1}{2} (s - \hat{s})^2$, are displayed versus accumulative flight and learning time t during the flight processes of the drone driven by the ANN shown in (G) with a learning rate of r_L equal to (H) 10^{-8} , (I) 2×10^{-8} , (J) 5×10^{-8} , and (K) 10^{-7} . The $F - t$ curves are best-fitted by $F(t) = (F(0) - F_e) e^{-t/T_L} + F_e$ (green lines) to extrapolate the average learning times T_L . (L) T_L is displayed versus the leaning rate r_L in the ANN shown in (A) (with no hidden layer, green circles) and (G) (with 1 hidden layer, red circles).

3.1.3. Discussion

We introduced an intelligent system based on a synaptic resistor circuit with concurrent real-time inference ($\mathbf{I} = \mathbf{W} \mathbf{x}$) and learning ($\dot{\mathbf{W}} = \alpha \mathbf{z} \otimes \mathbf{x}$) functionality of the brain, operating in Super-Turing analog parallel mode. Theoretical analysis indicates that the simultaneous execution of inference and learning can minimize the objective function $F = \frac{1}{2} (\mathbf{s} - \hat{\mathbf{s}})^2$ of the system, tuning the system states ($\mathbf{s}$) towards their target states ($\hat{\mathbf{s}}$) by modifying $\mathbf{W}$ toward $\widehat{\mathbf{W}} =$

$\arg \min_w F$. A synaptic resistor circuit was fabricated by integrating a heterojunction of a Si channel/SiO₂ dielectric layer/defect-reduced ferroelectric Hf_{0.5}Zr_{0.5}O₂ layer/defect-enhanced WO_{2.8} reference electrode vertically, and Schottky contacts between the Si channel/TiSi_{0.9}/Ti input and output electrodes laterally in each synaptic resistor. Paired voltage pulses (x_m and z_n) applied on the input and output electrodes of a synaptic resistor progressively modify the individual ferroelectric domains within the Hf_{0.5}Zr_{0.5}O₂ layer, thereby adjusting the synaptic resistor conductance in analog mode based on the learning rule ($\dot{W} = \alpha \mathbf{z} \otimes \mathbf{x}$); a single voltage pulse (x_m) applied to the input electrode of a synaptic resistor induces a current through the synaptic resistor channel according to the inference algorithm ($\mathbf{I} = \mathbf{W} \mathbf{x}$). The voltage primarily drops laterally across the Schottky junction and does not modify the ferroelectric domains within the Hf_{0.5}Zr_{0.5}O₂ layer, or the synaptic resistor conductance (i.e. $\dot{w}_{nm} = 0$ when $x_m \neq 0$ and $z_n = 0$) as per the learning rule ($\dot{W} = \alpha \mathbf{z} \otimes \mathbf{x}$). Compared to other neuromorphic devices like memristors^{50–53}, which can only perform inference and learning sequentially, the synaptic resistor circuit uniquely enables concurrent execution of inference and learning, allowing the circuit to operate in the Super-Turing mode. The defect-reduced Hf_{0.5}Zr_{0.5}O₂ layer with multiple switchable ferroelectric domains effectively improves the device performance and facilitates the precise, reliable, fast, and repetitive modification of synaptic resistor conductance with 1000 analog conductance levels within a conductance range of 0 – 60 nS, a learning accuracy of 36 pS, 1.6×10^{11} repetitive learning cycles, and a rapid modification time within 10 ns. In experiments aimed at navigating a drone towards target positions while avoiding obstacles in a simulated complex aerodynamic environment with time-varying wind, without any prior learning, a synaptic resistor circuit and humans executed real-time learning and inference concurrently in Super-Turing mode, in comparison with an ANN run on a computer executing

inference and learning sequentially in Turing mode. During concurrent real-time inference and learning in a synaptic resistor circuit, the conductance (w_{nm}) of each synaptic resistor can be dynamically adjusted based on signals (x_m and z_n) received from the input and output electrodes connected by the synaptic resistor in parallel analog mode. The learning process aims to optimize the objective function (F) in response to environmental changes in real-time. In contrast, during sequential inference and learning processes in the ANN or other neuromorphic resistor circuits operating in Turing mode, their $\mathbf{W}$ matrixes cannot be adjusted during the inference according to real-time environmental changes. The optimal $\mathbf{W}$ matrices are derived during sequential learning processes through serial statistical evaluation of numerous variations of $\mathbf{W}$ matrices and their performance datasets during inference. Hence, the experimental learning times (T_L) of both the synaptic resistor circuit and humans are four orders of magnitude shorter than those of the ANN. The ANN consistently failed to adapt to the complex environment with trees and strong winds, resulting in collisions in multiple trials. In contrast, the synaptic resistor circuit and the majority of humans successfully adapted to the environment with trees and strong winds, driving the drone to its final target without collisions. As evaluated by the equilibrium objective function (F_e) when $\mathbf{W}$ has been modified to $\widehat{\mathbf{W}}$, the performance of the drone driven by the synaptic resistor circuit and humans is significantly superior to that of the drone driven by the ANN. A single-layer synaptic resistor circuit can execute learning and inference concurrently in real-time, dynamically optimizing the $\mathbf{W}$ matrix, triggering optimal output actuation signals ($\mathbf{y}$) to minimize the objective function (F). Conversely, during their sequential inference and learning processes, the ANN or other neuromorphic resistor circuits expend additional energy and time to store the $\mathbf{W}$, $\mathbf{x}$, and $\mathbf{y}$ data, execute the learning algorithms in its computing units, and transfer the data back-and-forth between different circuits. The power consumption of the synaptic

resistor and neuron circuits (158 nW) for real-time concurrent inference and learning is seven orders of magnitude lower than the aggregate power consumption (6.3 W) of the computer executing the learning and inference algorithms sequentially in the ANN. The conductance and power consumption of the synaptic resistors can be further reduced by scaling down the synaptic resistors. On the basis of simulations, a HfZrO-based synaptic resistor can be scaled down to a channel length of 100 nm (fig. S10), and a HfZrO-based synaptic resistor crossbar circuit can be scaled up to 10^8 synaptic resistors with 10^4 input and output channels using the nanoscale fabrication techniques of the HfZrO-based ferroelectric transistor circuit. Synaptic resistor circuits offer a brain-inspired Super-Turing computing platform for AI systems with extremely low power consumption, high-speed real-time learning and inference, and agile adaptability to dynamic complex environments.

3.2. Associative Memory Circuit (AMC)

In this study, we analyzed and derived the hardware-compatible update rule of the modern Hopfield network and introduced an associative memory circuit (AMC) based on the synaptic resistor crossbar array (Figure 30). Then we trained the network to classify the MNIST hand-written digits by Hebbian/anti-Hebbian learning and unlearning algorithm, which shows the capability to learn the prototypes of different features and classify the digits with learned memories. The learning algorithm based on the localized learning rule of synaptic resistors enables concurrent inference and learning functionality, which demonstrates its potential as a brain-like computing circuit.

