UNIVERSITY OF CALIFORNIA

Los Angeles

# Self-Programming Neuromorphic Integrated Circuit for Intelligent Systems

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Mechanical Engineering

by

Christopher Michael Shaffer

2020

ABSTRACT OF THE DISSERTATION

Self-Programming Neuromorphic Integrated Circuit for Intelligent Systems

by

Christopher Michael Shaffer

Doctor of Philosophy in Mechanical Engineering

University of California, Los Angeles, 2020

Professor Yong Chen, Chair

Artificial neural networks (ANN) have demonstrated performance beyond human capability in challenging games like Go and chess. However, they are still limited by the von Neumann bottleneck, which requires massive overhead for data transmissions between logic and memory units. The end of Moore's Law calls for novel approaches to meet the accelerating computational demands of big data and machine learning. Neuromorphic circuits are promising candidates, inspired by the speed, parallelism, and efficiency of the human brain. A new synaptic device, the synstor, uses Schottky barriers at its I/O terminals and charge trap memory to combine the synaptic functions of convolutional signal processing, Hebbian learning, and nonvolatile analog memory. Unlike circuits based on other synaptic devices, such as memristors and phase change memory, the synstor circuit can spontaneously program its synaptic weights (conductances) via Hebb’s rule, without external computational circuits or complex unit cells. A

20 × 20 crossbar array of synstors is fabricated and connected to artificial neurons to form a self-programming neuromorphic circuit. By applying equal amplitude voltage pulses to synstor input and output electrodes during inference and learning, both processes can run concurrently in a synstor circuit. A 4 × 2 synstor crossbar with 2 "neurons" performs speech recognition with energy efficiency of ~$10^{17}$ FLOPS/W, surpassing existing computing circuits. Additionally, a 2 × 2 synstor circuit programs itself in real-time to stabilize a morphing wing by modulating its camber in a dynamic wind speed environment. The synstor circuit demonstrates performance superior to human participants and a computer-based controller in this task after repeated trials. If scaled up, synstor circuits have the potential to bypass the von Neumann bottleneck of transistor computing circuits, leading to a new computing platform with real-time self-programming and intelligence in complex dynamic environments.

The dissertation of Christopher Michael Shaffer is approved.

Lei He

Yongjie Hu

Veronica Santos

Yong Chen, Committee Chair

University of California, Los Angeles

2020

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support, guidance, and encouragement of my advisors, colleagues, friends, and family.

I would like to thank my advisor, Dr. Yong Chen for his perpetual support and guidance throughout my research. His dedication was vital in the direction and advancement of the project. His advice helped me to grow as a researcher and an individual.

To my committee members, Professor Lei He, Professor Yongjie Hu, and Professor Veronica Santos for their guidance and insights toward this completion.

To my present group members, Mr. Dhruva Nathan, Mr. Rahul Shenoy, Mr. Zixuan Rong, Mr. Dawei Gao, Mr. Jungmin Lee, and Mr. Atharva S. Deo for their ongoing support and collaboration in my research project.

To my past group members Dr. Cameron Danesh, Dr. Andrew Tudor, Mr. Macan Tadayon, Dr. Alex M. Shen, Dr. Kyunghyun Kim, Ms. Yvette Lin, Mr. Dongwon Lee, for their effort prior to my involvement and for their advice in my early career.

Chapters 2-4 include results from the published article "Danesh, et. al., Advanced Materials, vol. 31, 1808032 (2019)"

I acknowledge support given by the Air Force Office of Scientific Research (AFOSR) under the program, "Intelligent Neuromorphic Network" (contract number: FA9550-15- 1-0056) and "Avian-Inspired Multifunctional Morphing Vehicles" (contract number: FA9550-16- 1-0087).

# VITA

| | |
|---|---|
| 2011 | B.S. Mechanical Engineering<br>University of California, Berkeley |
| 2011-2013 | Integration Engineer<br>KLA-Tencor |
| 2015 | M.S. Mechanical Engineering<br>University of California, Los Angeles (UCLA) |
| 2015-2020 | Ph.D. Candidate and Graduate Student Researcher<br>Mechanical and Aerospace Engineering Department<br>University of California, Los Angeles (UCLA) |

# PUBLICATIONS AND POSTERS

**C. M. Shaffer**, A. Tudor, C. D. Danesh, and Y. Chen, “Self-Programming Neuromorphic Integrated Circuit Based on Carbon Nanotube Composite Synapstors” Gordon Research Conference, Multifunctional Materials & Structures Poster Session (2016)

C. D. Danesh*, **C. M. Shaffer***, D. Nathan, R. Shenoy, A. Tudor, M. Tadayon, Y. Lin, and Y. Chen, “Synaptic resistors for concurrent inference and learning with high energy efficiency” Advanced Materials, vol. 31, 1808032 (2019) **(*equally contributed authors)**

**C. M. Shaffer**, A. Deo, A. Tudor, R. Shenoy, C. D. Danesh, D. Nathan, L. L. Gamble, D. J. Inman, and Y. Chen, “Self-programming synaptic resistor circuit for intelligent systems” In Progress (2020)

R. Shenoy, A. Tudor, D. Nathan, **C. M. Shaffer**, C. D. Danesh, A. Deo, and Y. Chen, “Self-programming circuit with real-timing learning functions” In Progress (2020)

D. Nathan, D. Gao, J. Lee, Z. Rong, R. Shenoy, **C. M. Shaffer**, and Y. Chen, "A silicon-based synaptic resistor for scalable analog neuromorphic circuits" In Progress (2020)

# 1. Introduction

## 1.1. The von Neumann Bottleneck and the End of Moore's Law

When Turing established his universal computing model, his overarching ambition was to emulate the human brain.[1] Following Moore's law, the miniaturization of the transistors has exponentially improved the performance and energy efficiencies of computers[2,3] leading to an information revolution and artificial intelligent systems that can simulate learning functions of the human brain.[4] Based on the Turing model, digital computers execute algorithms in serial mode by physically separated logic and memory transistors (Figure 1b), and the computing energy is predominantly consumed by data memory and signal transmissions between memory and logic units,[5-8] referred to as the "von Neumann bottleneck."[9] Transistor-based circuits with parallel computing architectures and distributed memories, such as graphics processing units (GPUs) from Nvidia,[10] tensor processing units (TPUs) from Google,[4,11] field-programmable gate arrays (FPGAs) from Intel,[12] and the TrueNorth neuromorphic circuit from IBM[13] have been developed to improve their energy efficiencies (Figure 1a) to the range of $10^{10} - 10^{11}\ FLOPS/W$ (floating point operations per second per watt) by increasing parallelism and reducing global data transmission. However, their energy efficiencies are fundamentally limited by the energy consumptions on memory ($\sim 10^{-15}\ J/bit$) and signal transitions ($\sim 10^{-11}\ J/bit$) in digital computing circuits.[6,7] When transistors approach the limitations of their minimal sizes near the end of Moore's law, the energy efficiencies of transistor-based computing circuits are asymptotically saturated[5-7,14,15] (**Figure 1**). Meanwhile, the information industry generates "big data" with exponentially increasing volumes, and leads to exponentially increasing power requirements for computations.[5,7,15] This trajectory is unsustainable as it would exceed the entire global power production in one or two decades[16] (**Figure 1**). It is imperative to develop a new

platform to facilitate inference and learning from "big data" in emerging intelligent systems with significantly higher energy efficiency than that of the transistor-based Turing computing platform.

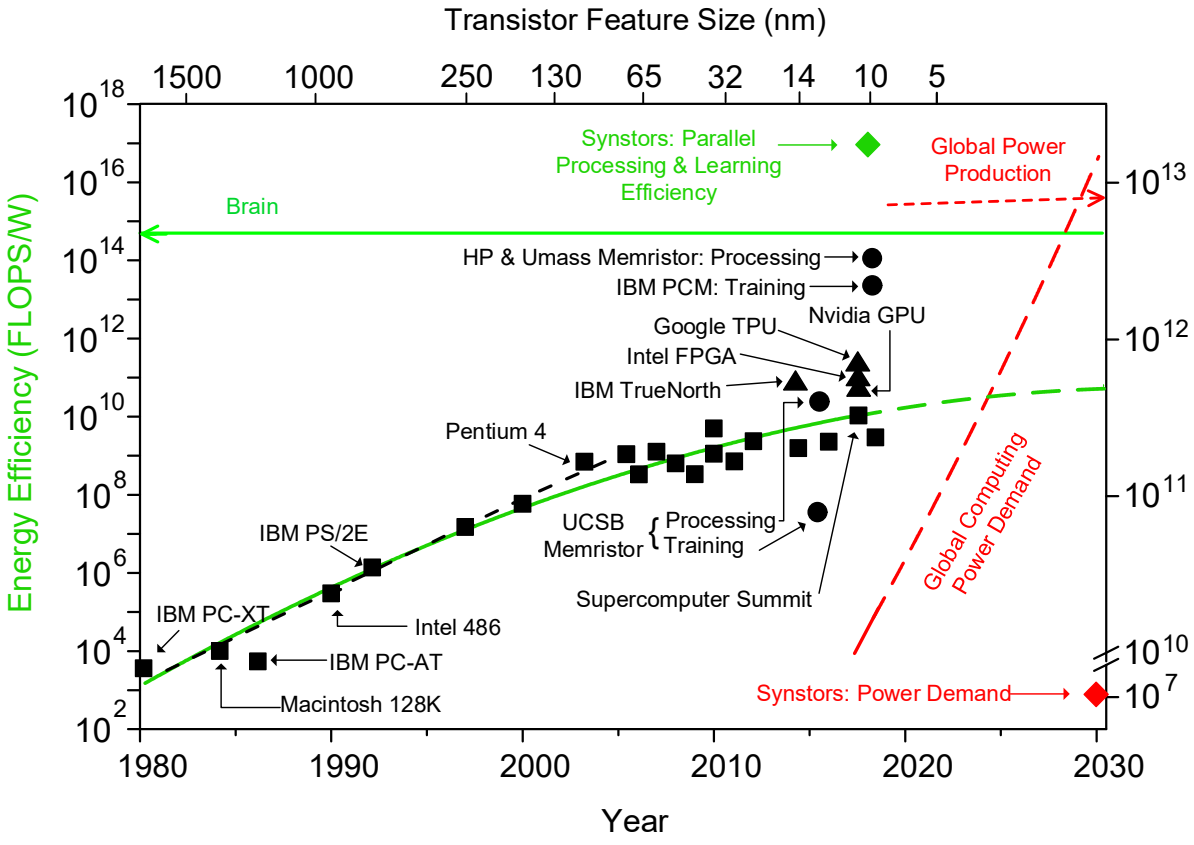


**Figure 1.** Comparison of energy efficiencies. The energy efficiencies of the human brain (green line), Summit supercomputer, personal computers (squares), Volta V100 graphics processing units (GPUs) from Nvidia, tensor processing units (TPUs) from Google, Stratrix 10 field-programmable gate array (FPGA) from Intel, TrueNorth neuromorphic circuit from IBM (triangles), memristor circuits from UCSB and UMass/HP, phase change memory (PCM) circuit from IBM (circles), and a synstor circuit reported in this work (green diamond) are shown (left y-axis) in the unit of floating point operations per second per watt ($FLOPS/W$) versus their introduction years. The different energy efficiencies of the memristor and PCM circuits for

signal inference and training are displayed separately. The trend lines of the energy efficiencies of digital computers based on Si-transistors from 1975 to 2009 (black dashed line) and from 1980 to 2018 (green line) are displayed. The projected global power production (dot-dashed red line) and global computing power demands (red dashed line) based on exponentially increasing data volume and the energy efficiencies of digital computers are also displayed. The global power demand in 2030 based on the energy efficiency of synstor circuits reported in this work is shown as a red diamond.

## 1.2. Neural Networks of the Human Brain

The human brain performs inference and learning from "big data" with an estimated speed ($\sim 10^{16}\ FLOPS$)[17] comparable to the speed ($\sim 10^{17}\ FLOPS$) of the fastest supercomputer, Summit,[18] but consumes much less power ($\sim 20\ W$) than the supercomputer ($\sim 10^{7}\ W$), and is much more energy-efficient ($\sim 10^{15}\ FLOPS/W$) than the supercomputer ($\sim 10^{10}\ FLOPS/W$, Figure 1). By contrast, the human brain concurrently performs spatiotemporal inference and learning in analog parallel mode[17,19-21] (Figure 2a) via a network of neurons connected by $\sim 10^{14}$ synapses (Figure 2b).

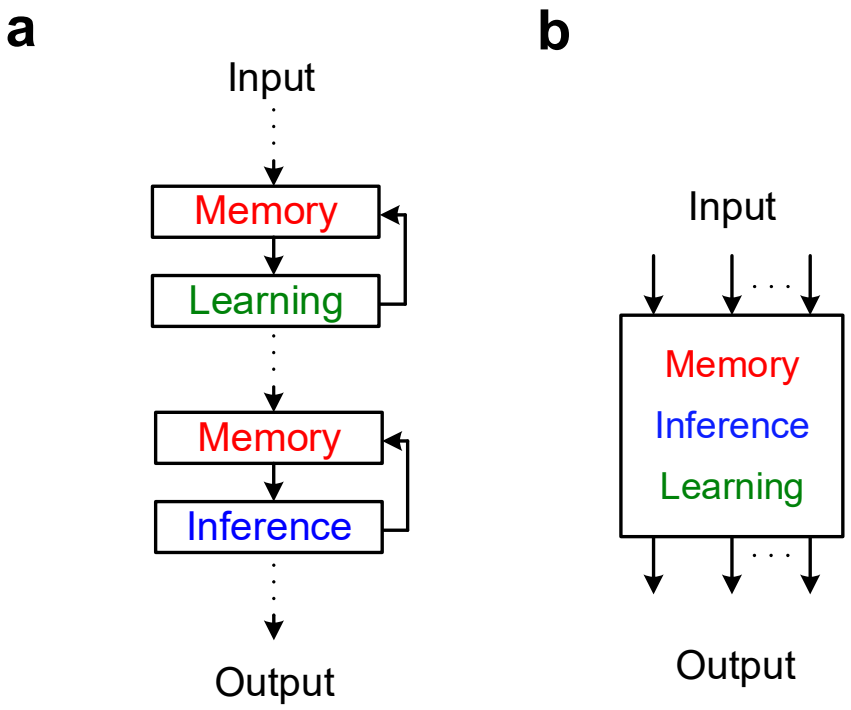


**Figure 2.** Turing model versus synapse and synstor circuits. a) Based on the Turing model, a computer executes inference and learning algorithms on separated logic and memory units in serial mode with data transitions between them. b) By integrating analog convolutional processing, correlative learning, and nonvolatile analog memory functions on each synapse (or synstor), a circuit of synapses (or synstors) perform inference and learning on multi-dimensional signals concurrently in parallel mode.

For inference, a wave of voltage pulses, $V_i^m(t)$, in the m$^{th}$ presynaptic neuron is processed by a synapse connected with the m$^{th}$ presynaptic and n$^{th}$ postsynaptic neurons, and induces a current in the n$^{th}$ postsynaptic neuron[22], $I^{nm} = \kappa^{nm} \circledast (w^{nm} V_i^m)$, where $w^{nm}$ denotes the synaptic weight (conductance), $\kappa^{nm}(t)$ denotes a temporal kernel function, and $\kappa^{nm} \circledast (w^{nm} V_i^m)$ represents the temporal convolution between $\kappa^{nm}$ and $w^{nm} V_i^m$. For spatiotemporal parallel

inference, a wave of voltage pulses in presynaptic neurons induces a collective current via synapses in the n$^{th}$ postsynaptic neuron (Figure 3), which can be expressed as,[22]

$$I^n(t) = \sum_m \kappa^{nm} \circledast (w^{nm} V_i^m) \quad (1)$$

and the current induces voltage pulses, $V_o^n(t)$, in the n$^{th}$ postsynaptic neuron. When the voltage pulse is fired in the postsynaptic neuron ($V_o^n \neq 0$), the postsynaptic current $I^n = 0$. The $w^{nm}$ matrix is also modified concurrently by the spatiotemporal waves of voltage pulses in the presynaptic and postsynaptic neurons for learning,[20,22,23]

$$\dot{w}^{nm} = \alpha\, V_i^m V_o^n \quad (2)$$

where $\alpha$ denotes the conductance modification coefficient, and $V_o^n$ and $V_i^m$ voltage pulses have the same amplitudes and durations. $w^{nm}$ is modified when $V_i^m = V_o^n$, with the learning coefficient $\alpha > 0$ in Hebbian learning, and $\alpha < 0$ in anti-Hebbian learning. $\alpha$ is a function of the timing difference between $V_i$ and $V_o$ pulses in the learning based on synaptic spike-timing-dependent plasticity (STDP). Based on Equation 2, general correlative learning algorithms in machine learning[20] can also be implemented (Section 3.4). Following Equation 2, when $V_i^m \cdot V_o^n = 0$ (e.g. $V_i^m \neq 0$ and $V_o^n = 0$ during inference), $\dot{w}^{nm} = 0$, i.e. $w$ remains nonvolatile for memory. By integrating the analog convolutional processing (Equation 1), correlative learning (Equation 2), and nonvolatile memory functions in a single synapse, the brain circumvents the fundamental limitations such as physically separated memory units, data transmission between memory and logic units in computers, and concurrently executes the inference (Equation 1) and learning

(Equation 2) algorithms in a neural network in analog parallel mode with an energy efficiency more than five orders of magnitudes higher than that of the Summit supercomputer.