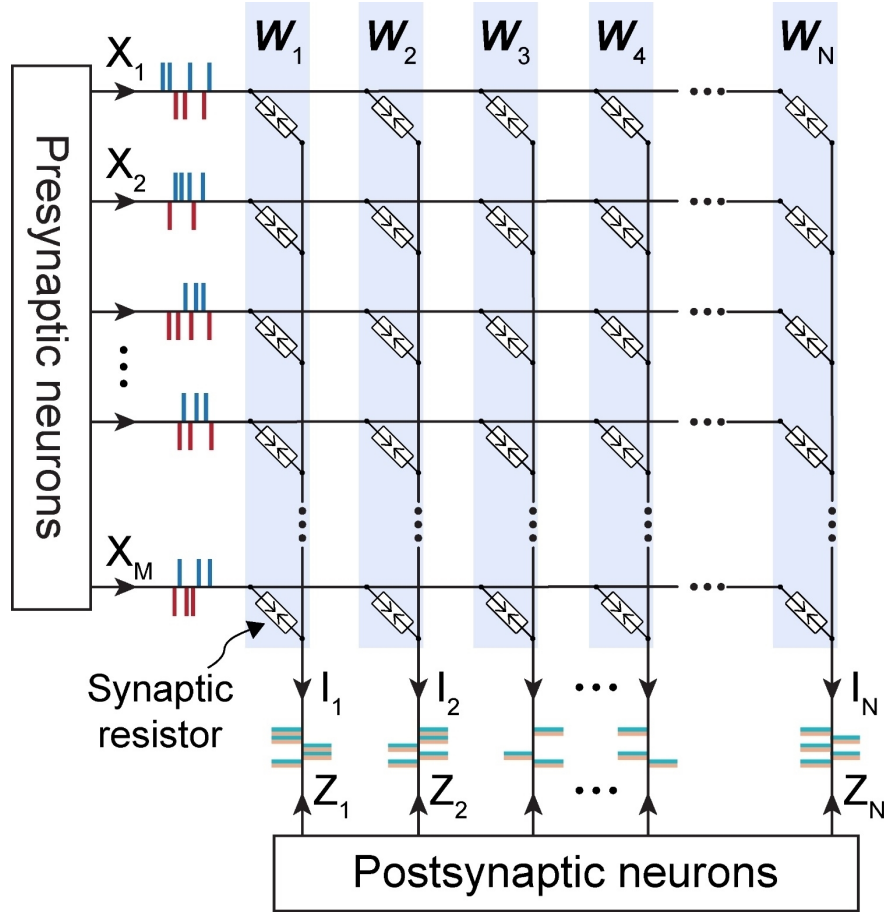


Figure 30. A schematic diagram of associative memory circuit based on the synaptic resistors.

(A) Associative memories (W) is modeled with a crossbar array of $M \times N$ devices to represent N memories with each memory have M components. A crossbar array is connected to M presynaptic neurons and N postsynaptic neurons, where the presynaptic neurons feed the input pulses and postsynaptic neurons integrate and fire the feedback signal for training.

3.2.1. Hopfield Network and Associative Memory

Hopfield network is a neural network that takes incomplete or corrupt patterns as inputs and then uses the update rule to complete or reconstruct the input patterns with the memorized patterns⁵⁴. It is an energy-based network that recalls the memorized states by minimizing the

energy of the input state, updating each component recurrently. It is a biologically plausible neural network, with a simple update rule that is compatible with crossbar circuits⁵⁵. However, the classical Hopfield network suffers from spurious minima states generated from the crosstalk between memories, which significantly lowers its memory capacity. To overcome this limitation, the modern Hopfield network utilizes the higher order polynomial energy function F to increase the memory capacity than its classical counterpart⁵⁶. The energy function of the modern Hopfield network is given below:

$$F(x) = x^n, \text{ with } n \geq 2 \quad (3)$$

When $n = 2$, the network is equivalent to the classical Hopfield network. With a given energy function, the Hopfield energy is defined as Equation 4.

$$E = - \sum_{\mu=1}^K F\left(\sum_{i=1}^N \xi_i^{\mu} X_i\right) \quad (4)$$

In this formula, X_i is the i th element of the input pattern ($\mathbf{X}$) with size N and ξ_i^{μ} is the i th element of the μ th memorized pattern among total K memorized patterns with size N . Each element in the pattern, X_i and ξ_i , is either 1 or -1. To visualize the memory update dynamics of Hopfield networks, four alphabet letters (U, C, L, and A) are memorized to the network, as shown in Figure 31. A total of 31 input patterns have been selected to visualize the energy landscapes and memory capacity for Hopfield networks with different energy functions. The sequence of patterns represents a single pathway on a multi-dimensional hyperspace, which

connects the four memorized patterns. The small arrow in Figure 31A shows the updated direction of the input pattern, where it minimizes the energy of the system along the energy landscape by the update rule Equation 5⁵⁶.

$$X'_i = \text{Sign}[\sum_{\mu=1}^K \{F(\xi_i^\mu + \sum_{j \neq i} \xi_j^\mu X_j) - F(-\xi_i^\mu + \sum_{j \neq i} \xi_j^\mu X_j)\}] \quad (5)$$

The transition from X_i to updated element X'_i always lowers the energy of the input pattern $\mathbf{X}$, thus leading to the local minima on an energy landscape with recurrent updates. For $n = 2$, which is equivalent to the classical Hopfield network, the spurious minima states have occupied the lowest energy along the energy landscape. A 2D energy landscape using multidimensional scaling (Figure 31B) shows that the spurious minima states (red dots) are generated in the middle of memorized states (blue dots) due to their overlapping influence. In contrast, the energy landscape is much sharper in the $n = 30$ case, successfully memorized the patterns without unwanted spurious minima. Moreover, a 2D energy landscape of $n = 30$ implicates that there is still more space for additional memories. Despite the obvious advantages of the modern Hopfield network and a growing interest in its machine learning applications⁵⁷, there was no previous attempt at its hardware implementation due to the nonlinear functions and global summations, which are hardly compatible with the linear circuit elements. In this work, we propose the associative memory circuit (AMC) based on the modern Hopfield network, using a circuit-friendly update rule.

In the memory update rule of the modern Hopfield network given in Equation 5, X_i is the i th element of the input pattern $\mathbf{X}$, where it updates to X'_i with the update rule. F is the energy function, ξ_i^μ is the i th element of the μ th pattern among total K memorized patterns. Each

element in the pattern, X_i and ξ_i , is either -1 or 1. With the quadratic energy function $F(x) = x^2$, the update rule is equivalent to the one of the classical Hopfield network. The update rule for classical Hopfield network can be easily implemented in the hardware with its matrix form using the linear circuit elements, where they perform multiply-and-accumulate (MAC) operations. In contrast, a modern Hopfield network with $n \geq 3$ is relatively complicated to implement in the hardware, due to the nonlinear energy function F . To successfully implement the modern Hopfield network in a hardware, a circuit friendly update rule can be derived from Equation 5 by choosing the special energy function.

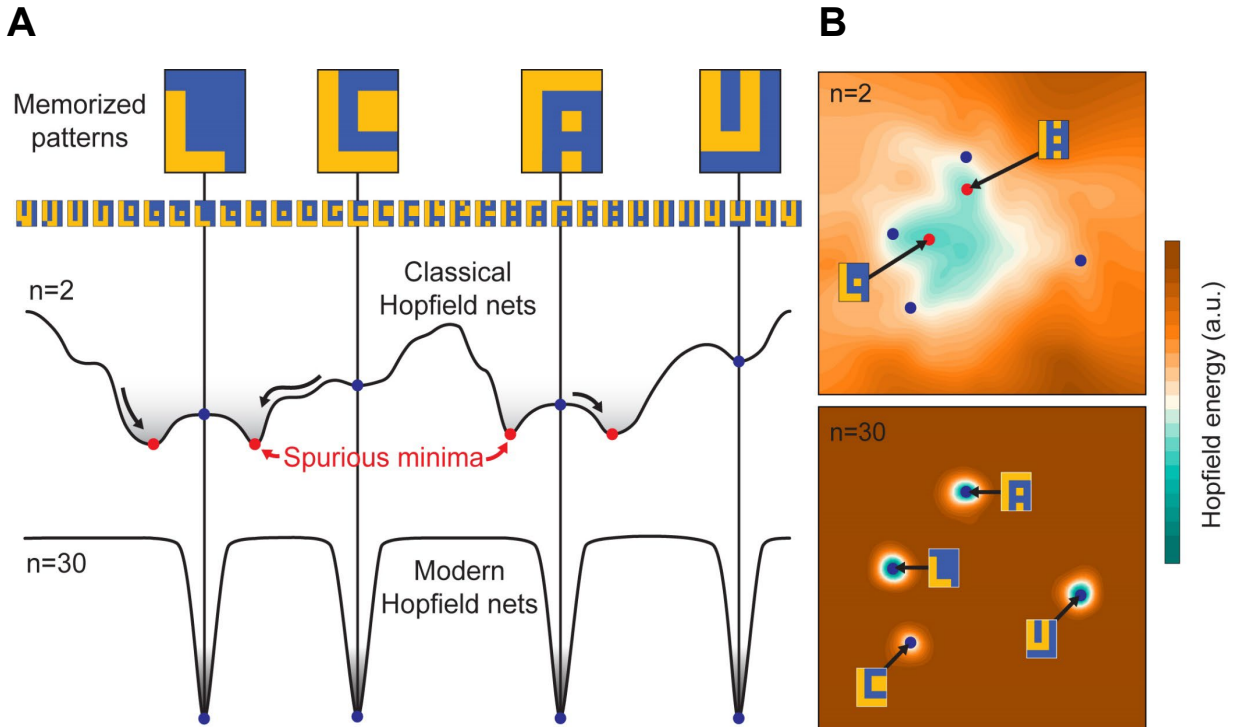


Figure 31. Characteristics of modern Hopfield network with different energy functions. (A) Four different memorized patterns (U, C, L, and A) are memorized to the networks. A total of 31 input

patterns are selected to demonstrate the energy landscape for Hopfield networks. Spurious minima states were denoted as red dots, while blue dots represent the memorized pattern along the energy landscape. (B) A 2D energy landscape using multidimensional scaling (MDS) technique. The blue/red dots represent the location of memorized patterns and spurious minima on an energy landscape for the corresponding energy functions.

Herein, to derive a circuit friendly update rule, the energy function of $F(x) = x^{2n+1}$ has chosen, where n is a sufficiently large integer. With such energy function, the contribution of a single pattern $\mu_{\max}^{[1]}$, which yields the maximum scalar product with the input vector, significantly outweighs the contributions of the remaining patterns (Equation 6).

^[1] The assumption used in the derivation may not be applicable, in case of there are multiple patterns with a same maximum scalar product with the clamped input vector, which are $\mu_{\max 1}, \mu_{\max 2}, \mu_{\max 3}, \dots, \mu_{\max N}$.

However, the same derivation approaches still hold for such cases and give the following update rule:

$$X'_i = \text{Sign}(\xi_i^{\mu_{\max 1}} + \xi_i^{\mu_{\max 2}} + \dots + \xi_i^{\mu_{\max N}})$$

This update rule suggests that, even with an infinitely sharper energy function, the spurious minima can show up when there is an excessive number of memories stored within a memory, e.g. $K \approx 2^N$. Note that this is a special case with binary states, which hardly has an implication to the memory with continuous states with a large number of elements.

$$\begin{aligned}
& \left| F(\xi_i^{\mu_{max}} + \sum_{j \neq i} \xi_j^{\mu_{max}} X_j) - F(-\xi_i^{\mu_{max}} + \sum_{j \neq i} \xi_j^{\mu_{max}} X_j) \right| \\
& \gg \left| \sum_{\mu \neq \mu_{max}}^K [F(\xi_i^\mu + \sum_{j \neq i} \xi_j^\mu X_j) - F(-\xi_i^\mu + \sum_{j \neq i} \xi_j^\mu X_j)] \right| \quad (6) \\
& , \text{ where } \mu_{max} = \operatorname{argmax}(X^T \cdot \mathbf{W})
\end{aligned}$$

Where $\mathbf{W}$ is the $N \times K$ matrix with K memorized patterns. The update rule is then reduced to Equation 7 without the summation across the entire set of patterns.

$$X'_i = \operatorname{Sign}[F(\xi_i^{\mu_{max}} + \sum_{j \neq i} \xi_j^{\mu_{max}} X_j) - F(-\xi_i^{\mu_{max}} + \sum_{j \neq i} \xi_j^{\mu_{max}} X_j)] \quad (7)$$

Since the $F(x)$ is a monotonic function, $\operatorname{Sign}(F(x)-F(y))$ is equivalent to $\operatorname{Sign}(x-y)$, which further simplifies the update rule:

$$X'_i = \operatorname{Sign}[(\xi_i^{\mu_{max}} + \sum_{j \neq i} \xi_j^{\mu_{max}} X_j) - (-\xi_i^{\mu_{max}} + \sum_{j \neq i} \xi_j^{\mu_{max}} X_j)] = \operatorname{Sign}(\xi_i^{\mu_{max}}) \quad (8)$$

The simplified update rule suggests that the input pattern will be recalled to the memorized pattern μ_{max} which has the highest similarity to the input pattern.

Although Equation 8 provides the hardware-friendly update rule, it only applies to the binary memory states. To generalize the update rule for analog hardware circuits, the update rule for binary memory states needs to be revised to the memory with analog states. Again, the intuition for the binary update rule applies the same for the analog memory states. The analog

input pattern shall recall to one of the memories with the highest similarity to the input pattern.

The memory update rule with the analog memory states can be proposed as follows:

$$X'_i = |\xi_i^\mu| \times \text{Sign}[\sum_{\mu=1}^K \{F(\xi_i^\mu + \sum_{j \neq i} \xi_j^\mu X_j) - F(-\xi_i^\mu + \sum_{j \neq i} \xi_j^\mu X_j)\}] \quad (9)$$

Since every derivation step for the binary memory update rule is valid on the analog memory case, the update rule now simplified as below:

$$X'_i = |\xi_i^\mu| \times \text{Sign}(\xi_i^{\mu_{max}}) = \xi_i^{\mu_{max}} \quad (10)$$

Equation 10 offers a simple update rule, which can be implemented in hardware with linear circuit elements. Using this update rule, nonlinear operations such as the XOR gate can be solved using Hopfield networks. By memorizing the truth table in its binary weight matrix $\mathbf{W}$, AMC can also implement various binary operations including the XOR gate (Figure 32).

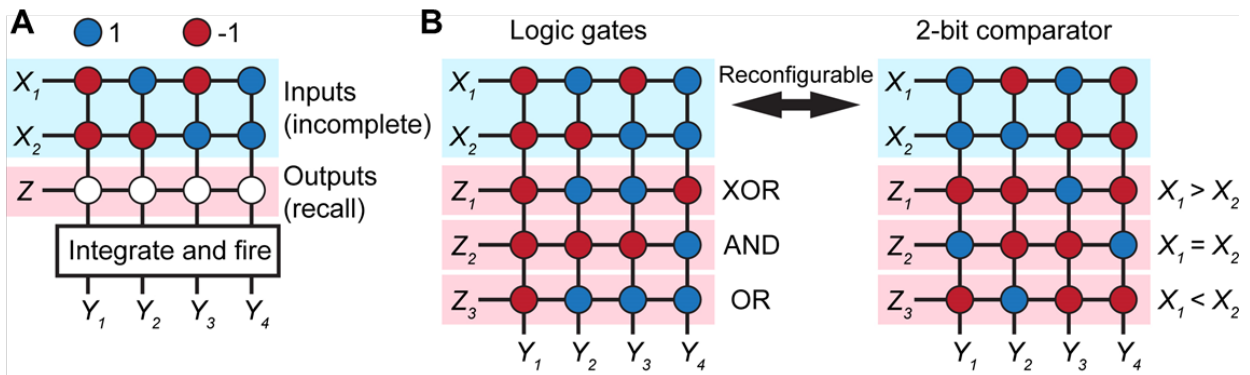


Figure 32. Binary operators based on AMC. (A) Input signals X_1 and X_2 are given to the AMC to select the output Y which fires first, to recall the result Z of a binary operation. (B) Binary operators implemented on AMC with their truth table memorized on the circuit.