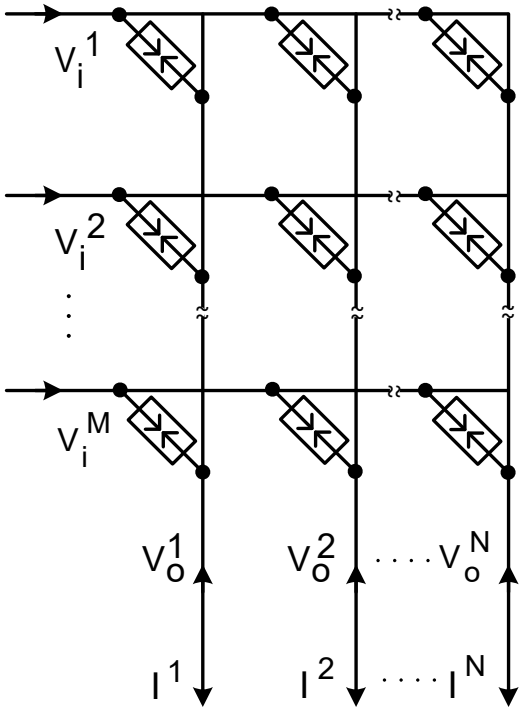


**Figure 3.** An M×N synapse array

## 1.3. Prior Art of Neuromorphic Circuits

Analog memory devices such as memory transistors,[24-27] memory resistors (memristors)[28-34], and phase change memory (PCM)[35,36] have been developed to emulate synapses. For inference, the neuromorphic circuits based on these devices processed input voltage pulses, $V_i^m$, and generated an output current by following Ohm's law, $I^n = \sum_m w^{nm} V_i^m$, in parallel analog mode with energy efficiencies of $\sim 10^{10} - 10^{14}\ FLOPS/W$,[29-36] which significantly exceeded the energy efficiencies of digital circuits (**Figure 1**). However, these devices lacked the function to trigger currents that lasted over time after the $V_i^m$ pulses ended, as described by Equation 1, which prevented the devices from performing convolutional signal processing of dynamic signals for

spatiotemporal inference.[22] For learning, the conductances of a memory transistor or memristor were modified by simultaneously applying writing voltage signals on their input and output electrodes, and correlative learning algorithms such as STDP were executed on individual devices by applying tailored voltage signals,[25,26,34] but a writing voltage signal applied on its input or output electrode alone would also change its conductance, therefore the devices not follow Equation 2 (i.e. $\dot{w}^{nm} = \alpha\, V_i^m V_o^n \neq 0$ when $V_i^m \neq 0$ and $V_o^n = 0$ or $V_o^n \neq 0$ and $V_i^m = 0$). To avoid change of conductance during signal inference, the voltage signals for inference were decreased to significantly smaller magnitudes than the voltage signals for learning, thus when the inference algorithm was executed in the circuits, the learning algorithm was interrupted, and vice versa.[24-36] Moreover, in order to modify device conductances accurately in a circuit, learning algorithms were executed in external digital circuits to obtain targeted conductance values, then devices were modified to the targeted conductance values by applying different writing voltages on different devices sequentially in iterative writing and reading processes.[29-36] The energy efficiencies for the writing processes were $\sim 10^5 - 10^{13}\ FLOPS/W$[29-36] (**Figure 1**), but the energy and time for the external digital computing circuits to execute correlative learning algorithms from "big data" with $M$-dimensional variables increase versus $M$ exponentially,[4,8] referred to as the "curse of dimensionality".[37] Although the circuits of the existing analog memory devices execute inference algorithms with high speeds and energy efficiencies, the differences between voltage signals for inference and learning prevent the circuits from executing inference (Equation 1) and learning (Equation 2) algorithms concurrently, and require separated memory and logic circuits, and signal transmissions between the circuits to execute learning algorithms, which limits their speeds and energy efficiencies for learning ($\lesssim 10^{11}\ FLOPS/W$).[15]

## 1.4. Research Goals

The goals of this research are 1) to develop a novel synaptic device that integrates the brain-like functions of signal processing, learning, and memory into a single device, 2) to implement these devices into a crossbar circuit with artificial neurons to form an artificial neural network, 3) demonstrate a highly energy efficient process of parallel inference and learning for speech recognition and 4) to implement concurrent inference and learning by self-programming of the network in real time, without the use of external circuits for calculating synaptic weight updates.

Fundamentally, the device must have analog tunable nonvolatile memory. Additionally, the materials and structure of the device must be optimized to enable the reading of the device conductance state using the same voltage amplitudes as for learning, without disrupting the nonvolatile memory. This is necessary to implement concurrent signal processing and learning of the synapse network without interrupting each mode. Then, the synapses can be modified efficiently by correlative Hebbian learning using the input pulse signals and backpropagation of the output pulse signals. The device conductance range must be controlled to be low, while still fulfilling the above requirements. The reduced overhead of peripheral circuits for programming the weights, implementation of a "synapse" in a simple 1-device unit cell, low device conductance range, and pulse operation will drastically reduce power consumption.

The artificial synapses need to be arranged into a crossbar circuit so that the network can compute multiply-accumulate (MAC) operations for matrix vector multiplication between the input vector and weight matrix by the intrinsic properties of the circuit. The synapses must have temporal dependent current, to enable convolutional signal processing. Then, in a crossbar, the synapse circuit can execute spatiotemporal inference. Additionally, this architecture allows for

backward propagation of the output vector to modify the matrix of synaptic weights by the correlation of input and output vectors. In order to solve classification problems, filter noise, and stabilize the optimized output, the synstor crossbar must be integrated with “neuron” circuits with nonlinear activation functions. Here, integrate-and-fire with sigmoidal output signals fulfill this requirement.

In this work, a synstor crossbar chip with the above properties will be designed, fabricated and tested. The output electrodes of the synstor chip will be connected to integrate-and-fire “neuron” circuits to form a Self-Programming Neuromorphic Integrated Circuit (SNIC). Lastly, this research aims to validate the capability and versatility of this platform with two demonstrations; speech recognition with energy efficiency five orders of magnitude higher than that of the Summit supercomputer; and real-time self-programming to optimize a real nonlinear feedback control system in a dynamic environment in with performance superior to computer-based and human controllers. In the latter demonstration, the shape of a morphing wing is optimized by the SNIC using sensing and actuation signals in a wind tunnel with dynamic wind speed.

# 2. Synstor Device

## 2.1. Structure

The structure of a carbon nanotube (CNT) synstor is shown in Figure 4. The synstor has an input electrode, an output electrode, as well as a reference electrode as a common electric ground like a synapse. The synstor is composed of a 20 μm-wide p-type semiconducting CNT network which forms Schottky contacts with the Al input and output electrodes. The CNT networks are fabricated on a 6.5 nm-thick $HfO_2$ dielectric layer on a 2.5 nm-thick and 10 μm-

wide $TiO_2$ charge storage layer on a 22 nm-thick $HfO_2$ dielectric layer on a 50 nm-thick and 10 μm-wide Al reference electrode. There is a 5 μm lateral space between the $TiO_2$ charge storage layer/Al reference electrode and Al input/output electrodes. The synstor has a transistor-like structure, but its reference electrode is always grounded, and it operates as a two-terminal device.

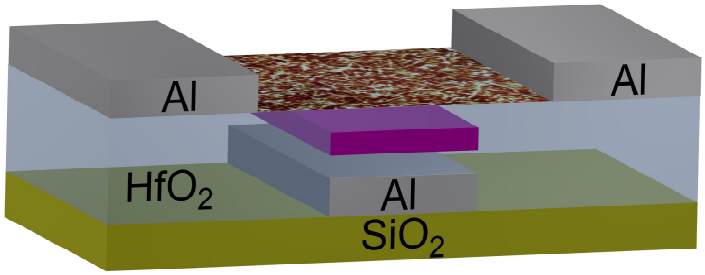


**Figure 4**. Synstor device structure. The synstor has Al input and feedback electrodes (gray), a randomly oriented semiconducting single-walled carbon nanotube (SWCNT) network channel (orange), $HfO_2$ dielectric barrier layers (blue-gray), a $TiO_2$ charge storage layer (magenta), and an Al reference electrode (gray).

## 2.2. Device Operation

As shown in Figure 5a, during inference mode, a negative voltage is applied to the input electrode $V_i$, while the feedback and reference electrodes are grounded. Electrons in the $TiO_2$ charge storage layer modulate the concentration of holes in the p-type doped CNT channel, controlling the channel conductance $w$ and current $I$ of holes flowing across the CNT layer, output from the feedback electrode. When voltage pulses are applied to the input electrode, the I-V relation of the device can be modeled as a series RC circuit of channel resistance and reference-channel capacitance with current $I = \kappa \circledast (wV_i)$, where $\kappa$ represents a temporal kernel function, $w$ represents the synstor weight (conductance) and $\circledast$ represents the temporal convolution. The kernel function $\kappa$ is a function of the voltage pulse profile and the device

resistance and capacitance. During inference mode, the conductance remains nonvolatile, and $\dot{w} = 0$.

During learning mode, equal amplitude voltages $V_i$ and $V_o$ are simultaneously applied to the input and feedback electrodes, respectively, while the reference electrode is grounded (Figure 5b). Current across the channel $I = 0$, while electrons diffuse between the CNT channel and $TiO_2$ charge storage layer as current $I_r \neq 0$ due to the relative channel-reference electrode voltage difference, modifying the channel conductance $w$ as the electric field from electrons in the $TiO_2$ charge storage layer attract or repel holes in the CNT layer.

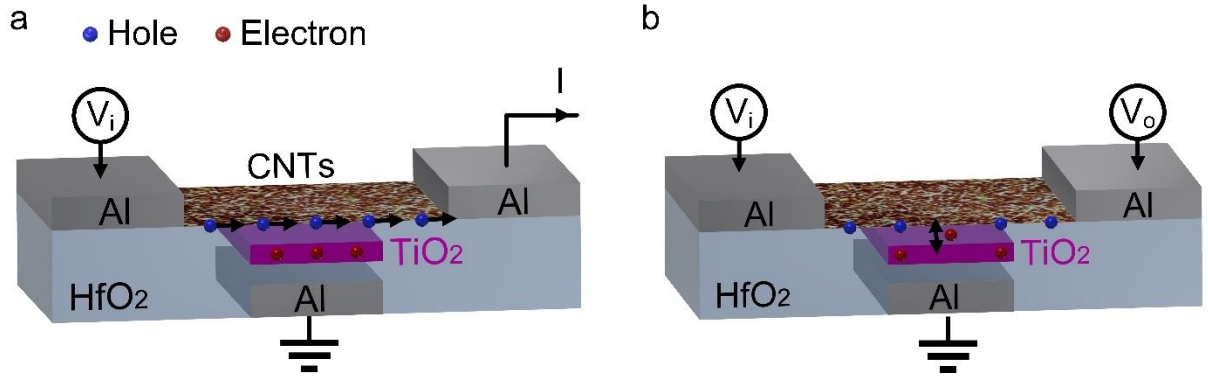


**Figure 5.** Synstor during inference and learning modes. a) Synstor in inference mode. b) Synstor in learning mode.

## 2.2. Energy Band Simulations

The devices were simulated by Technology Computer-Aided Design (TCAD) simulator (Sentaurus Device, Synopsys). The simulator performed numerical simulations of device physics based on partial differential equations of electrostatics, quantum mechanics, and carrier transport under a set of boundary conditions defined by device structures, and electronic properties and band diagrams were extracted from the simulations. Transient and quasi-stationary simulations were

conducted under different voltage biases on the Al input/output electrodes with respect to the grounded reference electrodes.

In Figure 6a, a scheme shows the cross-sectional structure of a synstor with a scale to mark the lateral distance, L. Simulated electronic band diagrams are plotted along the Al input electrode, the CNT network (orange), and the Al output electrode in the synstor under. Figure 6b displays the energy bands under the conditions of $V_i = V_o = -1.75\ V$ (green), $V_i = V_o = 1.75\ V$ (red), and Figure 6c shows the conditions of $V_o = 0$, $V_i = -1.75\ V$ (top), and $V_i = 1.75\ V$ (bottom). CNT energy-band diagrams with negative charge (green lines), and positive charge (red lines) stored in the $TiO_2$ layer are also shown in Figure 6c. $E_f^i$ , $E_f^o$, $E_f^r$, and $E_f^{TiO_2}$ denote the Fermi energies of the Al input, output, reference electrodes, and $TiO_2$ charge storage layer, respectively. The electronic charge is represented by "e". $E_c$ and $E_v$ denote the edges of the CNT conduction and valence bands, respectively. Electrons injected into or depleted from the $TiO_2$ layer are illustrated as the filled green circles, and holes in CNTs, $TiO_2$, or transported laterally along CNTs are illustrated as the open blue circles. The purple dot-dashed lines represent the Fermi energy of the $TiO_2$ charge storage layer. The black dot-dashed lines represent the Fermi energies of the CNT network and the Al input and output electrodes.

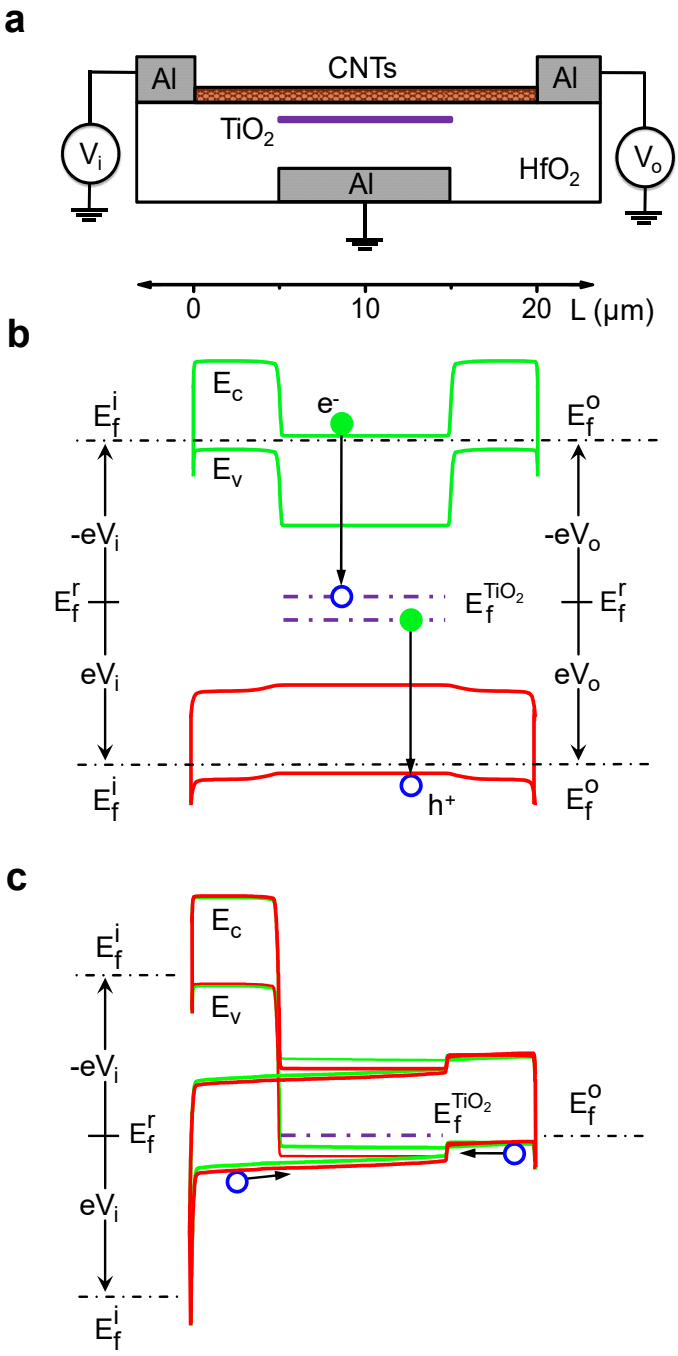


**Figure 6.** Simulated energy band structures during inference and learning.a) The synstor device cross section. b) Energy band diagrams under learning conditions. c) Energy band diagrams under read conditions.

In Figure 7a, on the left, a scheme shows the cross-sectional structure of a synstor composed of a $TiO_2$ charge storage layer (purple) embedded in a $HfO_2$ dielectric layer, an electrically grounded Al reference electrode (grey), and a carbon nanotube (CNT) network (orange) connected with an Al input electrode (grey) and an Al output electrode (grey). A voltage, $V_i$, is applied on the input electrode, and a voltage, $V_o$, is applied on the output electrode. On the right, a scheme shows a cross-sectional structure of a synstor along a single CNT.

Figure 7b-g display electronic energy-band structures along the Al input electrode, the CNT network, and the Al output electrode, on the left, and along CNT, the $HfO_2$ dielectric layer, the $TiO_2$ charge storage layer (orange), the $HfO_2$ dielectric layer, and the Al reference electrode on the right, under the conditions of $V_i = V_o = 0$ (Figure 7b), $V_i = V_o = -1.75\ V$(Figure 7c), $V_i = V_o = 1.75\ V$ (Figure 7d), $V_i = -1.75\ V$ and $V_o = 1.75\ V$ (Figure 7e), $V_i = -1.75\ V$ and $V_o = 0$ (Figure 7f), and $V_i = 1.75\ V$ and $V_o = 0$ (Figure 7g). CNT energy-band diagrams with no charge (orange), negative charge (green lines), and positive charge (red lines) stored in the $TiO_2$ layer are also shown in Figure 7f- g. $E_c$ and $E_v$ denote the edges of the CNT conduction and valence bands. $E_f^i$ , $E_f^o$, $E_f^r$, $E_f^{CNT}$, and $E_f^{TiO_2}$ denote the Fermi energies of the Al input, output, reference electrodes, CNT, and $TiO_2$ charge storage layer, respectively. The L and D scales in Figure 7b represent the lateral distance from the input electrode to the output electrode, and the vertical distance from CNT to the reference electrode, respectively. The differences between $E_f^i$, $E_f^o$, and $E_f^r$ are marked in Figure 7b-g on the left. The differences between $E_f^{CNT}$ and $E_f^{TiO_2}$ are marked in Figure 7c-g, on the right. Electrons (e) driven by the large differences between $E_f^{CNT}$ and $E_f^{TiO_2}$ to hop across the $HfO_2$ dielectric layer between the CNT network and the $TiO_2$ layer are illustrated by green dashed lines in Figure 7c and d).