This task is equivalent to modeling the binary operator as an associative memory update process, where the inputs X_1 and X_2 are treated as a part of an incomplete pattern, and the result of a binary operation Z is treated as an updated pattern. The inference process is as follows: first, the AMC is given a set of input signal X , either +1 or -1, which induces the current responses $\mathbf{X}'\mathbf{W}$ to each output column $\mathbf{Y}$. Subsequently, the first column Y_f to fire is selected as an output to recall Z . Then, the reading signal +1 is applied to Z , and the output result of a binary operation can be determined by reading the current response from Y_f . For instance, when the input $[X_1, X_2] = [1, -1]$ is given to the AMC in Figure 32B (left), it will induce the current $\mathbf{X}'\mathbf{W}$ on outputs $[Y_1, Y_2, Y_3, Y_4] = [0, 2, -2, 0]$, which makes the output Y_2 to fire. And then, a reading signal is given to Z_1 , then the output Y_2 gives the result of the XOR logic operation, which is 1. With a large flexibility in the configuration, AMC can be programmed and reprogrammed to any binary operator with a truth table compatible with the circuit size, such as a 2-bit comparator shown in Figure 32B (right) reconfigured from the logic gates Figure 32B (left). The way AMC solves the logic gates resembles the way human brains solve the binary logic gates; most of them just memorize the truth table and recall the result. However, binary logic is not the best use for AMC since the density of information in the memory is at its lowest.

3.2.2. Self-supervised Learning of Associative Memory Circuit

Self-supervised learning of an associative memory circuit (AMC) is designed to train the underlying structure or patterns in the data by shaping the energy landscape. The loss function (L) for the circuit is formulated as follows:

$$L = -\mathbf{X}^T \mathbf{W} \mathbf{Z}_\mu + \frac{1}{2} \beta (\mathbf{V} \mathbf{V}^T), \quad (11)$$

$$\text{where } \mathbf{V} = \vec{\mathbf{1}} \cdot \mathbf{W},, \text{ and } \mu = \operatorname{argmax}(\mathbf{X}^T \cdot \mathbf{W})$$

$\mathbf{W}$ is $N \times K$ weight matrix for K different memories with a size of N . $\mathbf{X}$ is an input training data vector with a size of N . $\vec{\mathbf{1}}$ is a row vector with all ones to calculate the columnar sum $\mathbf{V}$ of $\mathbf{W}$. $\mathbf{Z}_\mu$ is a selection vector with all 0 except for a single 1 for the μ th element, and β is a hyperparameter related with global unlearning. The first argument in Equation 11 is a quantitative measure of how well the memory models the given training data by evaluating the scalar product between the training data $\mathbf{X}$ and a memory $\mathbf{W} \mathbf{Z}_\mu$. Using this argument alone would lead the weight matrix to diverge without a limit, which could be regulated by introducing the second term. The second argument in Equation 11 is penalizing the memory columns with large conductance, thus prohibiting the weight matrix from divergence and keeping the weight values within the reasonable conductance range of the synaptic resistors. The hyperparameter β is a term for conductance range adjustment, which is set to be small when the device has a wide conductance range but relatively high when the device experiences a limited range of conductance states.

Learning the associative memory is achieved by shaping the Hopfield energy landscape to have local/global stable states at the positions where they could represent the underlying

features of the training datasets. It utilizes the gradient of the loss function to update the weight matrix, thus shaping the Hopfield energy landscape (Equation 12, 13). To represent the signed memory value using synaptic resistors, AMC utilizes two sets of complementary memory matrices $\mathbf{W}^+$ and $\mathbf{W}^-$ for the memory matrix ($\mathbf{W} = \mathbf{W}^+ - \mathbf{W}^-$), as well as $\mathbf{V}^+$ and $\mathbf{V}^-$.

$$\Delta \mathbf{W}^+ = -\tau \frac{\partial L}{\partial \mathbf{W}^+} = \tau \mathbf{X} \otimes \mathbf{Z}_\mu - \tau \beta (\vec{\mathbf{1}} \otimes \mathbf{V}^+) \quad (12)$$

$$\Delta \mathbf{W}^- = -\tau \frac{\partial L}{\partial \mathbf{W}^-} = -\tau \mathbf{X} \otimes \mathbf{Z}_\mu - \tau \beta (\vec{\mathbf{1}} \otimes \mathbf{V}^-) \quad (13)$$

τ is a learning rate which is closely related to the synaptic resistor's tuning behavior. In Equation 2,3, first term $\pm \tau \mathbf{X} \otimes \mathbf{Z}_\mu$ is for learning the memory by moving the position of a current memory state $\mathbf{WZ}_\mu$ slightly to a training data $\mathbf{X}$. The second term is for unlearning the memory. It regulates the device conductance to a reasonable range and reduces the conductance of unused weights in each memory. Both terms are implemented in the circuit using the inherent learning rule of a synaptic resistor, where it learns only when there are the same-polarity voltage pulses from pre- and post-synaptic neuron circuits, but it does not learn otherwise. Note that this energy formulation with a derived learning rule is for self-supervised learning. The learning rule for supervised learning of AMC is different as presented in the hand-written digit classification demonstration, where it appends the learning rule to Hebbian/Anti-Hebbian learning, which offers faster convergence and enabling AMC to extract a larger number of features by competing for the memories of different digit classes (Figure 33C, 33E).

During the learning and unlearning phase, the conductance of the device is updated based on the interaction between the input signal from the pre-synaptic neuron and a feedback signal from the post-synaptic neuron (Figure 33C, 33D, 33E). The feedback signal is given to the column selected for learning based on their response during the inference phase. Based on the device's learning rule, conductance changes only when it experiences both the input signal and feedback signal, which enables concurrent inference and learning. This scheme enables three biologically plausible features. First, the network learns through the locally available information given from pre- and post-synaptic neurons. The input signal is the same as the inference phase, and the feedback signal is a bias signal over a fixed cycle. Second, the learning process for the whole network is executed in parallel with inference. In particular, Hebbian and anti-Hebbian learning can be done parallel to the inference except for the unlearning, where it needs a negative bias voltage from the presynaptic neurons and the negative postsynaptic neuron signals V_i .

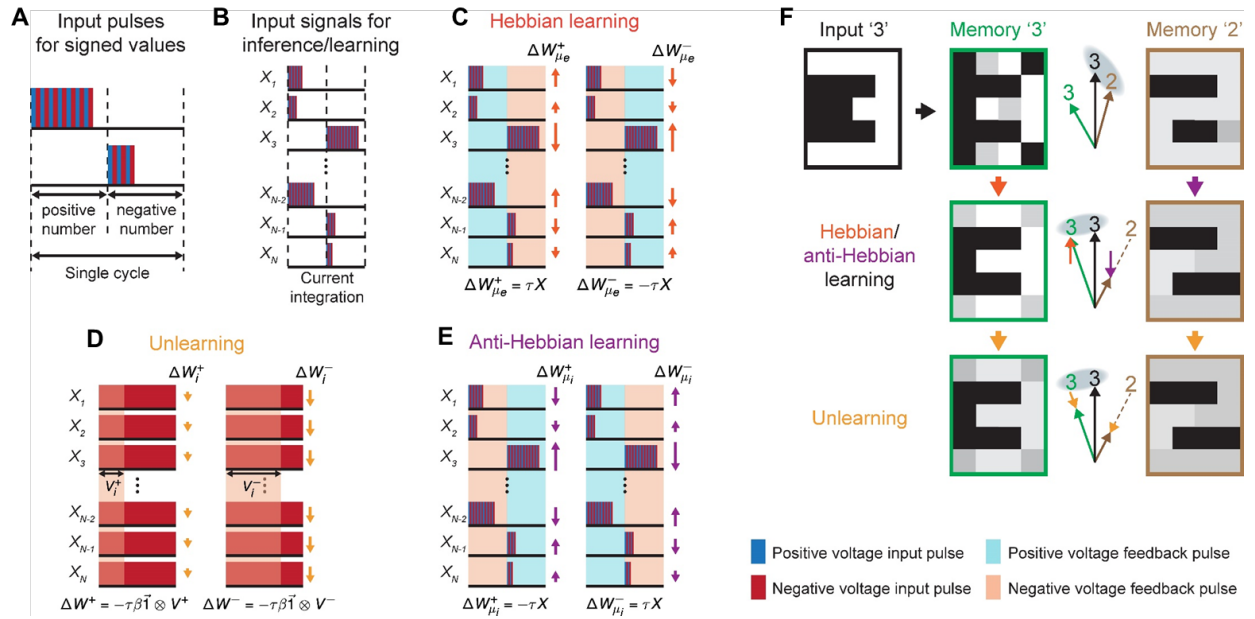


Figure 33. Hebbian learning, anti-Hebbian learning, and the global unlearning scheme of AMC.

(A) Input pulses for signed values with alternating voltage pulses. (B) Schematics for the input

signals for concurrent inference/learning. For the columns where the post-synaptic neurons fire, (C) Hebbian learning is executed by applying feedback pulses for $W_{\mu_e}^+$ and $W_{\mu_e}^-$ respectively, and (E) anti-Hebbian learning could be implemented with the reversed order of feedback pulses with respect to Hebbian learning. (D) Global unlearning is to reduce the overall conductance of the memory when their current response (V) exceeds the threshold, with the feedback pulse width relative to the columnar response V_i .

Learning of AMC is based on stochastic gradient descent without external memories and computations. The batch learning process requires external memory and circuitry to store and process the gradient, increasing the size and complexity by a few folds.

3.3.3. MNIST Hand-written Digits Classification

The MNIST digit classification is a pattern recognition task for evaluating machine learning models. It consists of 60,000 training images and 10,000 test images of handwritten digits (0-9). The learning algorithm of AMC was simulated using synaptic resistor device models and evaluated the accuracy with test datasets. Every device is modeled as having a conductance range from 0.1 nS to 60 nS, and the training images are zero-centered and normalized as a pre-processing.

The supervised learning of AMC with 900 memories for the image size of 784 (network size: 784×900) begins with the network initialized to random noise (Figure 34). Over time, the system evolved to learn and classify handwritten digits with increasing accuracy. At the start, the synaptic weights are initialized randomly, resulting in chaotic and unstructured representations within the network. This stage serves as a baseline, where no meaningful patterns or digit representations are evident. After three epochs, the network begins to form abstract

representations of the digits. These abstractions are not yet fully formed prototypes, but they inform that the network is starting to discern the basic features and patterns of the digits. The learning process at this stage involves significant adjustments to the synaptic weights, guided by the learning rules applied. By the fifth to sixth epoch, the network has evolved to create prototypes for each of the digits. These prototypes represent the different variations of handwritten digits and form the basis for high-accuracy classification. At this stage, the synaptic weights of associative memories have stabilized, and the network is fine-tuning its representations afterward.

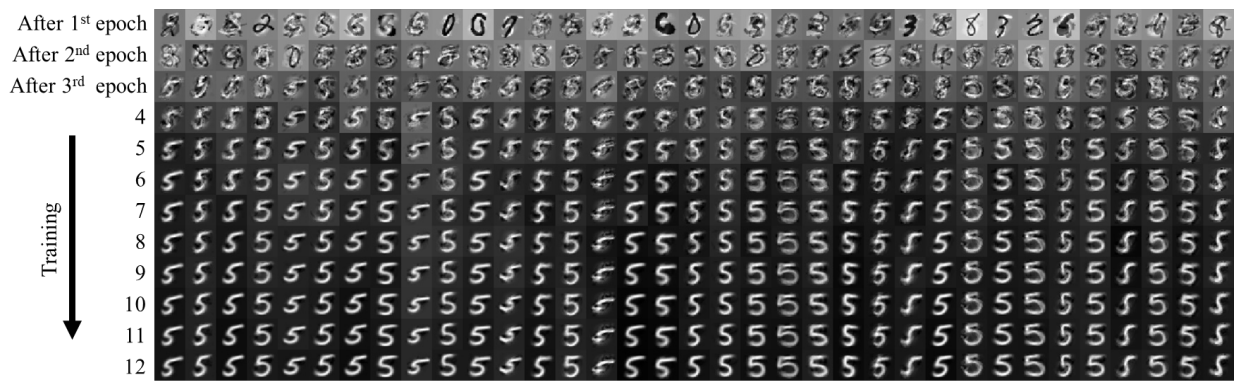


Figure 34. Evolution of associative memories assigned for digit ‘5’ after each training epoch. It begins with the network initialized to random noise and evolves under Hebbian learning, anti-Hebbian learning, and unlearning rules. After three epochs, the memories begin to form abstract representations of the digits, and they have evolved to create prototypes for each of the digits after the fifth ~ sixth epoch.

The classification accuracy and learning efficiency are significantly influenced by the learning rules applied. In this study, we employed three primary learning rules: Hebbian learning, anti-Hebbian learning, and unlearning. Each of these rules contributes differently to the learning process. Firstly, Hebbian learning strengthens synaptic connections when the input

pattern from the presynaptic neurons induces the current responses and fires one of the postsynaptic neurons that matches the label. Anti-Hebbian learning, in contrast, weakens the synaptic connections when the input pattern from the presynaptic neurons fires one of the postsynaptic neurons that does not match the label. This mechanism is crucial for memories to accurately represent the features of the assigned digit by lowering the similarity from the training images of different digits. Lastly, the unlearning process reduces the strength of synaptic connections of the memories with the large current response, maintaining the conductance of the synaptic resistor to be within the reasonable range and enabling better feature extraction.

In Figure 35, 36, and 37, the effect of using different combinations of these learning rules is demonstrated to evaluate their impact on learning behavior and classification accuracy. If AMC is trained only based on Hebbian learning, the network learns associations well but suffers from overgeneralization, leading to a moderate train accuracy of 69%, with the number of memories used for classification remaining only 30 out of 900 memories. Secondly, Hebbian learning with unlearning showed regularized behavior of memories, which improved the memories to have higher resolution representations relative to the digitized shape of the “only Hebbian” case, achieving train accuracy of 76% with 32 used memories.

With anti-Hebbian learning, the AMC could extract a much larger number of different features. Anti-Hebbian learning lets the associative memories assigned for different digits compete with each other, improving the network's inclination to overgeneralize and capture different features and evolve them into unique prototypes and scores the train accuracy of 95% with 539 memories being actively used. Combining all three learning mechanisms, AMC achieved the highest train accuracy of 99%, test accuracy of 96%, and generated 780 active memories by balancing generalization, overfitting, and regularization.