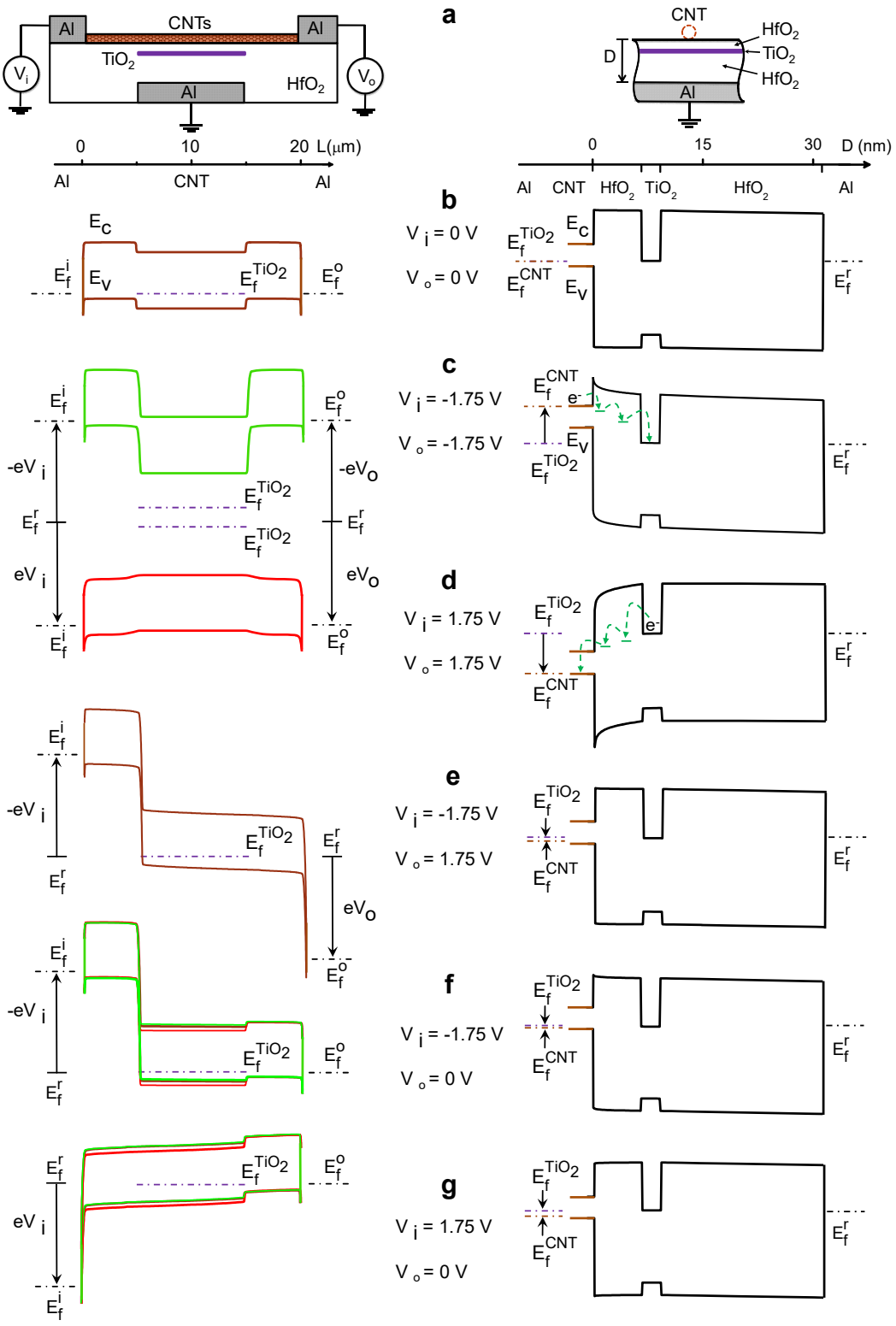


**Figure 7.** Device cross-section and energy band diagrams.

## 2.3. Fabrication

A 200 mm diameter [1 0 0] single crystal single crystal silicon boule was grown by the Czochralski method and diced to form a 725 μm thickness wafer (Silicon Quest International). The Si wafer was doped p-type by boron ions and oxidized by thermal oxidation to grow a 100 nm film of SiO2 (Silicon Quest International). The $Si/SiO_2$ wafer was diced into 3 cm x 3 cm square pieces and scribed with alignment marks for photolithography (DISCO Corporation). A 10 μm long and 50 nm thick Al reference electrode (Figure 8a) was deposited by electron beam (e-beam) evaporation (CHA Industries, CHA Mark 40), and patterned by photolithography and wet chemical etching with tetramethylammonium hydroxide (TMAH) based photoresist developer (AZ 300 MIF Developer). A 22 nm thick HfO2 barrier layer and a 2.5 nm thick $TiO_2$ charge storage layer (Figure 8b) were deposited by thermal atomic layer deposition (Cambridge NanoTech, Fiji Thermal and Plasma ALD). The $TiO_2$ film was patterned (Figure 8c) by photolithography and $CF_4/O_2$ (5:1 pressure) reactive ion etching (Technics RIE) using 200 W RF power at 150mTorr pressure for 2 minutes to form a 10 μm long pattern aligned to the Al reference electrode. A 6.5 nm thick HfO2 barrier layer (Figure 8d) was deposited by ALD, encapsulating the patterned $TiO_2$ charge storage layer. The surface was coated by an adhesion monolayer of poly (L-lysine) (PLL) from a 0.1% weight/volume (w/v) aqueous solution of PLL (Sigma Aldrich) by immersion of the wafer for 30 minutes to adsorb the PLL to the chip surface. A randomly oriented network of semiconducting single walled carbon nanotubes (SWCNTs) was deposited by immersion coating (Figure 8e) in an aqueous 99.9% pure semiconducting SWCNT aqueous (0.001% w/v CNTs) surfactant solution (Nanointegris, IsoNanotubes-S99.9%). Residual surfactant was removed from the surface by immersion in isopropanol (IPA) for 1 hour, rinsing with IPA, and drying by nitrogen blow dry. CNTs were doped p-type by adsorbing $O_2$ acceptors

from atmosphere. A 50 nm Al film (Figure 8f) was deposited by e-beam evaporation and patterned by the same process used for the Al reference electrode to form input and feedback electrodes. The CNTs were capped by a Parylene-C (PLC) polymer passivation layer deposited by thermal evaporation (Figure 8g) with a 175 °C vaporizer temperature, 25 mT pressure (Specialty Coating Systems, 2010 Parylene Vacuum Deposition System). The CNT network and PLC layer were patterned by SU-8 photoresist photolithography and $O_2$ RIE to form a 20 μm long CNT channel (Figure 8g). The SU-8 is an etch mask for the CNTs and PLC during $O_2$ RIE and encapsulates the CNTs and PLC prevent ambient doping of CNTs by atmosphere.

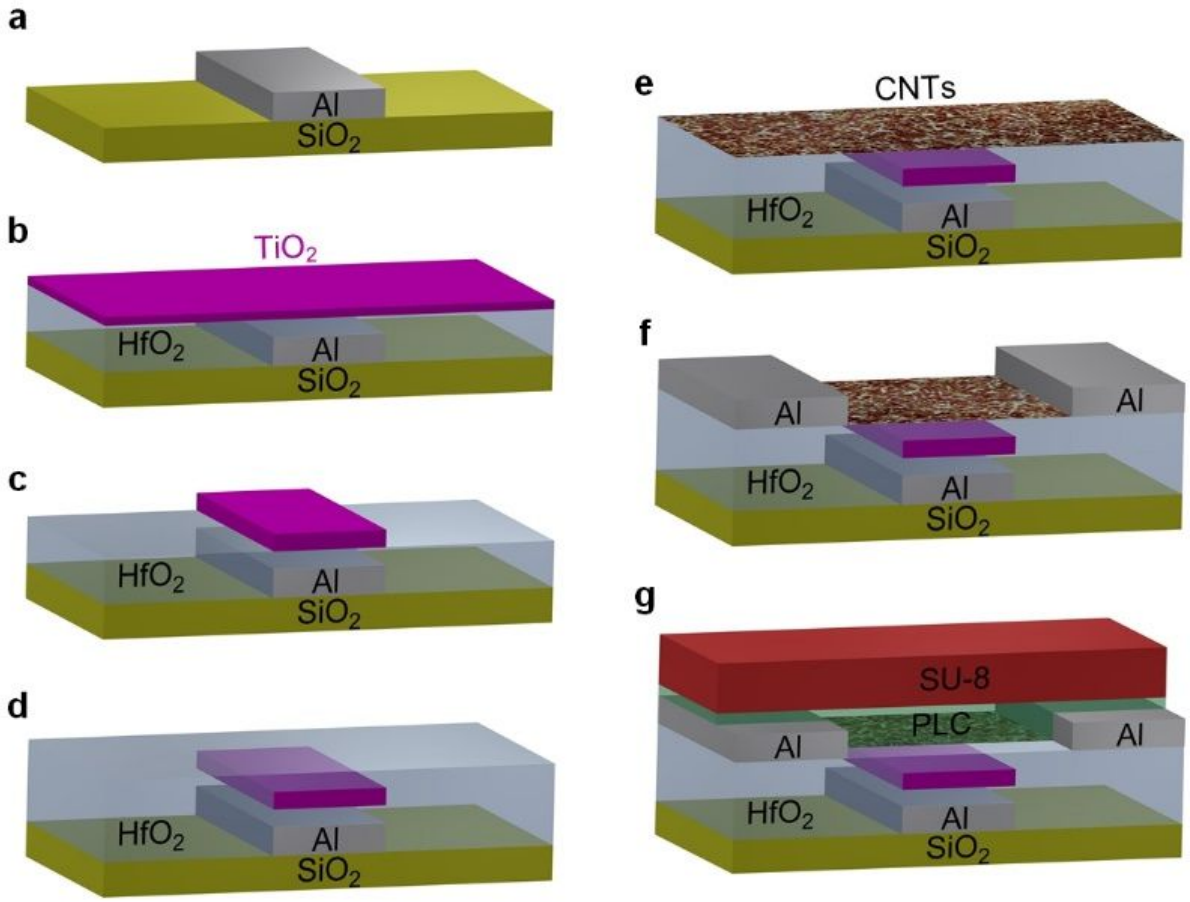


**Figure 8.** The fabrication process flow of a synstor. a) Al reference electrode deposition. b) $HfO_2$ barrier layer and $TiO_2$ charge storage layer deposition. c) $TiO_2$ charge storage layer patterning. d)

$HfO_2$ barrier layer deposition. e) CNT channel deposition. f) Al input and output electrode deposition. g) Parylene-C passivation layer deposition, SU-8 photoresist encapsulation layer deposition, and patterning of Parylene-C and CNT layers.

An AFM image in Figure 9a shows the active area of a synaptic resistor (synstor) with a 50 nm-thick Al input electrode, a 50 nm-thick Al output electrode, and a 6.5 nm-thick hafnium oxide ($HfO_2$) dielectric layer on a 2.5 nm-thick and 10 $\mu m$-wide titanium oxide ($TiO_2$) charge storage layer on a 22 nm-thick $HfO_2$ dielectric layer on a 50 nm-thick and 10 $\mu m$-wide Al reference electrode. The AFM image also shows 5 $\mu m$-wide spaces between the Al input/output electrodes and the $HfO_2/TiO_2/HfO_2$/Al reference electrode. In Figure 9b, a high-resolution AFM image shows randomly oriented carbon nanotubes (CNTs) in the 20 $\mu m$-wide CNT network.

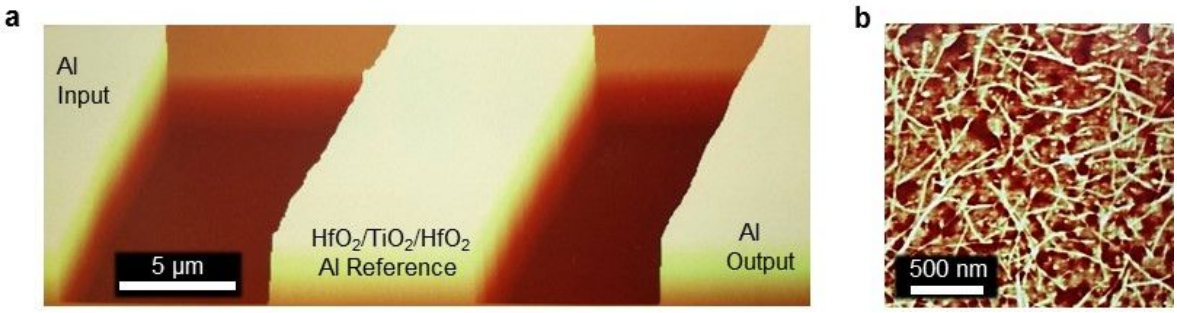


**Figure 9.** AFM images of synstor. a) AFM image of the active area of a synstor. b) AFM image of a randomly oriented CNT network on a synstor.

## 2.4. Device Testing

### 2.4.1. I-V Characteristics

A synstor was tested with a continuous voltage sweep, $V_i$, on its input electrode and a grounded output electrode. The current flowing through a synstor, $I$, was measured as a function of $V_i$, and displayed in Figure 10. The nonlinear rectifying $I - V_i$ curves indicate that Schottky barriers were formed between the Al input/output electrodes and the p-type semiconducting CNTs, as reported previously.[38] The DC conductance $w$ of the synstor is a nonlinear function of $V_i$. The device was modified by applying 50 pairs of 5 ms-wide $V_i$ and $V_o$ voltage pulses on their input and output electrodes simultaneously with the same amplitude ($V_i = V_o$). As shown in Figure 10, after the device experienced 50 pairs of $V_i$ and $V_o$ pulses with $V_i = V_o = 1.75\ V$, $w$ was decreased; after the device experienced 50 pairs of $V_i$ and $V_o$ pulses with $V_i = V_o = -1.75\ V$, $w$ was increased.

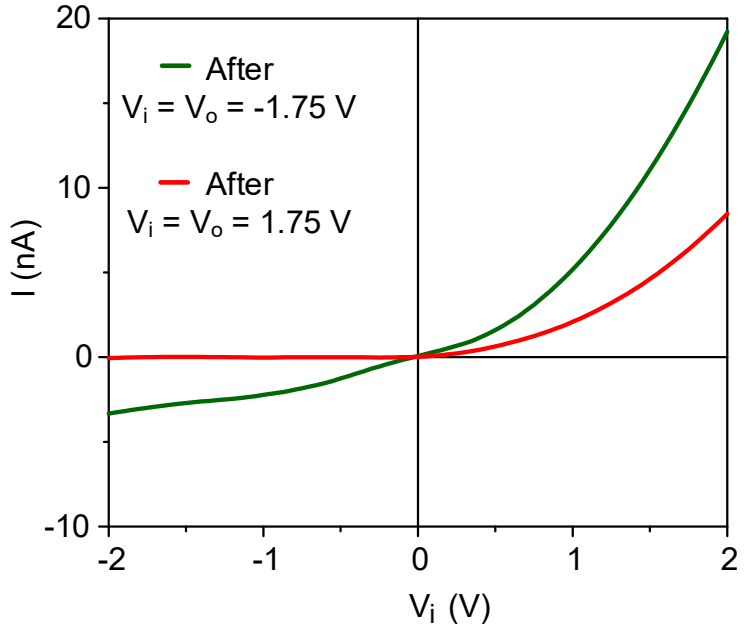


**Figure 10.** Synstor I-V characteristics after voltage pulses. Synstor currents, $I$, are plotted versus $V_i$, after the synstor was modified by 50 pairs of $5\ ms$-wide $V_i$ and $V_o$ pulses with $V_i = V_o = -1.75\ V$ (green line), and $V_i = V_o = 1.75\ V$ (red line) on the synstor.

## 2.4.2. Conductance Change by Voltage Pulses

The device was modified by applying 50 pairs of 5 ms-wide $V_i$ and $V_o$ voltage pulses on their input and output electrodes simultaneously with the same amplitude ($V_i = V_o$). As shown in Figure 11a, after the device experienced 50 pairs of $V_i$ and $V_o$ pulses with $V_i = V_o = 1.75\ V$, $w$ was decreased; after the device experienced 50 pairs of $V_i$ and $V_o$ pulses with $V_i = V_o = -1.75\ V$, $w$ was increased. The experimental data, $\Delta w(n)/w_0$, were fitted by $\Delta w(n)/w_0 = k_n^+ Ln\left(\frac{n}{n_0^+} + 1\right)$ with $k_n^+ = 7.5$ and $n_0^+ = 1.7 \times 10^3$ for $V_i = V_o = 1.75\ V$; and by $\Delta w(n)/w_0 = k_n^- Ln\left(\frac{n}{n_0^-} + 1\right)$ with $k_n^- = 15.3$ and $n_0^- = 1.76 \times 10^5$ for $V_i = V_o = -1.75\ V$ (Figure 11a). The percentage changes of the synstor conductance, $\Delta w/w_0$, induced by various $10\ ns$-wide $V_i$ and $V_o$ pulses with $V_i = V_o = -1.75\ V$ (green triangles), $V_i = V_o = 1.75\ V$ (red triangles), $V_i = V_o = 0$ (purple), $V_i = 0$ and $V_o = -1.75\ V$ (orange), $V_i = 0$ and $V_o = 1.75\ V$ (cyan), $V_i = 1.75\ V$ and $V_o = 0$ (yellow), and $V_i = -1.75\ V$ and $V_o = 0$ (blue) are plotted versus the applied pulse numbers, $n$.