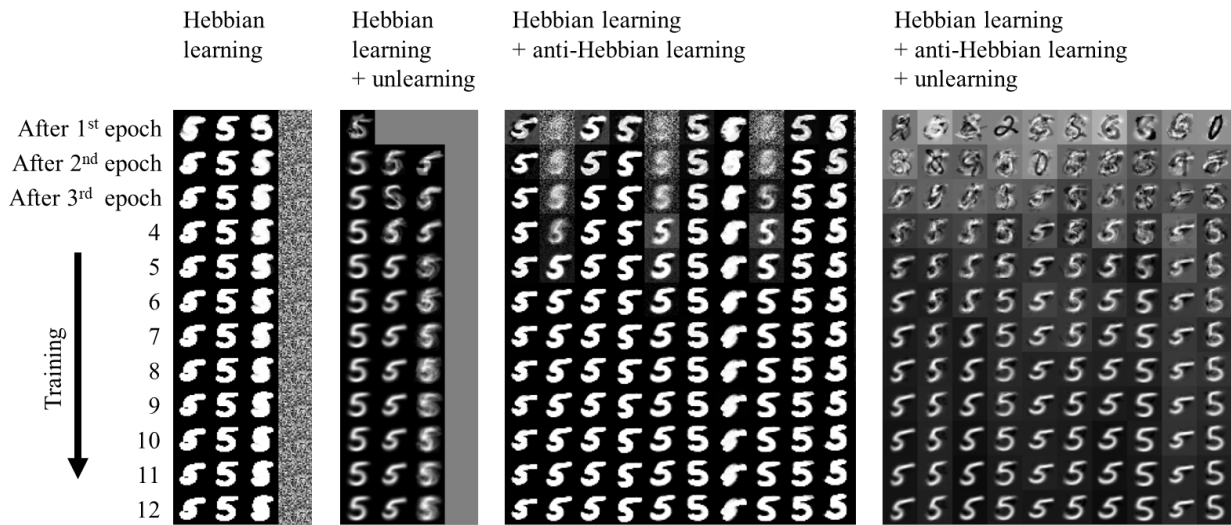


Figure 35. Training AMC with different combinations of learning mechanisms. For the cases “Hebbian + anti-Hebbian” and “Hebbian + anti-Hebbian + unlearning”, only 10 memories are displayed.

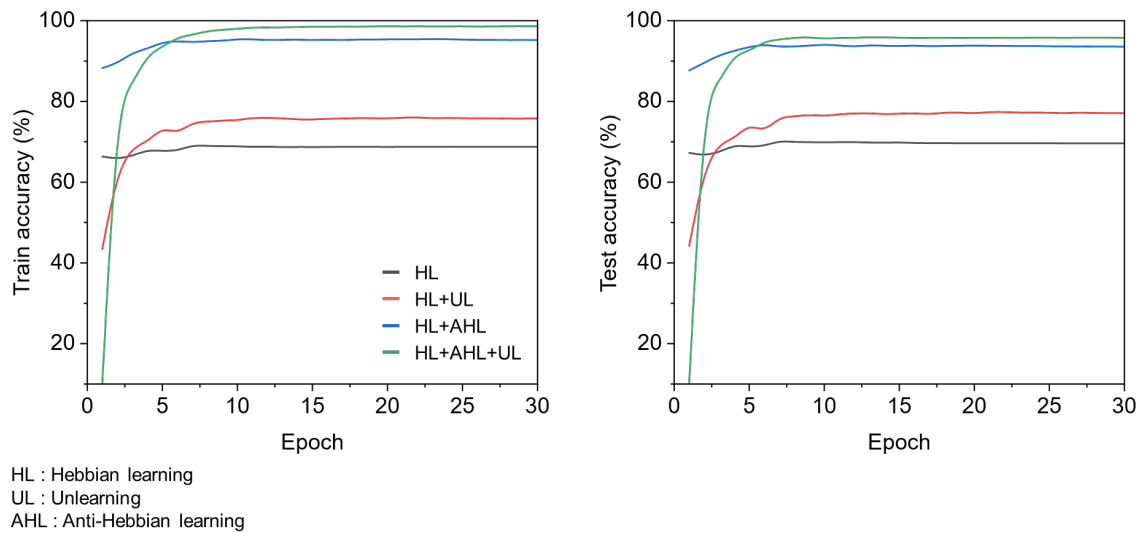


Figure 36. Train and test accuracy for the network under different learning algorithms. Train/test accuracy at epoch 30, HL: 69%/69%, HL+UL: 76%/77%, HL+AHL: 95%/94%, HL+AHL+UL: 99%/96%.

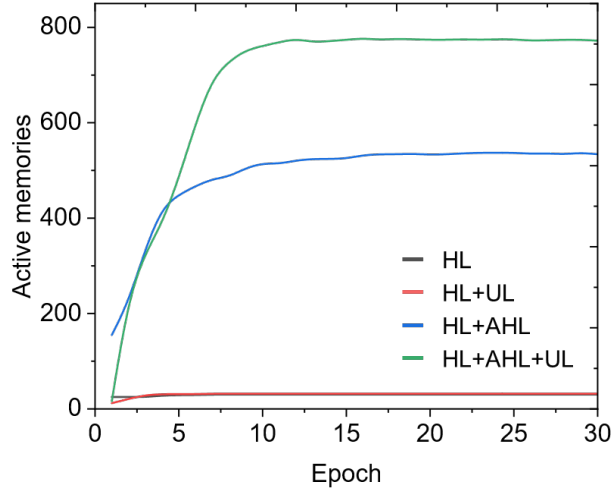
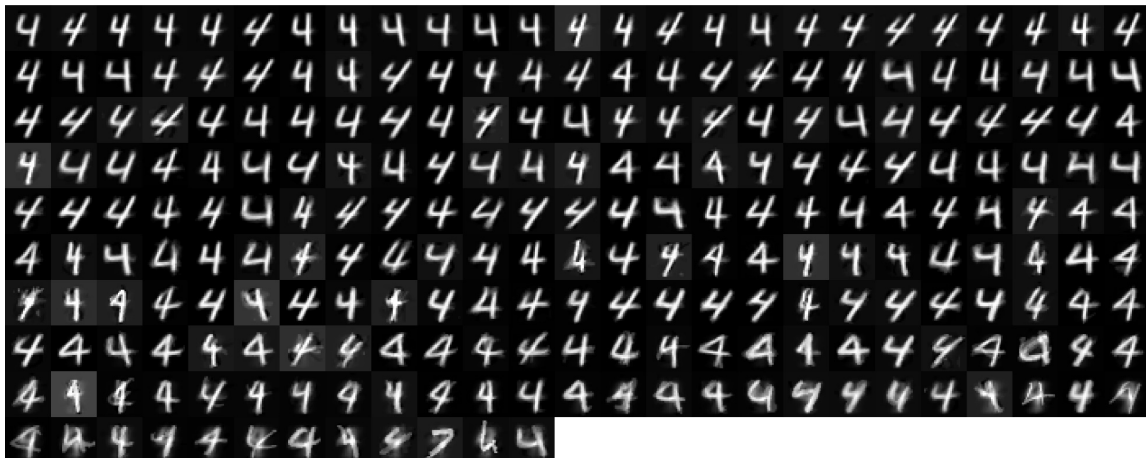
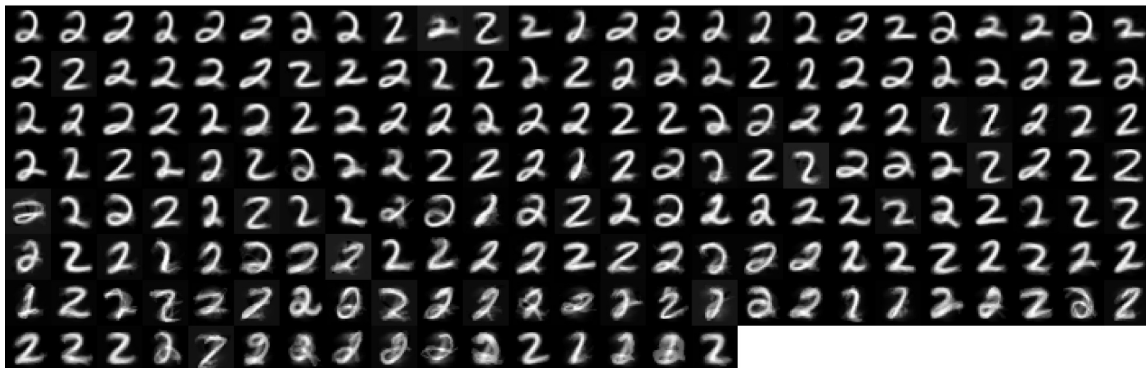
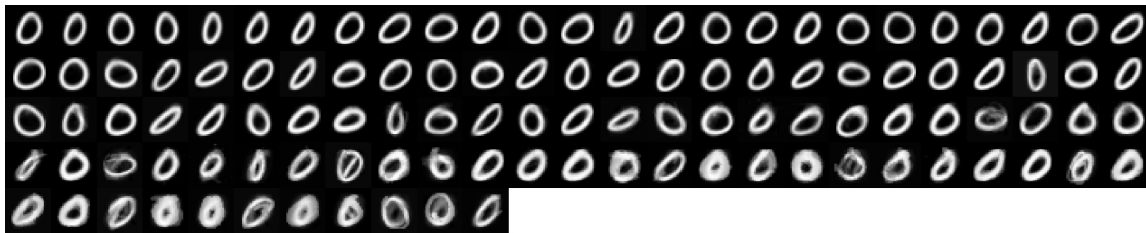


Figure 37. Total number of memories used for classification among 900 associative memories. HL: 30, HL+UL: 32, HL+AHL: 539, and HL+AHL+UL: 780. Using a combined learning algorithm, the AMC extracted 780 prototypical memories (images) for all the digits.

The results of our MNIST classification application demonstrate the potential of AMC with biological learning rules in neuromorphic systems. Starting from random noise, the network evolves to form abstract and then prototypical representations of digits, achieving high classification accuracy. The combination of Hebbian learning, anti-Hebbian learning, and unlearning proves to be the most effective, resulting in a hardware-friendly, high-accuracy classification system. This approach paves the way for efficient and scalable implementations of neuromorphic systems for complex pattern recognition tasks.



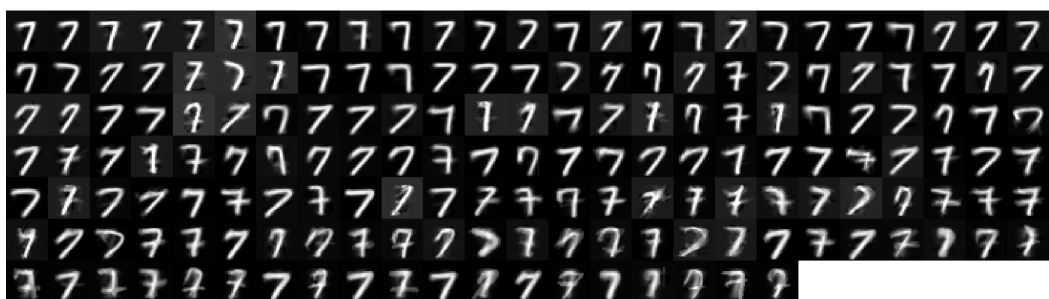
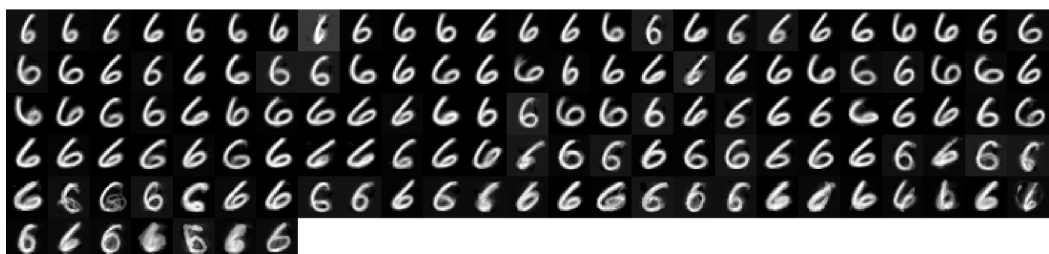


Figure 38. Extracted prototypical images of different digits using AMC network with 2700 memories (network size 784×2700). Out of 2700 memories, only the active memories that participate in classifying the digits are displayed.

4. Conclusions

This dissertation presents a comprehensive study on HfZrO-based synaptic resistor circuits for neuromorphic computing. A new type of synaptic resistor based on HfZrO ferroelectric material was successfully fabricated, demonstrating stable multi-level, non-volatile memory characteristics. The device structure, which integrates a $\text{WO}_{2.8}/\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ heterojunction with a silicon channel, allows for precise conductance modulation. Detailed material and device characterizations using advanced techniques such as STEM, EDX, and EELS confirmed the presence of ferroelectric and dielectric phases in the HfZrO layer, crucial for the device's performance. With its learning rule, the HfZrO-based synaptic resistor circuit uniquely enables concurrent execution of inference and learning, mimicking the real-time adaptability of biological synapses. Theoretical analyses and experimental results demonstrated that this concurrent functionality is critical for dynamic optimization of system states, enhancing the circuit's ability to adapt to changing environments and tasks efficiently.

The synaptic resistor circuit was applied to various neuromorphic tasks, showcasing its potential in real-world scenarios. Notably, the circuit was used to navigate a drone through a simulated environment with dynamic obstacles and wind conditions. The results showed that the synaptic resistor circuit outperformed both human operators and traditional computer-based artificial neural networks (ANNs) in terms of learning time, adaptability, and power consumption.

Lastly, an associative memory circuit (AMC) based on the modern Hopfield network was proposed and implemented using the synaptic resistor crossbar array. The AMC, trained with a combination of Hebbian learning, anti-Hebbian learning, and unlearning, demonstrated the capability to classify the MNIST handwritten digit dataset with high accuracy. The network evolved from random noise to abstract representations and finally to prototypical digit representations, achieving a test accuracy of 96%.

In conclusion, the HfZrO-based synaptic resistor circuit represents a significant step toward realizing brain-inspired computing systems. Its ability to perform concurrent real-time learning and inference with high energy efficiency and scalability makes it a promising candidate for next-generation artificial intelligence applications. This dissertation lays the groundwork for further advancements in neuromorphic computing, potentially leading to more intelligent and adaptable AI systems.

REFERENCES

1. Turing, A. M. Computing machinery and intelligence. *Mind* **LIX**, 433–460 (1950).
2. Mei, Q., Xie, Y., Yuan, W. & Jackson, M. O. A Turing test of whether AI chatbots are behaviorally similar to humans. *Proc. Natl. Acad. Sci. U. S. A.* **121**, 1–8 (2024).
3. Semiconductor Research Corporation. *SRC decadal plan 2020*. (2020).
4. Camus, A. *The Plague*. (1948).
5. Mill, J. S. On Liberty. (1885).
6. Erich Fromm. To Have or to Be? (1976).
7. IEEE. *International Roadmap for Devices and Systems: More Moore. International Roadmap for Devices and Systems* (2019).
8. Danesh, C. D. *et al.* Synaptic Resistors for Concurrent Inference and Learning with High Energy Efficiency. *Adv. Mater.* **1808032**, 1808032 (2019).
9. Nathan, D. *et al.* Si-based self-programming neuromorphic integrated circuits for intelligent morphing wings. *J. Compos. Mater.* **56**, 4561–4575 (2022).
10. Gao, D. *et al.* Synaptic Resistor Circuits Based on Al Oxide and Ti Silicide for Concurrent Learning and Signal Processing in Artificial Intelligence Systems. *Adv. Mater.* **35**, 1–11 (2023).
11. Von Neumann, J. *The computer and the brain*. (Yale university press, 1958).
12. Minsky, Marvin L., and S. A. P. *Perceptrons: expanded edition*. (1988).
13. Hertz, J., Krogh, A., Palmer, R. G., & Horner, H. *Introduction to the theory of neural computation*. (1991).
14. Mead, C. *Analog VLSI implementation of neural systems*. (1989).
15. Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory*. (1949).