The changes of $w$, $\Delta w = w - w_0$, were also measured before and after the synstors experienced 50 pairs of 5 ms-wide $V_i$ and $V_o$ pulses with various amplitudes ranged between $-2\ V < V_i = V_o < 2\ V$, and the percentage changes of $w$, $\Delta w/w_0$, are plotted versus the pulse amplitudes in Figure 11b. The $w$ is modified by following Equation 2, $\dot{w} = \alpha\ V_i \cdot V_o$. When $V_i = V_o \gtrsim 1.0\ V$, $w$ was decreased ($\alpha < 0$); when $V_i = V_o \lesssim -0.8\ V$, $w$ was increased ($\alpha > 0$); when $-0.8\ V \lesssim V_i = V_o \lesssim 1.0\ V$, $\dot{w} \approx 0$ ($\alpha \approx 0$). The experimental data, $\Delta w/w_0$, (Figure 11b) were fitted (solid black lines) by $\Delta w/w_0 = e^{\beta_v^+ (V_i - V_t^+)} - 1$ with $\beta_v^+ = 4.06/V$ and $V_t^+ = 1.05\ V$

under $V_i = V_o > V_t^+$; and $\Delta w/w_0 = e^{-\beta_v^-(V_i - V_t^-)} - 1$ with $\beta_v^- = 3.69/V$ and $V_t^- = -0.81\ V$ under $V_i = V_o < V_t^-$; $\Delta w \approx 0$ under $V_t^+ > V_i = V_o > V_t^-$.

After synstors were modified to their analog conductances, the synstors were tested under $V_i \cdot V_o \leq 0$ by applying 50 pairs of various 5 ms-wide $V_i$ and $V_o$ pulses under the conditions: (1) $-2\ \mathrm{V} < V_i < 2\ V$ and $V_o = 0$; (2) $-2\ \mathrm{V} < V_o < 2\ V$ and $V_i = 0$; and (3) $-2\ \mathrm{V} < V_i = -V_o < 2\ V$. It was observed that $w$ remained unchanged under these conditions (Figure 11b), which indicates that the synstor has a nonvolatile memory of $w$ (i.e. $\dot{w} \approx 0$ (Equation 2)) under $V_i \cdot V_o \leq 0$. A series of 10 $ns$-wide $V_i$ and $V_o$ pulses with the same amplitude ($V_i = V_o$) were applied on a synstor simultaneously (without the timing difference between $V_i$ and $V_o$ pulses), and its DC conductance, $w$, was measured before and after the pulses were applied. As shown in Figure 11a, when a series of 10 $ns$-wide paired $V_i$ and $V_o$ pulses with $V_i = V_o = 1.75\ V$ (or $V_i = V_o = -1.75\ V$) were applied on the synstor, $w$ was gradually decreased (or increased) versus the number of pulse pairs, $n$. When a series of 10 $ns$-wide $V_i$ and $V_o$ pulses with $V_i = 0$ or $V_o = 0$ were applied on the synstor, the average percentage changes of the conductances, $\overline{|\Delta w/w_0|} \lesssim 3.3\ \%$.

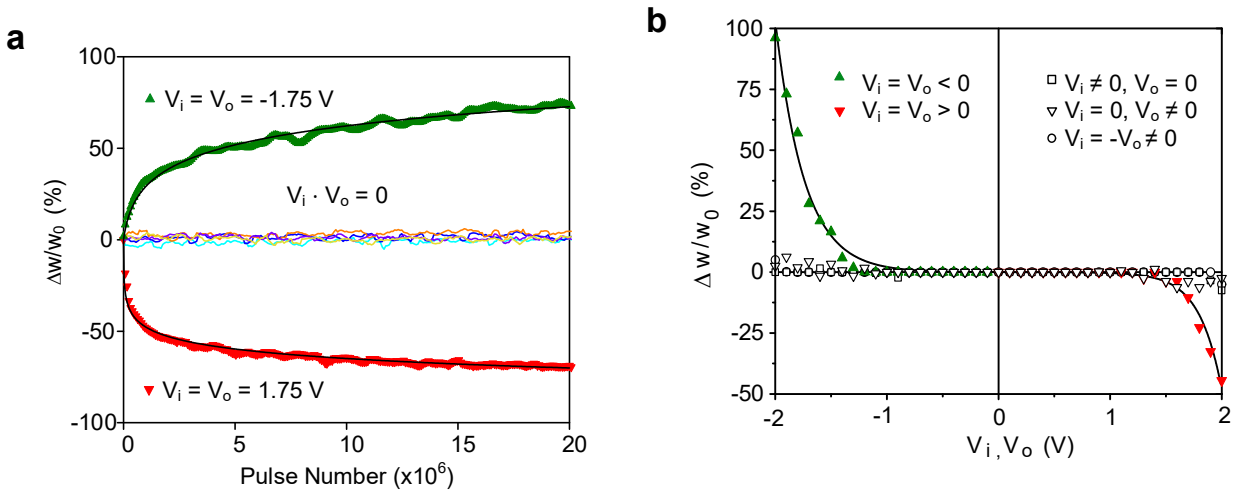


Figure 11. Conductance change versus pulse number (a) and voltage amplitude (b).

To derive the synstor conductance $w$ in the above experiments, input pulses $V_i = -1.75\ V$ were applied to a synstor with output electrode connected to an operational amplifier (Microchip Technologies, MPC6022) in an inverting op-amp configuration. The output voltage of the op-amp $V_{out}$ was measured to calculate $w = \frac{I}{V_i} = -\frac{V_{out}}{V_i R_f}$, where $R_f$ was a feedback resistor connecting the inverting input and output terminals of the op-amp.

## 2.4.3. Charge Density and Capacitance Measurements

As shown in the simulated electronic band structures in Figure 6b, a pair of negative (or positive) $V_i$ and $V_o$ voltages with $V_i = V_o = -1.75\ V$ (or $+1.75\ V$) increases (or decreases) the Fermi energy of the CNT network, and induces a difference of $+1.66\ eV$ (or $-1.59\ eV$) between the CNT and $TiO_2$ Fermi energies (Figure 12a), which injects electrons into (or depletes electrons from) the $TiO_2$ charge-storage layer by electronic hopping through the $HfO_2$ dielectric barrier layer[39,40] (Figure 7). The changes of the charge density in the $TiO_2$ layer, $\Delta\rho_s$, induced by the paired $V_i$ and $V_o$ voltages with $V_i = V_o$ were measured by capacitance-voltage tests on a synstor (Figure 13), and plotted as a function of $V_i$ and $V_o$ in Figure 12b. When the synstor experienced $V_i$ and $V_o$ voltages with $V_i = V_o \gtrsim V_t^+$ and $V_i = V_o \lesssim V_t^-$, $\Delta\rho_s$ increased versus $V_i, V_o$. The $\Delta\rho_s - V_i, V_o$ data were fitted by $\Delta\rho_s = k_\rho^+[V_a - V_t^+]$ under $V_i = V_o \gtrsim V_t^+$ or $\Delta\rho_s = k_\rho^-[V_a - V_t^-]$ under $V_i = V_o \lesssim V_t^-$ with $k_\rho^+ = -145\ nF/cm^2$, $k_\rho^- = -106\ nF/cm^2$, $V_t^+ = 0.92\ V$, and $V_t^- = -0.85\ V$ as the positive and negative threshold voltages to modify the charges in their storage layer (Equation 4). When the synstor experienced $V_i$ and $V_o$ voltages with $V_t^+ \gtrsim V_i = V_o \gtrsim V_t^-$, the external voltage could not drive a significant amount of electrons to overcome the potential barrier in the $HfO_2$ layer, therefore, no significant charge modification in the charge storage layer was

observed ($\Delta\rho_s \approx 0$) (Figure 12b). When the synstor experienced $V_i$ and $V_o$ voltages with $V_i = V_o \gtrapprox V_t^+$ and $V_i = V_o \lessapprox V_t^-$, the negative (or positive) charges in the storage layer attract (repel) the holes in the p-type semiconducting CNT network (Figure 6c), and increase (or decrease) the device conductance $w$ exponentially versus the magnitudes of $V_i$ and $V_o$ voltages (Equation 5). When a series of paired $V_i$ and $V_o$ pulses with $V_i = V_o$ were applied, $\rho_s$ was also modified by the external potential as a function of the number of the applied voltage pulses, $n$. $\rho_s$ also gradually builds up an internal potential against the external potential, resulting in $\Delta w(n)/w_0$ to change as a logarithm function of $n$ (Equation 5). A single-wall CNT with an average diameter of ~1 $nm$ locally forms a capacitor with the $TiO_2$ layer with an extremely low capacitance (~$10^{-19}\ F/nm$)[41]. The voltages applied on the CNT network with respect to the Al reference predominantly drop across the CNT/$HfO_2$/$TiO_2$ capacitor locally, driving electrons to hop through the $HfO_2$ dielectric layer. The low capacitance also allows the charge stored in the $TiO_2$ layer to influence the hole concentration and conductance of the CNT network significantly. The Fermi energies of the p-type CNT, Al, and $TiO_2$ materials are approximately equal (with differences less than $0.2\ eV$),[38,42] resulting in the symmetric Fermi energy differences, and similar charge and conductance modification rates by the positive and negative voltages with the equal magnitude following the learning algorithm $\dot{w} = \alpha\ V_i \cdot V_o$ (Equation 2).

When $V_i$ and $V_o$ voltages with $V_i \cdot V_o = 0$ were applied on a synstor, its electronic band structures were also simulated under the conditions of (1) $V_i = 1.75\ V$ and $V_o = 0$ (Figure 6c), (2) $V_i = -1.75\ V$ and $V_o = 0$ (Figure 6c), and (3) $V_i = 0$ and $V_o = 0$ (Figure 7b). Under these asymmetric $V_i$ and $V_o$ voltages, the positive $V_i$ or $V_o$ voltage mainly drops across the reverse-biased Schottky contact between the Al electrode and p-type CNTs, and the negative $V_i$ or $V_o$ voltage mainly drops across the hole-depletion region on the lateral space beyond the $TiO_2$

charge storage layer/Al reference electrode, which leads to small differences ($\lesssim 0.23\ eV$) between the Fermi energies of the CNT network and the recessed $TiO_2$ layer (Figure 12a). The changes of the charge density in the $TiO_2$ layer, $\Delta\rho_s$, induced by the $V_i$ and $V_o$ voltages under $-2\ V \leq V_i \leq 2\ V$ and $V_o = 0$ were measured by capacitance-voltage tests on a synstor (Figure 13), and plotted as a function of $V_i$ and $V_o$ in Figure 12. When the synstor experienced $V_i$ and $V_o$ voltages with $V_i \neq 0$ and $V_o = 0$, the observed charge density changes, $|\Delta\rho_s| < 25\ nF/cm^2$, which are less than $15\%$ of the charge density changes induced by the paired pulses with the same magnitude (Figure 11).

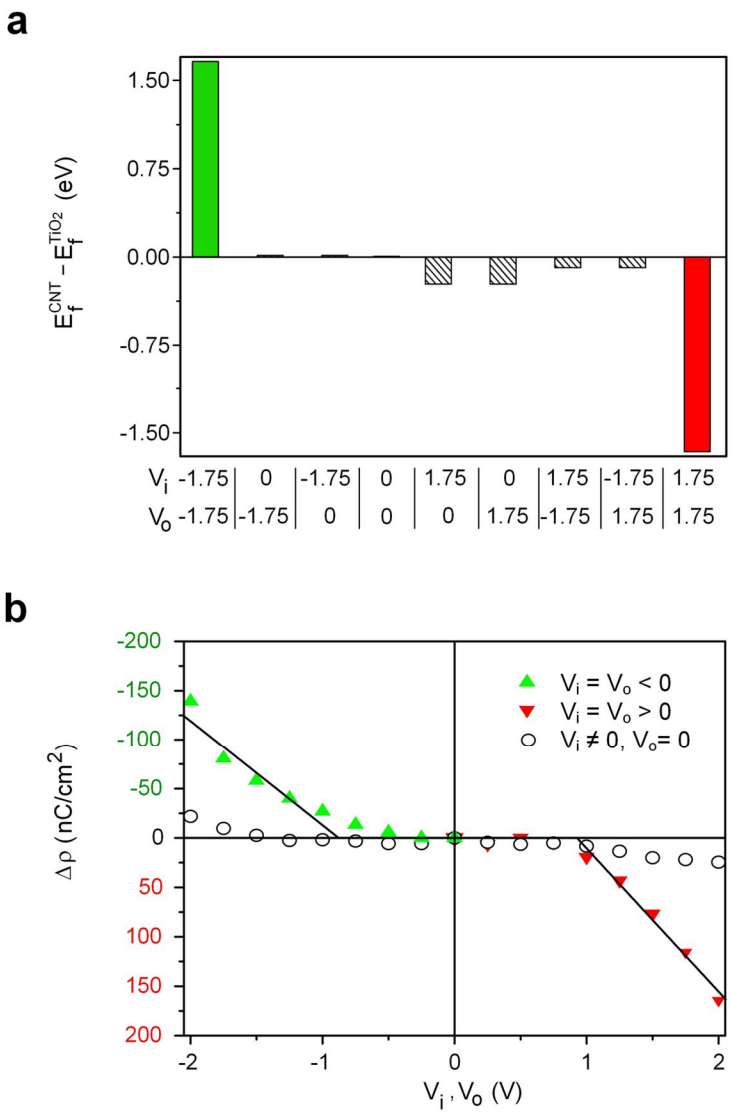


**Figure 12.** Fermi energy differences and change of charge density. a) The simulated differences between the average Fermi energies of the CNT network, $E_f^{CNT}$, and the $TiO_2$ charge storage layer, $E_f^{TiO_2}$, in a synstor under various combinations of $V_i$ voltages on its input electrode and $V_o$ voltages on its output electrode. b) The change of the charge density in the $TiO_2$ layer of a synstor, $\Delta\rho_s$, induced by various $V_i$ and $V_o$ voltages are measured by capacitance-voltage test and plotted versus the amplitudes of the $V_i$ and $V_o$ voltages.

Synstor capacitance, $C_s$, was measured by applying an AC sinusoidal voltage with an amplitude of $12.5\ mV$ and a frequency of $1\ kHz$, and a DC voltage bias, $V_D$, on its input and output electrodes with respect to its grounded reference electrode from an Agilent 4284A Precision LCR Meter. The $C_s - V_D$ curves are shown in Figure 13a. When $V_D < 0$, the holes in the CNT network above the Al reference electrode and $TiO_2$ charge storage layer were gradually depleted, and the measured capacitance, $C_s^-$, represented the capacitance of the capacitor of the depletion region in series with the capacitor between the CNT network and the Al reference electrode, $C_{CNT/Al}$. When $V_D$ decreased, the depletion region increased, resulting in the decrease of the capacitance of the depletion region and $C_s^-$. When $V_D > 0$, the holes in the CNT network above the Al reference electrode were gradually accumulated, and the CNT depletion region was decreased with increasing $V_D$, the measured capacitance, $C_s^+$, represented $C_{CNT/Al}$. When $V_D$ was swept between $2\ V$ and $-2\ V$, the charges in the charge storage layer were modified, which induced hysteresis loops in the $C_s - V_D$ curves (Figure 13a). $C_s$ was measured at $V_D = 0$ before and after the synstor experienced 50 5 $ms$-wide $V_i$ and $V_o$ voltage pulses on its input and output electrodes with various combinations of amplitudes, and the changes of the capacitance, $\Delta C_s$, are displayed in Figure 13b. When positive voltage pulses with $V_i = V_o > 0.92\ V$ were applied on the synstor, the electrons were depleted from the $TiO_2$ layer, leaving positive charge in the charge storage layer. The positive charge repelled holes in the CNT network, increased the depletion region, and decreased its capacitance, thus $\Delta C_s < 0$. When negative voltages $V_i = V_o < -0.85\ V$ were applied on the synstor, the electrons were injected into the $TiO_2$ layer, leaving negative charge in the charge storage layer. The negative charge attracted holes in the CNT network, decreased the depletion region, and increased its capacitance, thus $\Delta C_s > 0$. When the synstor experienced 50 5 $ms$-wide $V_i$ voltage pulses under $V_o = 0$, the observed $|\Delta C_s| < 100\ fF$, which are less than 15% of $|\Delta C_s|$

induced by the paired pulses with the same magnitude as $V_i$ (Figure 13b). From $\Delta C_s$ and the $C_s - V_D$ curves in Figure S9a, the shifts of the corresponding $C_s - V_D$ curves along the $V_D$ axis, $\Delta V_D$, were derived, and the changes of the charge density in the charge storage layer were extrapolated[40], $\Delta\rho_s = c_{TiO2/Al}\Delta V_D$, where $c_{TiO2/Al} \approx 1.0\ \mu F/cm^2$ denotes the capacitance per area of the 22-nm thick $HfO_2$ film sandwiched between the $TiO_2$ layer and Al reference electrode. $\Delta\rho_s$ is plotted as a function of $V_i$ and $V_o$ in Figure 12b.