16. W. S. McCulloch & W. Pitts. A logical calculus of the ideas immanent in nervous activity. *BULLETIN OF MATHEMATICAL BIOPHYSICS* **5**, 115–133 (1943).
17. Dan, Y. & Poo, M. M. Spike timing-dependent plasticity of neural circuits. *Neuron* **44**, 23–30 (2004).
18. Z. Chen, S. Haykin, J. J. Eggermont, S. B. *Correlative Learning: A Basis for Brain and Adaptive Systems*. (2007).
19. Frémaux, N. & Gerstner, W. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Front. Neural Circuits* **9**, (2015).
20. La Cerra, P. & Bingham, R. The adaptive nature of the human neurocognitive architecture: An alternative model. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 11290–11294 (1998).
21. OpenAI *et al.* GPT-4 Technical Report. **4**, 1–100 (2023).
22. Jaeger, H., Noheda, B. & van der Wiel, W. G. Toward a formal theory for computing machines made out of whatever physics offers. *Nat. Commun.* **14**, (2023).
23. Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. *Nature* **453**, 80–83 (2008).
24. Yang, J. J. *et al.* Memristive switching mechanism for metal/oxide/metal nanodevices. *Nat. Nanotechnol.* **3**, 429–433 (2008).
25. Kund, M. *et al.* Conductive bridging RAM (CBRAM): An emerging non-volatile memory technology scalable to sub 20nm. *Tech. Dig. - Int. Electron Devices Meet. IEDM* **2005**, 754–757 (2005).
26. Lankhorst, M. H. R., Ketelaars, B. W. S. M. M. & Wolters, R. A. M. Low-cost and nanoscale non-volatile memory concept for future silicon chips. *Nat. Mater.* **4**, 347–352

- (2005).
27. Esaki, A. L., R. B. Laibowitz, and P. J. S. Polar switch. *IBM Tech. Discl. Bull* **13**, 2161 (1971).
 28. Asapu, S. *et al.* Large remnant polarization and great reliability characteristics in W / HZO / W ferroelectric capacitors. **2011**, 1–8 (2022).
 29. Engel, B. N. *et al.* A 4-Mb toggle MRAM based on a novel bit and switching method. *IEEE Trans. Magn.* **41**, 132–136 (2005).
 30. Yi, S. I., Kumar, S. & Williams, R. S. Improved Hopfield Network Optimization Using Manufacturable Three-Terminal Electronic Synapses. *IEEE Trans. Circuits Syst. I Regul. Pap.* **68**, 4970–4978 (2021).
 31. Dünkel, S. *et al.* A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond. *Tech. Dig. - Int. Electron Devices Meet. IEDM* 19.7.1-19.7.4 (2018). doi:10.1109/IEDM.2017.8268425
 32. Rivnay, J. *et al.* Organic electrochemical transistors. *Nat. Rev. Mater.* **3**, (2018).
 33. Chua, L. Memristor-The missing circuit element. *IEEE Trans. circuit theory* **18**, 507–519 (1971).
 34. Prezioso, M. *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61–64 (2015).
 35. Hu, M., Chen, Y., Yang, J. J., Wang, Y. & Li, H. A Compact Memristor-Based Dynamic Synapse for Spiking Neural Networks. *IEEE Trans. Comput. Des. Integr. Circuits Syst.* **36**, 1353–1366 (2017).
 36. Yao, P. *et al.* Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).

37. Song, W. *et al.* Programming memristor arrays with arbitrarily high precision for analog computing. *Science (80-.)*. **383**, 903–910 (2024).
38. Du, S. *et al.* Multidirectional Associative Memory Neural Network Circuit Based on Memristor. *IEEE Trans. Biomed. Circuits Syst.* **17**, 433–445 (2023).
39. Ambrogio, S. *et al.* An analog-AI chip for energy-efficient speech recognition and transcription. *Nature* **620**, 768–775 (2023).
40. Fuller, E. J. *et al.* Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing. *Science (80-.)*. **364**, 570–574 (2019).
41. Danial, L. *et al.* Two-terminal floating-gate transistors with a low-power memristive operation mode for analogue neuromorphic computing. *Nat. Electron.* **2**, (2019).
42. Chen, Y. Brain-Inspired Synaptic Resistor Circuits for Self-Programming Intelligent Systems. *Adv. Intell. Syst.* **3**, 1–12 (2021).
43. Müller, J. *et al.* Nanosecond polarization switching and long retention in a novel MFIS-FET based on ferroelectric HfO₂. *IEEE Electron Device Lett.* **33**, 185–187 (2012).
44. Mulaosmanovic, H. *et al.* Evidence of single domain switching in hafnium oxide based FeFETs: Enabler for multi-level FeFET memory cells. *Tech. Dig. - Int. Electron Devices Meet. IEDM 2016-February*, 26.8.1-26.8.3 (2015).
45. Mulaosmanovic, H. *et al.* Novel ferroelectric FET based synapse for neuromorphic systems. *Dig. Tech. Pap. - Symp. VLSI Technol.* T176–T177 (2017).
doi:10.23919/VLSIT.2017.7998165
46. Oh, S. *et al.* HfZrO_x-based ferroelectric synapse device with 32 levels of conductance states for neuromorphic applications. *IEEE Electron Device Lett.* **38**, 732–735 (2017).
47. Yang, K. *et al.* Wake-up-mitigated giant ferroelectricity in Hf_{0.5}Zr_{0.5}O₂ thin films

- through oxygen-providing, surface-oxidized W electrode. *Mater. Sci. Semicond. Process.* **164**, 107565 (2023).
48. Gong, N. & Ma, T. P. Why Is FE-HfO₂ More Suitable Than PZT or SBT for Scaled Nonvolatile 1-T Memory Cell? A Retention Perspective. *IEEE Electron Device Lett.* **37**, 1123–1126 (2016).
 49. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117**, 500–544 (1952).
 50. C. Li et al. Analogue signal and image processing with large memristor crossbars. *Nat. Electron.* **1**, 52–59 (2018).
 51. Hu, M. *et al.* Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine. *Adv. Mater.* **30**, 1–10 (2018).
 52. Rao, M. *et al.* Thousands of conductance levels in memristors integrated on CMOS. *Nature* **615**, 823–829 (2023).
 53. Zhang, W. *et al.* Edge learning using a fully integrated neuro-inspired memristor chip. *Science (80-.).* **381**, 1205–1211 (2023).
 54. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 2554–2558 (1982).
 55. Yi, S. in, Kendall, J. D., Williams, R. S. & Kumar, S. Activity-difference training of deep neural networks using memristor crossbars. *Nat. Electron.* **6**, 45–51 (2023).
 56. Krotov, D. & Hopfield, J. J. Dense associative memory for pattern recognition. *Adv. Neural Inf. Process. Syst.* 1180–1188 (2016).
 57. Hoover, B. *et al.* Energy Transformer. *Adv. Neural Inf. Process. Syst.* **36**, (2023).