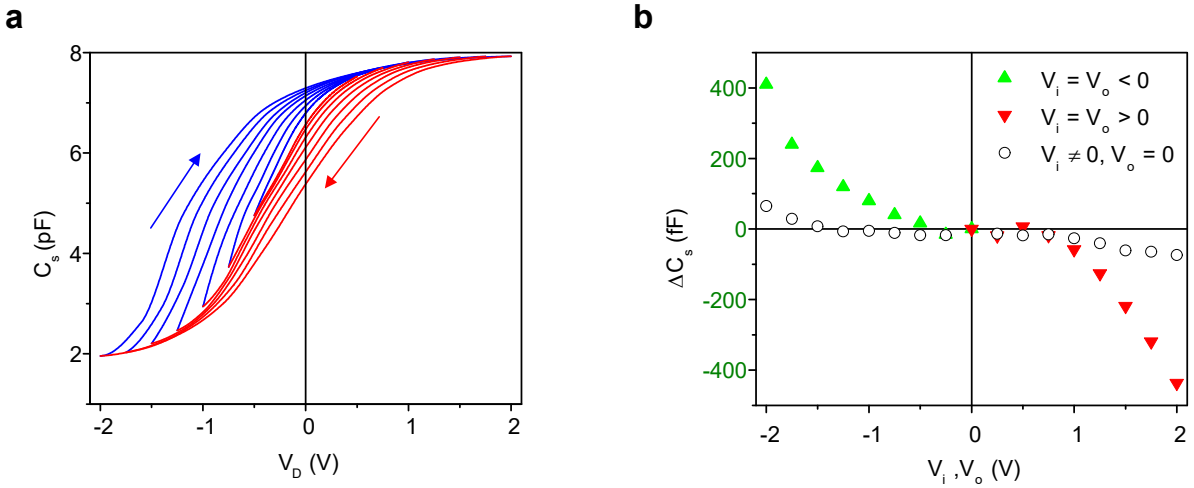


**Figure 13.** Capacitance measurements.a) The capacitance of a synstor, $C_s$, is displayed versus a DC voltage bias, $V_i = V_o = V_D$, applied on its input and output electrodes with respect to its grounded reference electrode when $V_D$ sweeps to different ranges from negative to positive voltages (blue) and from positive to negative voltages (red). . b) The changes of the capacitance, $\Delta C_s$, of a synstor measured at $V_D = 0$ after the synstor experienced 50 5 $ms$-wide $V_i$ and $V_o$ voltage pulses on its input and output electrodes

## 2.4.4. Pulse Currents for Convolutional Signal Processing

A series of periodic $V_i$ pulses with an amplitude of $-1.75\ V$, a period of $30\ ns$, and different durations ($8, 10, 15, 20\ ns$) were applied periodically on a synstor under $V_o = 0$, and the currents flowing through the synstor, $|I(t)|$, increased versus $t$ during the pulse, decayed versus time after the pulse (Figure 14 and Figure 15). $|I|$ also increased with increasing pulse number and duration.

A cross-section of the synstor with an Al/CNT/Al structure of a resistor and a CNT/$HfO_2$/$TiO_2$/$HfO_2$/Al structure of a capacitor is shown in Figure 6a. When a voltage pulse is applied on the input electrode of a synstor, it drives a current through the CNT network toward the grounded output electrode of the synstor, and simultaneously charges the capacitor between the CNT network and Al reference electrode. After the pulse ends, the capacitor is discharged, leading to a current through the output electrode. As shown in Figure 14 and Figure 15, the output current triggered by a series of $-1.75\ V$ periodic pulses with a period of $30\ ns$ and a duration $t_d = 8, 10, 15, 20\ ns$ from a synstor changed versus time by following Equation 1, $I(t) = (w\ \kappa) \circledast\ V_i$, with $V_i(t) = \sum_n V_a \delta(t - t_n)$, $V_a$ as the amplitude of the pulse, and $t_n$ as the moment to trigger the $n^{th}$ pulse. $\kappa(t) = 1 - e^{-\beta_p t}$ during the pulse ($t \leq t_d$), and $\kappa(t) = (1 - e^{-\beta_p t_d})e^{-\beta_d t}$ after the pulse ($t > t_d$), with $t_d$ as the duration of the pulse, $\beta_p$ and $\beta_d$ as the parameters related to the resistance and capacitance of the CNT network (Equation 3). $I(t)$ was fitted by $I(t) = w\ \kappa \circledast\ V_i$ with $\beta_d = 41.5\ MHz$, and $\beta_p = 0.73,\ 0.67, 0.47, 0.36\ MHz$ when $t_d = 8, 10, 15, 20\ ns$, respectively.

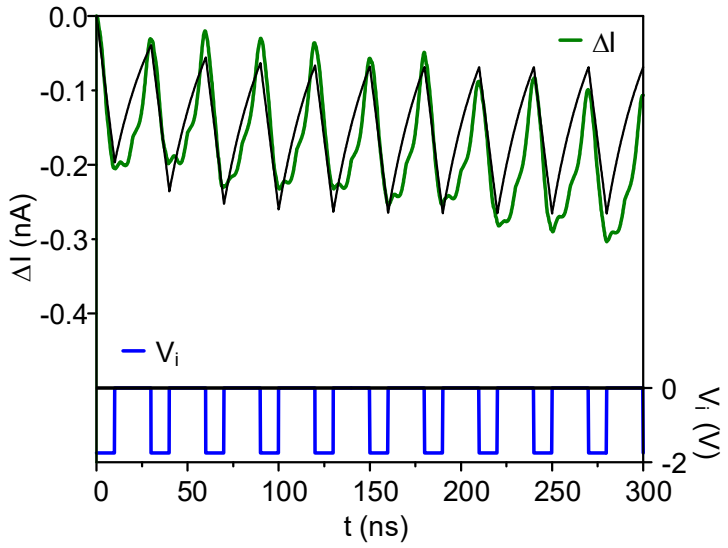


**Figure 14.** Output current of synstor after input voltage pulses.

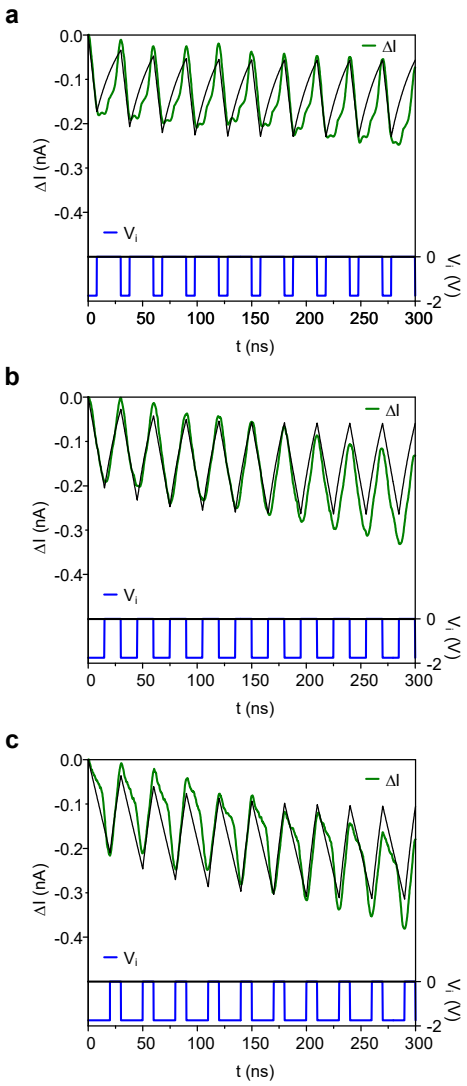


**Figure 15.** Currents triggered by input voltage pulses.The pulses have a firing rate of 33.3 $MHz$ and a duration $t_d$ equal to a) 8 $ns$; b) 15 ns; c) 20 ns.

## 2.4.5. Endurance, Uniformity, and Nonvolatile Memory

A synstor was modified to its high and low conductance values, $w_H$ and $w_L$, alternatingly in 2930 cycles, as shown in Figure 16. In each cycle, a synstor was modified to $w_H$ at first by applying 50 pairs of 5 ms-wide $V_i$ and $V_o$ pulses on its input and output electrodes with $V_i = V_o = -1.75\ V$, and then the synstor was modified to $w_L$ by applying 50 pairs of 5 ms-wide $V_i$ and $V_o$

pulses with $V_i = V_o = 1.75\ V$. After each set of pulses, $w_H$ and $w_L$ was measured by applying 10 ms-wide $V_i$ pulses with $V_i = -1.75\ V$ under $V_o = 0$. In each cycle, the percentage changes of the conductance, $\Delta w_H/w_{mid} = (w_H - w_{mid})/w_{mid}$ and $\Delta w_L/w_{mid} = (w_L - w_{mid})/w_{mid}$ are shown as green and red triangles, respectively, versus the number of modification cycles, with a mid-range conductance $w_{mid} = (w_H + w_L)/2$. Over the 2930 modification cycles, the average percentage conductance change, $\overline{|\Delta w/w_{mid}|} = (w_H - w_L)/w_{mid} = 51.6$ % with a cycle-to-cycle standard deviation of 1.2 %.

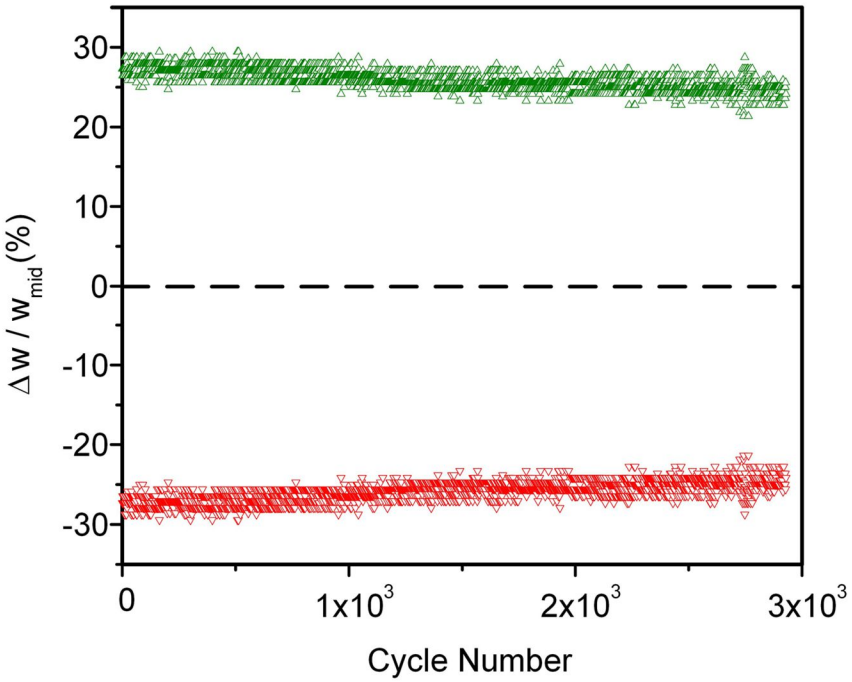


**Figure 16.** Synstor conductance changes in modification cycles. A synstor was modified to its high and low conductances, $w_H$ and $w_L$, iteratively in 2930 modification cycles

Figure 17,108 synstors on a chip were modified to their high DC conductance values, $w_H$, by applying 50 pairs of 5 ms-wide $V_i$ and $V_o$ pulses on their input and output electrodes with $V_i = V_o = -1.75\ V$, and the synstors were modified to their low DC conductance values, $w_L$, by

applying 50 pairs of 5 ms-wide $V_i$ and $V_o$ pulses with $V_i = V_o = 1.75\ V$. The synstor DC conductances, $w_L$ and $w_H$, were tested by applying 10 ms-wide $V_i$ pulses on their input electrodes with $V_i = -1.75\ V$ and $V_o = 0$. The distributions of $w_H$ (green) and $w_L$ (red) of the 108 synstors are plotted. The average $w_H$ value, $\overline{w}_H = 1.90$ nS, and the standard deviation of $w_H$, $\sigma_H = 0.44$ nS with $\sigma_H/\overline{w}_H \approx 0.23$. $w_L < 0.2$ nS, the limit of the electric module. Note the synstors can be modified to an arbitrary analog DC conductance within the range of $0.2\ nS \leq w \leq 1\ nS$.

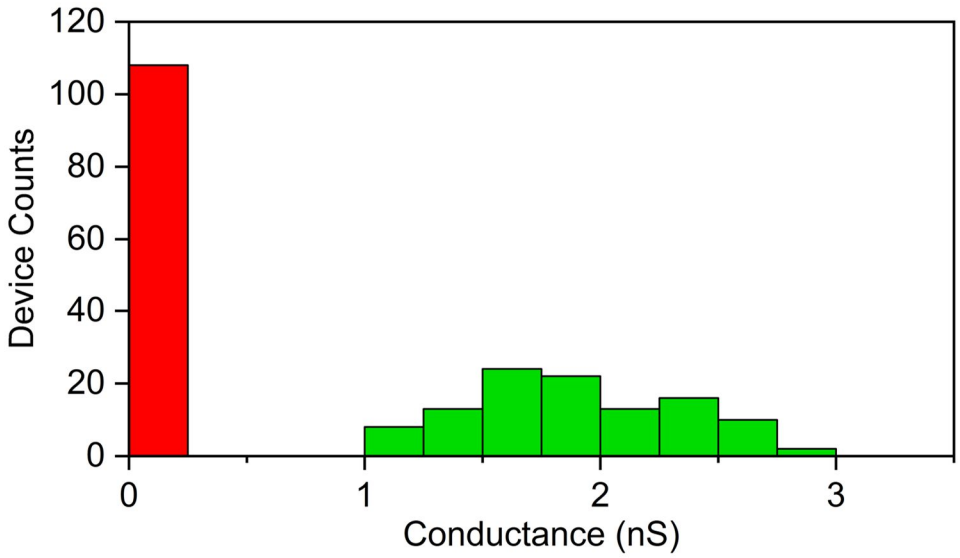


**Figure 17.** Synstor conductance distribution.

After synstors were modified to different initial analog conductances by applying pairs of 5 ms-wide $V_i$ and $V_o$ voltage pulses on its input and output electrodes simultaneously with $V_i = V_o = -1.75\ V$ or $V_i = V_o = 1.75\ V$, the nonvolatile memory of the synstors was examined by measuring their conductances versus time over $1.75 \times 10^5\ s$ at room temperature (Figure 18). The device DC conductances, $w$, were measured by applying 10 ms-wide $V_i$ pulses with $V_i = -1.75\ V$ under $V_o = 0$. $w$ is shown as a function of time on a logarithmic scale, and the experimental data (dots) are fitted and extrapolated (solid lines). Over the test period, the average

percentage conductance changes, $\overline{|\Delta w/w_i|} = \sim 3\ \%$, with a standard deviation of $\sim 2.5\ \%$. Based on the extrapolations of the experimental data, the analog conductances could be distinguished and preserved for more than 10 years, indicating their long-term nonvolatile analog memory.

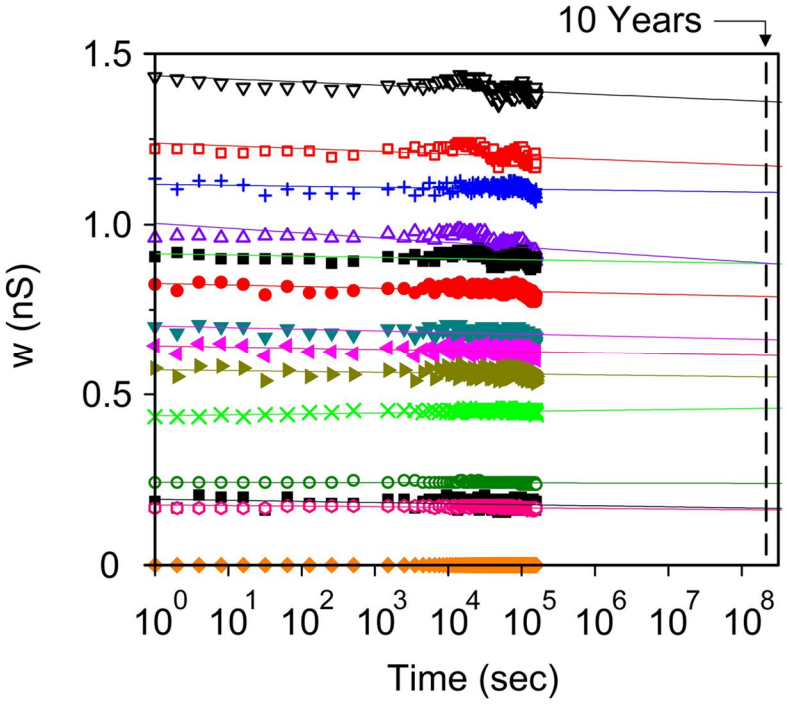


**Figure 18.** Nonvolatile memory test.

## 2.4.6. Methods

The reference electrodes of the synstors and control devices were always grounded during the tests. Current-voltage (I-V) characteristics were measured by a Keithley 4200 semiconductor parameter analyzer. The electrical voltage pulses ($V_i$ and $V_o$) applied to the input and output electrodes of the devices and circuits were generated by a field programmable gate array (FPGA, National Instruments, cRIO-9063), computer-controlled modules (National Instruments, NI-

9264), and a Tektronix AFG3152C waveform/function generator. Currents flowing through synstors were measured by a semiconductor parameter analyzer, computer-controlled circuit modules (National Instruments, NI-9205 and NI-9403), and oscilloscope (Tektronix TDS 3054B). The DC conductance of the devices was derived from the currents measured by applying negative voltage pulses with a duration of 5 ms and a magnitude equal to the voltages applied for learning. Testing protocols were programmed (NI LabVIEW) and implemented in an embedded field-programmable gate array (FPGA, Xilinx), a microcontroller, and a reconfigurable I/O interface (NI CompactRIO). During pulse tests, the device electrode pads (Figure 19) were contacted using a custom pogo-pin PCB adapter (Figure 20). A custom interface PCB connected the pogo-pin adapter to voltage generators and measurement equipment.

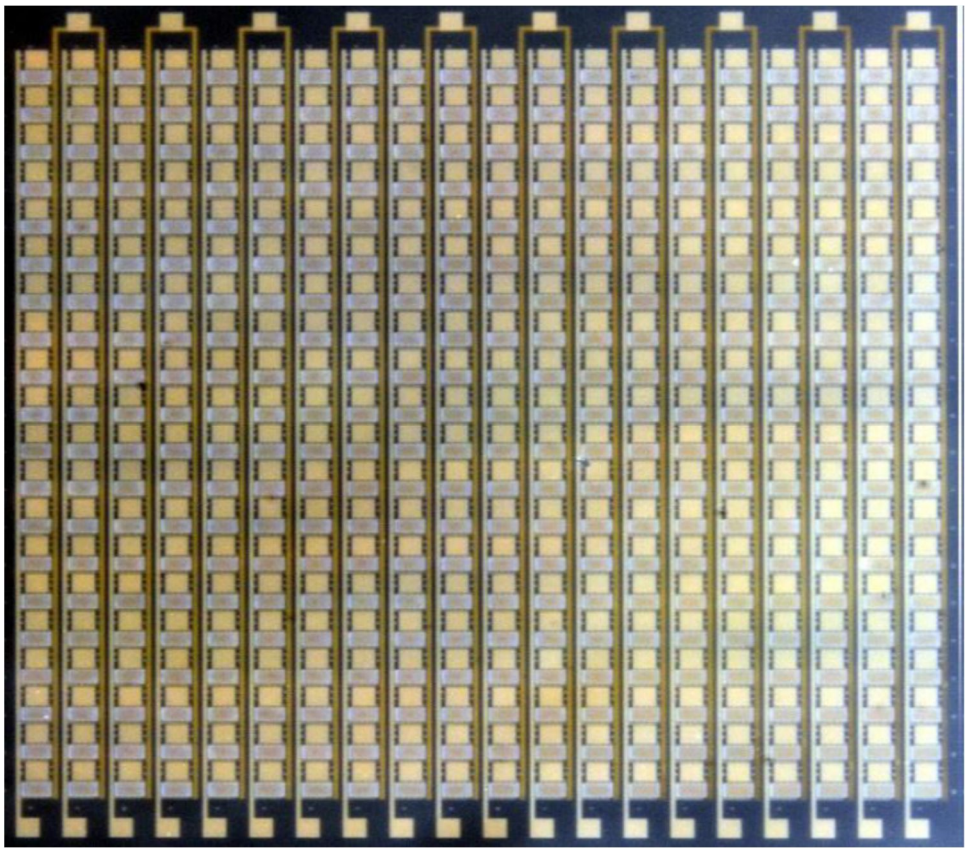

**Figure 19.** Light microscope image of synstor chip showing the 20 mm by 22 mm area device array containing 400 synstors. The total chip dimensions are 30 by 30 mm.

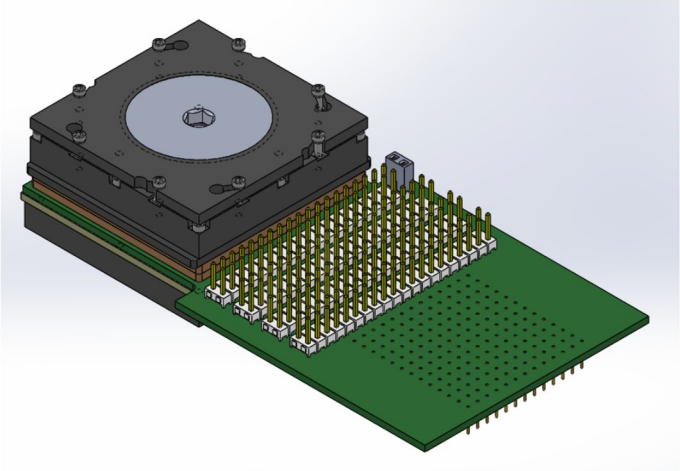

**Figure 20.** Custom pogo-pin adapter PCB for connecting synstor chip to testing and neuron circuit PCB.

## 2.5. Device Modeling

### 2.5.1. Convolutional Signal Processing

A voltage pulse with an amplitude of $V_a$ and a duration of $t_d$ applied on the input electrode of a synstor via a diode at the moment $t = t_n$ charges the capacitor between the CNT network and Al reference electrode, $c_{CNT/Al}$, and induces a change of the current through the CNT network toward the grounded output electrode of the synstor, $\Delta I(t) \approx w\, V_p\left[1 - e^{-\beta_p(t-t_n)}\right]$, where $w$ denotes the DC conductance of the synstor, and $\beta_p$ denotes a parameter related to $c_{CNT/Al}$ and the resistance to charge the CNT network. After the pulse ends at $t = t_n + t_d$, the capacitor is gradually discharged with a current flowing mainly toward the output electrode, $I(t) \approx I(t_n + t_d)e^{-\beta_d(t-t_n-t_d)}$, where $\beta_d$ denotes a parameter related to $c_{CNT/Al}$ and the resistance to discharge the CNT network. A series of voltage pulses trigger a current from the synstor, $I(t) =$

$(w\ \kappa) \circledast\ V_i$, where $\kappa \circledast V_i$ represents the temporal convolution between $\kappa(t)$ and $V_i(t)$, $V_i(t) = \sum_n V_a \delta(t - t_n)$, and the kernel function,

$$\kappa(t) = \begin{cases} 1 - e^{-\beta_p t} & when\ t \le t_d \\ (1 - e^{-\beta_p t_d}) e^{-\beta_d t} & when\ t > t_d \end{cases} \quad 3$$

The currents triggered by periodic pulses with an amplitude $V_a = -1.75\ V$, a period of $30\ ns$, and different durations ($8 - 20\ ns$) were amplified by an operational amplifier (Texas Instruments, LMC6482) and measured by an oscilloscope (Tektronix TDS 3054B) versus time (Figure 14 and Figure 15). The currents were fitted by Equations 1 and 3 with $\beta_d = 41.5\ MHz$, $\beta_p = 0.73\ MHz$ when $t_d = 8\ ns$; $\beta_p = 0.67\ MHz$ when $t_d = 10\ ns$; $\beta_p = 0.47\ MHz$ when $t_d = 15\ ns$; $\beta_p = 0.36\ MHz$ when $t_d = 20\ ns$. When the pulse duration increased, $\beta_p$ decreased. The CNT resistor connected with the input electrode to charge the capacitor has a much larger Schottky barrier and resistance than the CNT resistor connected with the output electrode to discharge the capacitor, leading to longer charging than discharging times, thus $\beta_p \ll \beta_d$. $\beta_p$ and $\beta_d$ could be modified by adjusting the resistance and capacitance values.

## 2.5.2. Physical Model of Charge and Conductance Modification

$V_i$ and $V_o$ voltages with an amplitude $V_i = V_o$ applied on the input and output electrodes of a synstor induce a voltage, $V_{CNT/TiO2}$, on the capacitor between the CNT network and $TiO_2$ charge-storage layer, which in turn modify the charge density, $\rho_s$, in the charge storage layer by electronic hopping through the $HfO_2$ layer[39] (Figure 6 and Figure 7). The $\rho_s$ modification rate is equal to the current density of electronic hopping through the $HfO_2$ layer,[40] $d\rho_s/dt =$

$\mathcal{J}$ exp $[\frac{q(\theta\sqrt{|V_{CNT/TiO2}|}-\phi_B)}{k_BT}]$, where $q$ denotes the charge of an electron, $k_B$ denotes the Boltzmann constant, $T$ denotes temperature, $\phi_B$ denotes the potential barrier for electrons to diffuse in the $HfO_2$ barrier layer, $\theta$ denotes a parameter related to the thickness of the $HfO_2$ layer, and $\mathcal{J}$ represents a parameter equal to current density under $|V_{CNT/TiO2}| = (\phi_B/\theta)^2$. The CNT/$HfO_2$/$TiO_2$/$HfO_2$/Al layers in a synstor are composed of two capacitors connected in series with the CNT/$HfO_2$/$TiO_2$ and $TiO_2$/$HfO_2$/Al sandwich structures and corresponding capacitance $c_{CNT/TiO2}$and $c_{TiO2/Al}$, and $V_{CNT/TiO2} = V_i/\nu - \rho_s/(\nu c_{TiO2/Al})$ with $\nu = (c_{CNT/TiO2} + c_{TiO2/Al})/c_{TiO2/Al}$. After substituting $V_{CNT/TiO2}$ in $d\rho_s/dt$, $d\rho_s/dt =$ $\mathcal{J}$exp $[\frac{q\theta(\sqrt{V_i-\rho_s/c_{TiO2/Al}}-\sqrt{V_t^+-\rho_s^0/c_{TiO2/Al}}}{\sqrt{\nu}k_BT}]$ under $V_i = V_o > V_t^+ > 0$; $d\rho_s/dt =$ $\mathcal{J}$exp $[\frac{q\theta(\sqrt{-V_i+\rho_s/c_{TiO2/Al}}-\sqrt{|V_t^--\rho_s^0/c_{TiO2/Al}|})}{\sqrt{\nu}k_BT}]$ under $V_i = V_o < V_a^{t-} < 0$, where $V_t^+ = \nu(\phi_B/\theta)^2 + \rho_s^0/c_{TiO2/Al} > 0$, $V_t^- = -\nu(\phi_B/\theta)^2 + \rho_s^0/c_{TiO2/Al} < 0$, and $\rho_s^0$ as $\rho_s$ before the voltage $V_a$ is applied. When $V_t^- < V_i = V_o < V_t^+$, the external voltage $V_i$ drives an insignificant amount of electrons to overcome the potential barrier in the $HfO_2$ layer, and $|d\rho_s/dt| < |\mathcal{J}| \approx 0$. When $V_i = V_o > V_t^+$ or $< V_t^-$, the external potential drives electrons through the potential barrier to modify $\rho_s$, and $\rho_s$ also gradually builds up an internal potential against the external potential. When $\rho_s$ is modified to balance the external potential with $V_i - \rho_s/c_{TiO2/Al} \approx V_t^+ - \rho_s^0/c_{TiO2/Al}$ or $V_i - \rho_s/c_{TiO2/Al} \approx V_t^- - \rho_s^0/c_{TiO2/Al}$, $d\rho_s/dt = \mathcal{J} \approx 0$, and $\rho_s$ reaches its saturation values with $\rho_s \approx \rho_s^0 + c_{TiO2/Al}[V_i - V_t^+]$ or $\rho_s \approx \rho_s^0 + c_{TiO2/Al}[V_i - V_t^-]$. In the capacitance-voltage test, after a synstor experiences multiple voltage pulses with $V_i = V_o$ on its input and output electrodes, the modification of the charge densities in the synstor (Figure 12b),

$$\Delta\rho_s \approx \begin{cases} c_{TiO2/Al}[V_i - V_t^+] & when\ V_i = V_o > V_t^+ > 0 \\ 0 & when\ V_a^{t-} < V_i = V_o < V_t^+ \\ c_{TiO2/Al}[V_i - V_t^-] & when\ V_i = V_o < V_t^- < 0 \end{cases} \quad 4$$

where $\Delta\rho_s = \rho_s - \rho_s^0$.

Multiple $V_i$ and $V_o$ voltage pulses with an amplitude $V_i = V_o = V_a$ applied on the input and output electrodes of a synstor also modify $\rho_s$, and the change of $\rho_s$, $\Delta\rho_s$ induces the change of the Fermi level of the CNT network, which in turn causes the change of device conductance[40] (Figure 11b), $w \approx w_0 \exp\left(\frac{q\Delta V_{CNT}^f}{k_B T}\right)$, and $\Delta w = w - w_0 = w_0[\exp\left(\frac{q\Delta V_{CNT}^f}{k_B T}\right) - 1]$, where $w_0$ denotes the initial device DC conductance before the charge modification, and $q\Delta V_{CNT}^f$ denotes the change of the CNT Fermi level induced by $\Delta\rho_s$. $\Delta V_{CNT}^f$ increases monotonically with $-\Delta\rho_s$ in the p-type CNTs, and its linear approximation gives $\Delta V_{CNT}^f \approx -\varepsilon_\rho^+ \Delta\rho_s$ when $\Delta\rho_s > 0$ and $\Delta V_{CNT}^f \approx -\varepsilon_\rho^- \Delta\rho_s$ when $\Delta\rho_s < 0$, where $\varepsilon_\rho^+$ and $\varepsilon_\rho^-$ denote constants related to the device structure, capacitance, CNT doping concentration. After substituting $\Delta V_{CNT}^f$ in $\Delta w$ by $\Delta\rho_s$ in Equation 3,

$$\frac{\Delta w}{w_0} \approx \begin{cases} e^{-\beta_v^+(V_i - V_t^+)} - 1 & when\ V_i = V_o \geq V_t^+ > 0 \\ 0 & when\ V_a^{t-} < V_i = V_o < V_t^+ \\ e^{-\beta_v^-(V_i - V_t^-)} - 1 & when\ V_i = V_o \leq V_t^- < 0 \end{cases} \quad 5$$

where $\beta_v^+ = q\varepsilon_\rho^+ c_{TiO2/Al}/k_B T$ and $\beta_v^- = q\varepsilon_\rho^- c_{TiO2/Al}/k_B T$.

When a series of $V_i$ and $V_o$ voltage pulses with $V_i = V_o$ are applied, $\rho_s$ is modified by the pulses as a function of the number of the applied pulses, $n$ (Figure 11a). The modification rate, $\frac{d\rho_s}{dn} = t_d \frac{d\rho_s}{dt} = \mathcal{J} t_d \exp\left[\frac{q(\theta\sqrt{|V_i - \rho_s/c_{TiO2/Al}|} - \sqrt{\nu}\phi_B)}{\sqrt{\nu} k_B T}\right]$ with $t_d$ as the pulse duration. The solution of

the differential equation gives $\rho_s(n) \approx \eta^+ Ln\left(\frac{n}{n_0^+}+1\right)+\rho_s^0$ with $\eta^+ = 2\sqrt{\nu}kTc_{TiO2/Al}\sqrt{V_i-\rho_s^0/c_{TiO2/Al}}/(q\theta)$, and $n_0^+ = \eta^+ \exp\left[\frac{q\phi_B}{k_BT}\right]/(\mathcal{J}t_d)$ under $V_a \geq V_t^+ > 0$. $\rho_s(n) \approx -\eta^- Ln\left(\frac{n}{n_0^-}+1\right)+\rho_s^0$ with $\eta^- = 2\sqrt{\nu}kTc_{TiO2/Al}\sqrt{-V_i+\rho_s^0/c_{TiO2/Al}}/(q\theta)$ and $n_0^- = \eta^- \exp\left[\frac{q\phi_B}{k_BT}\right]/(\mathcal{J}t_d)$ under $V_a \leq V_t^- < 0$. $\Delta\rho_s(n) = \rho_s(n) - \rho_s^0$ induces the shift of the CNT Fermi energy $q\Delta V_{CNT}^f(n)$, which in turn induces $\Delta w(n) = w_0\left[\exp\left(\frac{q\Delta V_{CNT}^f(n)}{k_BT}\right)-1\right] \approx \frac{w_0 q\Delta V_{CNT}^f(n)}{k_BT} = \frac{w_0 q\varepsilon_\rho^\pm}{k_BT}\Delta\rho_s(n) = \frac{w_0\eta^\pm q\varepsilon_\rho^\pm}{k_BT}Ln\left(\frac{n}{n_0^\pm}+1\right)$. Therefore,

$$\frac{\Delta w(n)}{w_0} \approx \begin{cases} \kappa^+ Ln\left(\frac{n}{n_0^+}+1\right) & when\ V_i = V_o \geq V_t^+ > 0 \\ 0 & when\ V_a^{t-} < V_i = V_o < V_t^+ \\ \kappa^- Ln\left(\frac{n}{n_0^-}+1\right) & when\ V_i = V_o \leq V_t^- < 0 \end{cases} \quad (6)$$

where $\kappa^+ = \frac{q\eta^+\varepsilon_\rho^+}{k_BT}$ and $\kappa^- = \frac{q\eta^-\varepsilon_\rho^-}{k_BT}$.

## 2.6. Au Electrode Control Device

A control device was fabricated with a structure similar to the synstor, but in which the Al input and output electrodes were replaced by Au ones. An Ohmic contact was formed between the Au electrode and the p-type CNTs, and a $V_i$ or $V_o$ voltage alone significantly modified the conductance of the control device (Figure 21), which confirms that the large Al/CNT Schottky barriers are necessary to keep $w$ nonvolatile under $V_i \cdot V_o = 0$.

Figure 21a displays a scheme showing the cross-sectional structure of the control device with a carbon nanotube (CNT) network (orange) connected by a Au input electrode (golden) and a Au output electrode (golden), which replaces Al input and output electrodes in a synstor. A titanium oxide ($TiO_2$) charge storage layer (purple) is embedded in a hafnium oxide ($HfO_2$) dielectric layer sandwiched between the CNT network and an Al reference electrode (grey), which is electrically grounded. $V_i$ and $V_o$ voltages are applied on the input and output electrodes, respectively. The L scale represent the lateral distance from the input electrode to the output electrode. In Figure 21b, $\Delta w/w_0$, the percentage changes of DC conductance $w$ of the control device, are plotted versus the amplitudes of fifty paired 5 ms-wide $V_i$ and $V_o$ pulses with $V_i = V_o$ (filled triangles), $V_i$ pulses with $V_o = 0$ (open triangles), and $V_o$ pulses with $V_i = 0$ (opened squares). When $V_i = V_o > 0.65\ V$, $w$ was decreased (filled red triangles); when $V_i = V_o < -1.0\ V$, $w$ was increased (filled green triangles). When $V_i > 1.25\ V$ and $V_o = 0$, $w$ was decreased. When $V_i < -1.23\ V$ and $V_o = 0$, $w$ was increased. In Figure 21c, simulated electronic energy-band diagrams of the control device are plotted along the Au input electrode, the CNT network, and the Au output electrode under the condition $V_i = -1.75\ V$ (green), and $V_i = 1.75\ V$ (red) with $V_o = 0$. $E_f^i$, $E_f^o$, $E_f^r$, and $E_f^{TiO_2}$ denote the Fermi energies of the Al input, output, reference electrodes, and the $TiO_2$ charge storage layer, respectively. The Fermi energy differences between $E_f^i$ and $E_f^r$ are marked. In Figure 21d, the current, $I$, through the control device is plotted versus the $V_i$ voltage after applying multiple pairs of $V_i$ and $V_o$ pulses with $V_i = V_o = 2\ V$ (red line) or $V_i = V_o = -2\ V$ (green line) on the control device. The linear $I-V_i$ curves measured from the control device and the simulated electronic energy-band diagrams indicate the formation of the Ohmic contact between the Au electrodes and the p-type semiconducting CNTs[38]. Without the Schottky barrier between the Au electrodes and the CNTs,

the $V_i$ or $V_o$ voltages alone generate a large Fermi energy difference ($\gtrsim 1\ eV$) between the CNTs and the $TiO_2$ to drive electrons through the $HfO_2$ dielectric layer to modify the charge stored in the $TiO_2$ layer, which in turn modifies $w$.

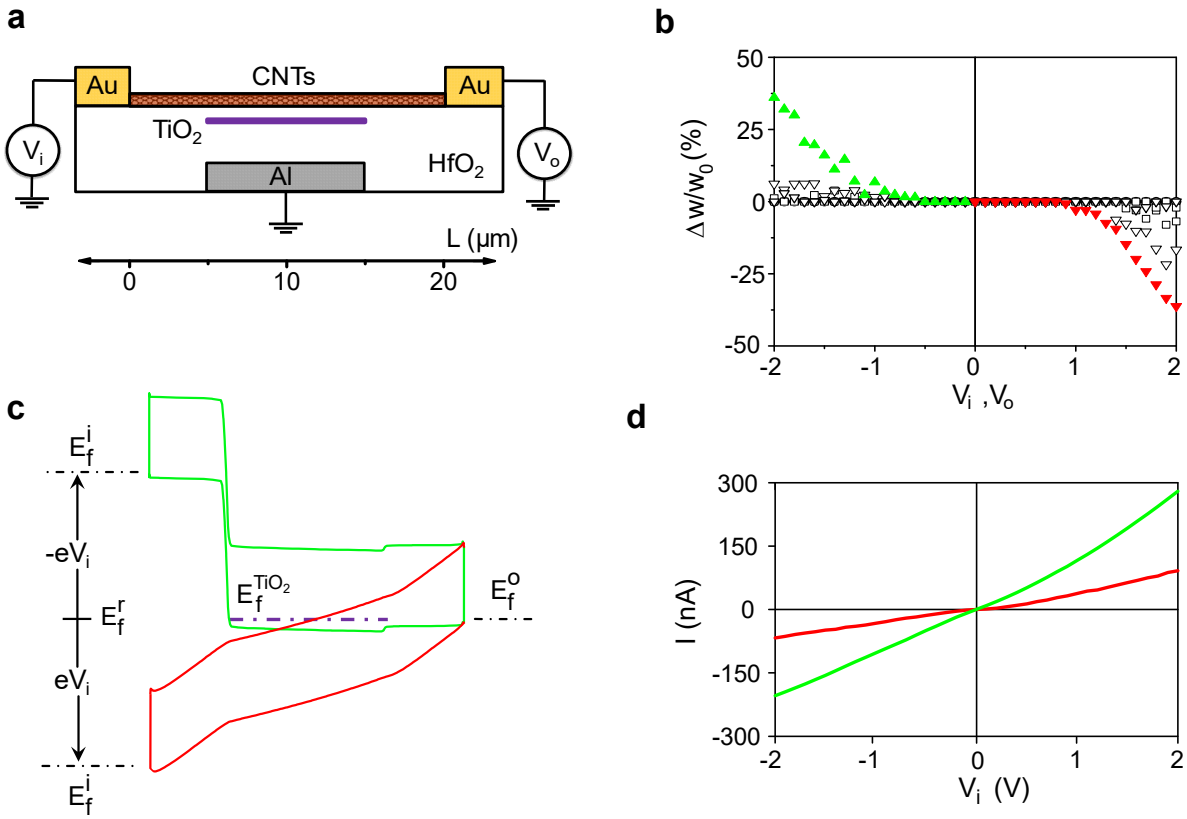


**Figure 21.** Au electrode structure and device properties. a) Cross-section diagram of Au electrode device. b) $\Delta w/w_0$ versus voltage amplitudes. c) Simulated electronic energy-band diagrams of the control device. d) The current, $I$ versus the $V_i$ voltage after applying pairs of $V_i$ and $V_o$ pulses

## 2.7. Simulations of 200 nm, 40 nm, and 20 nm Wide Synstors

Nanoscale synstors have been simulated by Technology Computer-Aided Design (TCAD) simulator (Sentaurus Device, Synopsys). The cross-sectional structure of a simulated synstor

composed of a 200 nm-wide p-type semiconducting CNT network formed contacts with an Al input and an Al output electrodes is shown in **Figure 22**a. The CNT network is on a 2.5 nm-thick $HfO_2$ dielectric layer on a 1.0 nm-thick $TiO_2$ charge storage layer on a 4.5 nm-thick $HfO_2$ dielectric layer on an Al reference electrode. There is a 50 nm lateral space between the Al reference electrode/$TiO_2$ charge storage layer and the Al input and output electrodes, and the $TiO_2$ charge storage layer is also laterally recessed with respect to the Al reference electrode by 20 nm at both ends.

The simulated electronic band diagrams along the Al input electrode/CNT/Al output electrode under various $V_i$, the potential on the input electrode, and $V_o$, the potential on the output electrode are shown in **Figure 22**b-f. When a pair of negative $V_i$ and $V_o$ pulses with $V_i = V_o = -1.75\ V$ are applied on the input and output electrodes of a synstor simultaneously, the negative potential on the CNT network with respect to the Al reference electrode inverts the p-type CNTs above the Al reference electrode to n-type CNTs, and moves the edge of the CNT conduction band close to the Fermi level of the Al input/output electrodes (**Figure 22**b). The capacitance between a single CNT in the network and the continuous $TiO_2$ charge-storage layer is much smaller than the capacitance between the $TiO_2$ layer and the Al reference electrode. Therefore, the negative potential between the CNT network with respect to the Al reference electrode predominantly drops across the CNT network and the $TiO_2$ charge-storage layer, which induces an average difference of $1.51\ eV$ between the CNT and $TiO_2$ Fermi energies (**Figure 23**a), and drives electrons from the CNT network into the $TiO_2$ charge-storage layer through the 2.5 nm-thick $HfO_2$ dielectric layer to increase the negative charge in the $TiO_2$ layer. Based on the TCAD simulation, the negative $V_i$ and $V_o$ potentials with $V_i = V_o = -1.75\ V$ modify the charge in the $TiO_2$ layer to a density of $-6.86\ \mu C/\mathrm{cm}^2$ (**Figure 23**b), which attracts holes in the CNT network, increasing the device

conductance, $w$. When a pair of positive $V_i$ and $V_o$ pulses with $V_i = V_o = 1.75\ V$ are applied on the input and output electrodes of a synstor simultaneously, the positive potential on the CNT network with respect to the Al reference electrode induces hole accumulation in the p-type CNTs above the Al reference electrode (**Figure 22**c). The positive potential between the CNT network with respect to the Al reference electrode predominantly drops across the CNT network and the $TiO_2$ charge storage layer, which induces a difference of $-1.51\ eV$ between the CNT and $TiO_2$ Fermi energies (**Figure 23**a), and drives electrons from the $TiO_2$ charge storage layer to the CNT network through the 2.5 nm-thick $HfO_2$ dielectric layer (**Figure 22**c). Based on the TCAD simulation, the positive $V_i$ and $V_o$ potentials with $V_i = V_o = 1.75\ V$ modify the charge in the $TiO_2$ layer to a density of $5.14\ \mu C/\mathrm{cm}^2$ (**Figure 23**b), which attracts holes in the CNT network, increasing the device conductance, $w$.

Electronic band diagrams were also simulated by TCAD when a pair of $V_i$ and $V_o$ pulses are applied on the input and output electrodes of the synstor with the 200 nm-wide CNT network under the different conditions: (1) $V_i = -1.75\ V$ and $V_o = 1.75\ V$ (**Figure 22**d); (2) $V_i = 1.75\ V$ and $V_o = -1.75\ V$ (not shown, but symmetric to **Figure 22**d); (3) $V_i = 1.75\ V$ and $V_o = 0$ (**Figure 22**e); (4) $V_i = 0\ V$ and $V_o = 1.75\ V$ (not shown, but symmetric to **Figure 22**e); (5) $V_i = -1.75\ V$ and $V_o = 0$ (**Figure 22**f); (6) $V_i = 0\ V$ and $V_o = -1.75\ V$ (not shown, but symmetric to **Figure 22**f). Under the $V_i \cdot V_o \leq 0$ conditions, the $V_i$ and/or $V_o$ potentials mainly drops across the hole-depletion region on the lateral space beyond the $TiO_2$ charge storage layer (**Figure 22**d-f), which leads to small Fermi energy differences ($\lesssim 0.15\ eV$) between the CNTs and the recessed $TiO_2$ layer across the $HfO_2$ dielectric layer (**Figure 23**a). Based on the TCAD simulation, the small Fermi energy differences cannot drive electrons to tunnel through the 2.5 nm-thick $HfO_2$ dielectric layer (**Figure 23**b). Thus, the charge stored in the $TiO_2$ layer cannot be modified by $V_i$ and $V_o$

pulses with $V_i \cdot V_o \leq 0$, and the synstor conductance, $w$, remains unchanged. Based on the TCAD simulation, when the synstor is scaled down to 200 nm, it still functions properly for inference and learning.

The cross-sectional structure of a simulated synstor composed of a 40 nm-wide p-type semiconducting CNT network formed contacts with an Al input and an Al output electrode is shown in **Figure 22**g. The CNT network is on a 1.5 nm-thick $HfO_2$ dielectric layer on a 0.5 nm-thick $TiO_2$ charge storage layer on a 2.5 nm-thick $HfO_2$ dielectric layer on an Al reference electrode. There is a 10 nm lateral space between the Al reference electrode and the Al input and output electrodes, and the $TiO_2$ charge storage layer is also laterally recessed with respect to the Al reference electrode by 2.5 nm at both ends (**Figure 22**g).

The simulated electronic band diagrams along the Al input electrode/CNT/Al output electrode under various $V_i$, the potential on the input electrode, and $V_o$, the potential on the output electrode are shown in **Figure 22**h-l. In the synstor learning mode, when a pair of negative $V_i$ and $V_o$ pulses with $V_i = V_o = -0.5\,V$ are applied on the input and output electrodes of a synstor simultaneously, the negative potential on the CNT network with respect to the Al reference electrode inverts the p-type CNTs above the Al reference electrode to n-type CNTs, and moves the edge of the CNT conduction band close to the Fermi level of the Al input/output electrodes (**Figure 22**h). The negative potential between the CNT network with respect to the Al reference electrode predominantly drops across the CNT network and the $TiO_2$ charge-storage layer, which induces a difference of $0.43\,eV$ between the CNT and $TiO_2$ Fermi energies (**Figure 23**c), and drives electrons from the CNT network into the $TiO_2$ charge-storage layer through the 1.5 nm-thick $HfO_2$ layer (**Figure 23**d). Based on the TCAD simulation, the negative $V_i$ and $V_o$ potentials with $V_i = V_o = -0.5\,V$ modify the charge in the $TiO_2$ layer to a density of $-3.06\,\mu C/\mathrm{cm}^2$ (**Figure**

**23**b), which attracts the holes in the CNT network above the $TiO_2$ layer, increasing the device conductance, $w$. When a pair of positive $V_i$ and $V_o$ pulses with $V_i = V_o = 0.5\ V$ are applied on the input and output electrodes of a synstor simultaneously, the positive potential on the CNT network with respect to the Al reference electrode induce hole accumulation in the p-type CNTs above the Al reference electrode (**Figure 22**i). The positive potential between the CNT network with respect to the Al reference electrode predominantly drops across the CNT network and the $TiO_2$ charge storage layer, which induces a difference of $-0.43\ eV$ between the CNT and $TiO_2$ Fermi energies (**Figure 23**c), and drives electrons from the $TiO_2$ charge storage layer to the CNT network through the 1.5 nm-thick $HfO_2$ dielectric layer (**Figure 22**i). Based on the TCAD simulation, the positive $V_i$ and $V_o$ potentials with $V_i = V_o = 0.5\ V$ modify the charge in the $TiO_2$ layer to a density of $2.93\ \mu C/\mathrm{cm}^2$ (**Figure 23**d), which repels the holes in the CNT network above the $TiO_2$ layer, decreasing the device conductance, $w$.

Electronic band diagrams were also simulated by TCAD when a pair of $V_i$ and $V_o$ pulses are applied on the input and output electrodes of the synstor with the 40 nm-wide CNT network under different conditions: (1) $V_i = -0.5\ V$ and $V_o = 0.5\ V$ (**Figure 22**j); (2) $V_i = 0.5\ V$ and $V_o = -0.5\ V$ (not shown, but symmetric to **Figure 22**j); (3) $V_i = 0.5\ V$ and $V_o = 0$ (**Figure 22**k); (4) $V_i = 0\ V$ and $V_o = 0.5\ V$ (not shown, but symmetric to **Figure 22**k); (5) $V_i = -0.5\ V$ and $V_o = 0$ (**Figure 22**l); (6) $V_i = 0\ V$ and $V_o = -0.5\ V$ (not shown, but symmetric to **Figure 22**l). Under the $V_i \cdot V_o \leq 0$ conditions, the $V_i$ and/or $V_o$ potentials mainly drops across the hole-depletion region on the lateral space beyond the $TiO_2$ charge storage layer (**Figure 22**j-l), which leads to small Fermi energy differences ($\lesssim 0.11\ eV$) between the CNTs and the recessed $TiO_2$ layer across the $HfO_2$ dielectric layer (**Figure 23**c). The small Fermi energy differences can not drive electrons to tunnel through the 1.5 nm-thick $HfO_2$ dielectric layer, causing ignorable changes of the charges in the

TiO2 layer (**Figure 23**d). Since the charge stored in the $TiO_2$ layer cannot be modified by $V_i$ and $V_o$ pulses with $V_i \cdot V_o \leq 0$, the synstor conductance, $w$, remains unchanged. Based on the TCAD simulation, when the synstor is scaled down to 40 nm, it still functions properly for concurrent inference and learning.

We have also simulated a synstor with a 20 nm-wide CNT network. In its learning mode, when a pair of $V_i$ and $V_o$ pulses with $V_i = V_o = -0.5\ V$ (or $V_i = V_o = 0.5\ V$) are applied on the input and output electrodes of the synstor simultaneously, the potentials induce a difference of $0.43\ eV$ between the CNT and $TiO_2$ Fermi energies (the band diagrams are not shown here), and drive electrons through the 1.5 nm-thick $HfO_2$ dielectric layer sandwiched between the $TiO_2$ charge storage layer and the CNT network to modify the charge in the $TiO_2$ charge storage layer, which in turn modifies the hole concentrations in the CNT network and the device conductance, $w$. Electronic band diagrams were also simulated by TCAD (not shown here) when a pair of $V_i$ and $V_o$ pulses are applied on the input and output electrodes of the synstor with the 20 nm-wide CNT network under the different conditions: (1) $V_i = -0.5\ V$ and $V_o = 0.5\ V$; (2) $V_i = 0.5\ V$ and $V_o = -0.5\ V$; (3) $V_i = 0.5\ V$ and $V_o = 0$; (4) $V_i = 0\ V$ and $V_o = 0.5\ V$; (5) $V_i = -0.5\ V$ and $V_o = 0$; (6) $V_i = 0\ V$ and $V_o = -0.5\ V$. Under the $V_i \cdot V_o \leq 0$ condition, the currents driven by the $V_i$ and/or $V_o$ potentials through the CNT network, and the potentials across the CNT network above the $TiO_2$ layer in the 20 nm synstor are significantly larger than those in the 40 nm synstor, which leads to a maximal Fermi energy difference of $\sim 0.18\ eV$ between the CNTs and the $TiO_2$ layer. The maximal Fermi energy differences can drive electrons to tunnel through the 1.5 nm-thick $HfO_2$ dielectric layer (**Figure 23**d), causing significant changes of the charges in the $TiO_2$ layer and the synstor conductance, $w$, even in its signal processing mode. Based on the TCAD simulation, when the synstor is scaled down to 20 nm, it loses its non-volatile memory during signal inference. The

structure and materials of sub-20 nm synstors need to be modified for concurrent interference and learning.

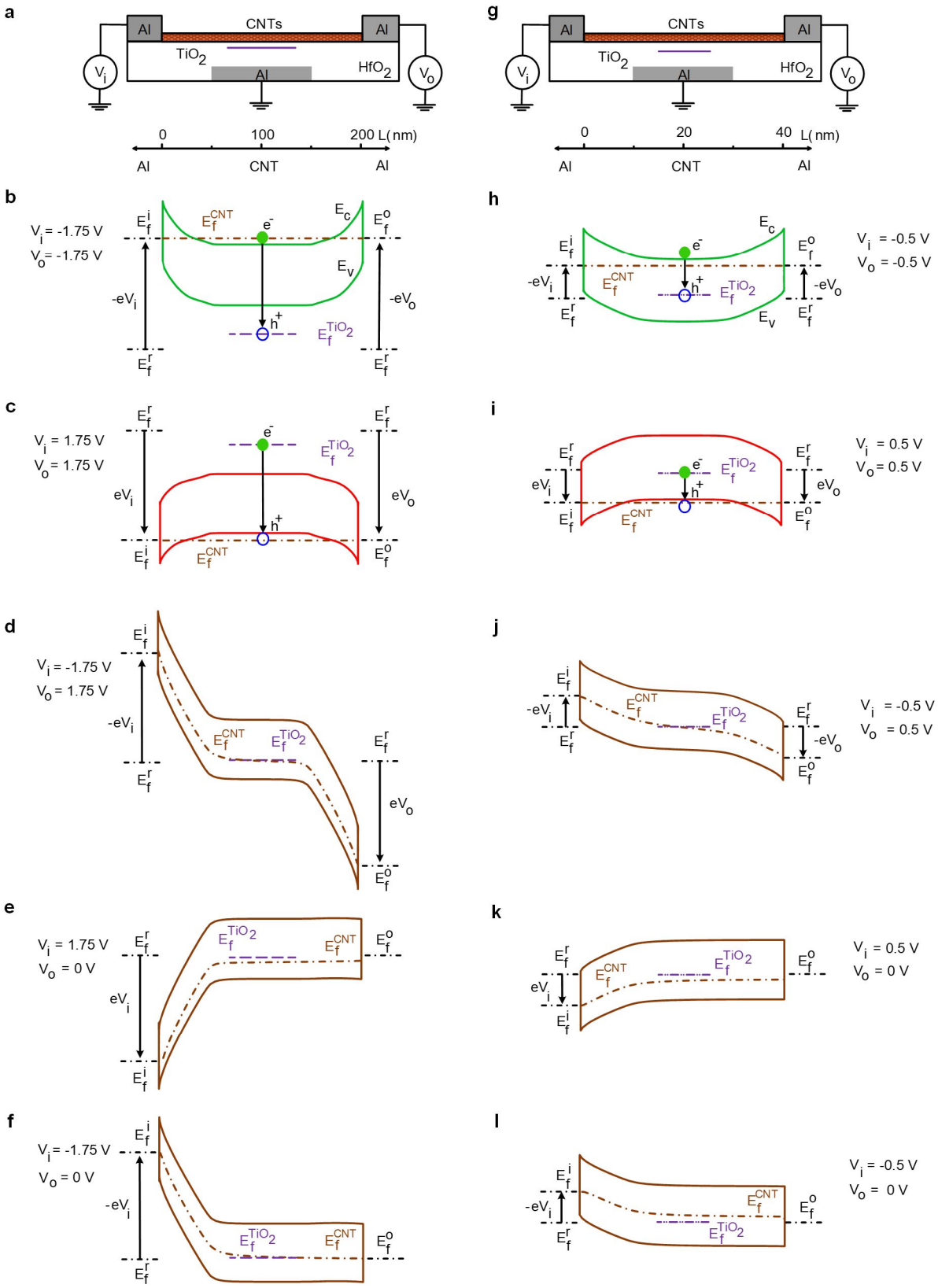
a
Al
CNTs
Al
TiO2
Al
HfO2
Vi
Vo
0
100
200 L(nm)
Al
CNT
Al
g
Al
CNTs
Al
TiO2
Al
HfO2
Vi
Vo
0
20
40 L(nm)
Al
CNT
Al
b
Vi = -1.75 V
Vo = -1.75 V
Ec
Ev
e-
h+
-eVi
-eVo
h
Vi = -0.5 V
Vo = -0.5 V
c
Vi = 1.75 V
Vo = 1.75 V
eVi
eVo
i
Vi = 0.5 V
Vo = 0.5 V
d
Vi = -1.75 V
Vo = 1.75 V
j
Vi = -0.5 V
Vo = 0.5 V
e
Vi = 1.75 V
Vo = 0 V
k
Vi = 0.5 V
Vo = 0 V
f
Vi = -1.75 V
Vo = 0 V
l
Vi = -0.5 V
Vo = 0 V

**Figure 22**. Cross-sections and band diagrams of simulated nanoscale synstors. a) Cross-section of 200 nm-wide synstor. Band diagrams under the conditions b) $V_i = V_o = -1.75\ V$ (green), c) $V_i = V_o = 1.75\ V$ (red), d) $V_i = -1.75\ V$ and $V_o = 1.75\ V$, e) $V_i = 1.75\ V$ and $V_o = 0$, and f) $V_i = -1.75\ V$ and $V_o = 0$. g) Cross-section of 200 nm-wide synstor. Band diagrams under the conditions h) $V_i = V_o = -0.5\ V$ (green), i) $V_i = V_o = 0.5\ V$ (red), j) $V_i = -0.5\ V$ and $V_o = 0.5\ V$, k) $V_i = 0.5\ V$ and $V_o = 0$, and l) $V_i = -0.5\ V$ and $V_o = 0$.

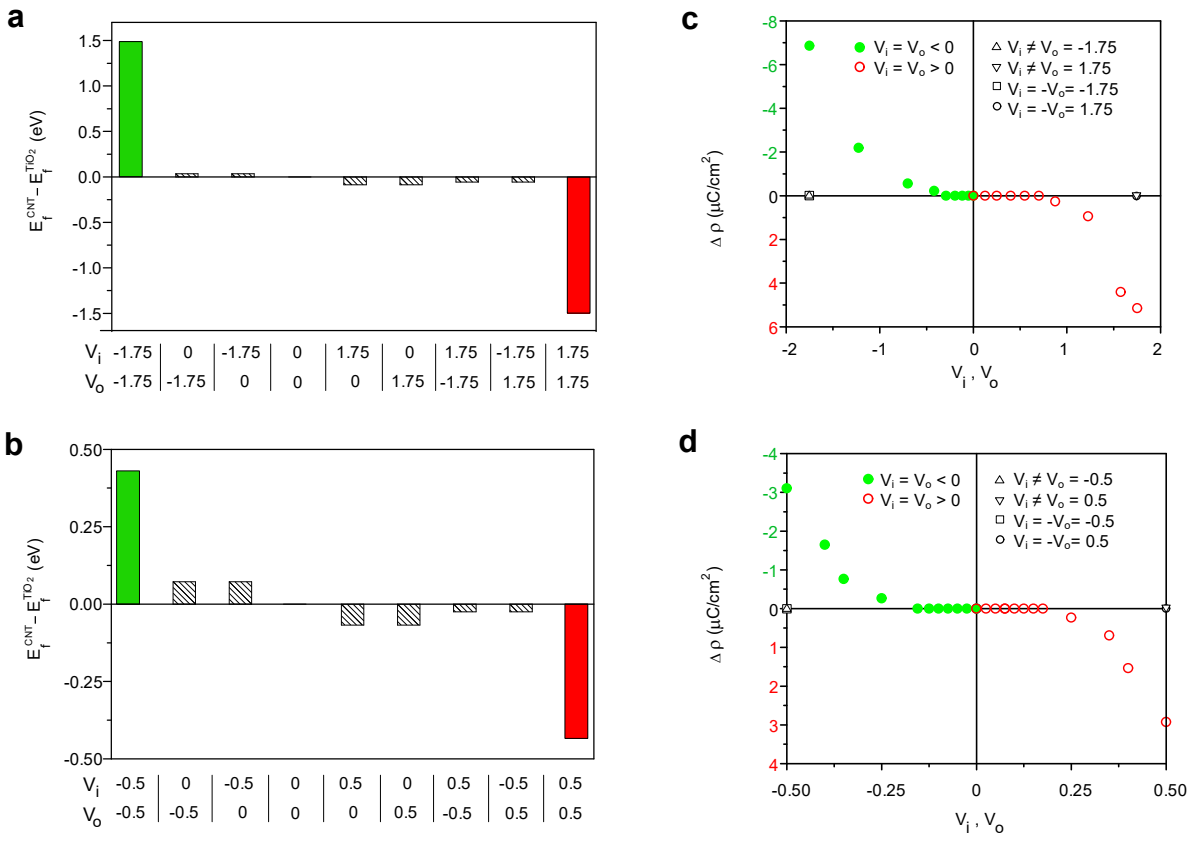


**Figure 23.** The simulated energy differences between the CNTs and$TiO_2$ charge storage layer in a synstor with a) a 200 nm-wide and b) a 40 nm-wide synstor. The simulated changes of net charge density, $\Delta\rho$, in the $TiO_2$ charge storage layer with c) a 200 nm-wide and d) a 40 nm-wide synstor.

# 3. Self-Programming Neuromorphic Circuit (SNIC)

## 3.1. Integrate-and-Fire Neuron Circuit

We designed and fabricated an integrate-and-fire circuit with the basic functions according to the Hodgkin–Huxley neuron model.[43] The "neuron" circuit is shown in Figure 24a. The collective current, $I$, from multiple synstors flows through a diode toward a capacitor $C_{IF}$, increasing the voltage, $V_C$, on the capacitor. A negative voltage, $V_L$, is applied to induce a leakage current, $I_L$, flowing through the resistor, $R_L$, to filter the thermodynamic and signal noises in the circuits. $V_C$ is proportional to the integration of $I - I_L$ with respect to time. When $V_C$ reaches a threshold value, a Schmitt trigger composed of transistors M1-M6 is switched back and forth to generate an output pulse from the output channel, $V_f$. The output pulse resets $V_C$ back to zero by switching transistors M7, M8, and M9, and the capacitor $C_{IF}$ restarts the integration of the current.

A "neuron" circuit with $C_{IF} = 1$ nF, $R_L = 50\ M\Omega$, $V_L = -0.25\ V$, and $R_{INV} = 0.5\ M\Omega$ was tested by applying a series of 10 ns-wide pulses with varied firing rates over a resistor to inject a current $I$ to $C_{IF}$, and the average firing rate of the output pulses triggered from the circuit, $r_f$, is plotted as a function of the average magnitude of the current, $\bar{I}$, in Figure 24b. When $\bar{I} <$ 5.3 nA, no pulse is output from the "neuron" circuit. When $5.3\ nA < \bar{I} < 24.3\ nA$, $r_f$ increases monotonically with increasing $\bar{I}$. When $\bar{I} > 24.3\ nA$, $r_f$ is saturated at $\sim 142\ Hz$. $r_f$ was fitted as a nonlinear sigmoid function of $\bar{I}$,

$$r_f = \frac{r_s}{1 + e^{-\chi[\bar{I} - I_0]}} \qquad 7$$

where $r_s = 152\ Hz$, $\chi = 0.26\ /\text{nA}$, and $I_0 = 16.0$ nA. The transistors in the circuit are operated in their subthreshold regions.

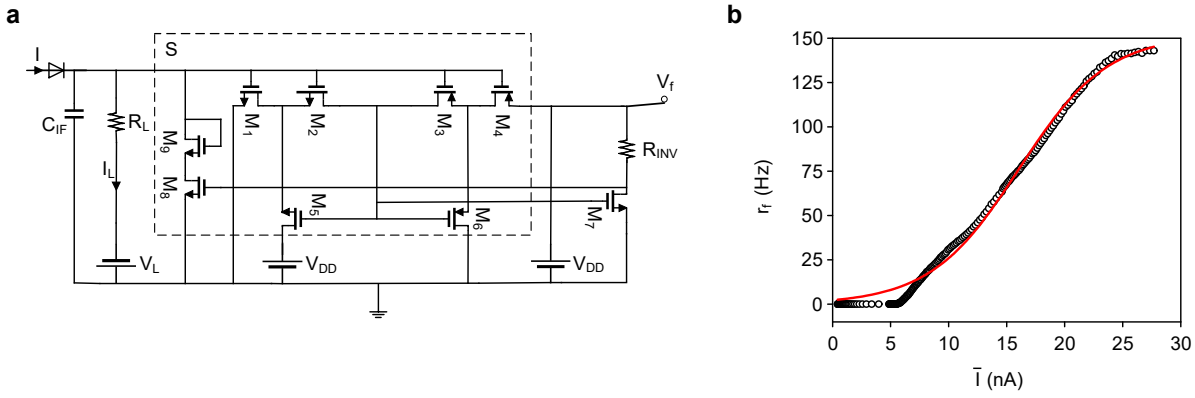


**Figure 24.** Integrate-and-fire circuit. a) The integrate-and-fire "neuron" circuit consists of a capacitor, $C_{IF}$, a diode, two resistors, $R_L$ and $R_{INV}$, and nine Si CMOS transistors ($M_1$-$M_9$). b) The firing rate of the pulses output from the "neuron" circuit, $r_f$, is plotted versus the average current, $\bar{I}$, input to the "neuron" circuit (open circles). The experimental data were fitted by $r_f = \frac{r_s}{1+e^{-\chi[\bar{I}-I_0]}}$ (red line) with $r_s = 152\ Hz$, $\chi = 0.26\ /\text{nA}$, and $I_0 = 16.0$ nA.

## 3.2. Concurrent Learning and Signal Processing

An $M \times N$ crossbar circuit of synstors (Figure 3) processes and learns from an $M$-dimensional $\vec{V}_i$ pulse wave and an $N$-dimensional $\vec{V}_o$ pulse wave simultaneously. In an $M \times N$ crossbar circuit (Figure 25), a synstor connected with the m$^{th}$ input and n$^{th}$ output electrodes experiences various combinations of $V_i^m$ and $V_o^n$ voltages: (i) When $V_i^m \neq 0$ and $V_o^n = 0$, a current $I^{mn}(t) = (w^{nm}\kappa^{nm}) \circledast V_i^m$ (Equation 1) is triggered via the synstor for inference, and $\dot{w}^{nm} = V_i^m V_o^n = 0$ (Equation 2) for learning. (ii) When $V_i^m = V_o^n \neq 0$, $I^{mn} = 0$ for inference, and $\dot{w}^{nm} = \alpha\, V_i^m V_o^n \neq 0$ for learning. (iii) Under all other conditions including $V_i^m = V_o^n = 0$, and $V_o^n \neq 0$ under $V_i^m = 0$, then $I^{mn} = 0$ for inference, and $\dot{w}^{nm} = 0$ for learning. When a pulse is fired on

the n[th] output electrode, $V_o^n \neq 0$, the n[th] output electrode is disconnected from the integrate-and-fire "neuron" circuit, and the current does not flow into the "neuron" circuit, thus $I^{mn} = 0$. Therefore, the inference algorithm $I^n(t) = \sum_m (w^{nm}\kappa^{nm}) \circledast V_i^m$ when $V_o^n = 0$; $I^n(t) = 0$ when $V_o^n \neq 0$ (Equation 1), and the learning algorithm $\dot{w}^{nm} = \alpha\, V_i^m V_o^n$ (Equation 2) can be executed concurrently in the synstor circuit without interrupting each other.

## 3.3. Phase Shifts and Sneak Currents in Synstor Circuit

A synstor circuit needs to satisfy the following conditions in order to execute signal inference and learning concurrently: (1) To avoid the phase shifts of $\vec{V}_i$ and $\vec{V}_o$ pulse waves on the input and output electrodes, $M^2 R_e C_e, N^2 R_e C_e \ll t_d$, where $t_d$ denotes the pulse duration, $R_e$ denotes the resistance of the electrode per synstor unit, and $C_e$ as the capacitance of the electrode per synstor unit. In the synstor circuit reported in this work, $C_e \approx 10^{-19}\ F$, $R_e \approx 0.27\ \Omega$, $t_d = 10\ ns$, thus $M, N \ll 10^6$. In a 40 nm synstor circuit, $C_e \approx 10^{-17}\ F$, $R_e \approx 0.54\ \Omega$, thus $M, N \ll 10^5$. (2) To avoid significant voltage variations on the input and output electrodes induced by the sneak currents passing through the synstors, the worst-case voltage variation on an input or output electrodes induced by the maximal sneak currents, $\Delta V_{Max} < k_B T/q \approx 26\ mV$, with $q$ as the charge of an electron, $k_B$ as the Boltzmann constant, $T$ as the room temperature. In the synstor circuit reported in this work, $\Delta V_{Max} \approx 0.5N(N+1)V_a \bar{w} R_e, 0.5M(M+1)V_a \bar{w} R_e$ with $\bar{w} \approx 1.9$ nS, $V_p = 1.75\ V$, and $R_e \approx 0.27\ \Omega$, thus $M, N \lesssim 10^4$. In a 40 nm synstor circuit, $\bar{w} \approx 1.9$ nS, $V_p = 1.75\ V$, $R_e \approx 0.54\ \Omega$, thus $M, N \lesssim 10^4$. $M, N$ can be increased by decreasing $\bar{w}$. To

reduce the currents in the circuit, no voltage pulses with opposite polarity are applied on the input and output electrodes simultaneously.

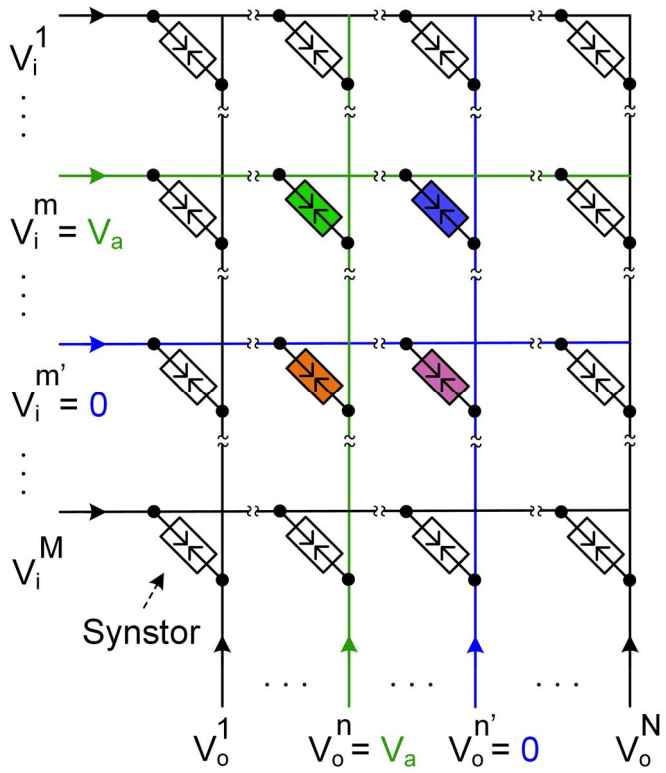


**Figure 25.** Voltage amplitudes to reduce current. An M×N crossbar circuit of synstors connected with M rows of input electrodes and N columns of output electrodes. $V_i^m$ denotes the amplitude of a voltage pulse on the m[th] input electrode, and $V_o^n$ denotes the amplitude of a voltage pulse on the n[th] output electrode. Synstors in the circuit experience $V_i^m$ and $V_o^n$ voltage pulses with the following possible combinations: (i) $V_i^m = V_a$ and $V_o^{n'} = 0$ (marked by blue color), (ii) $V_i^{m'} = V_o^{n'} = 0$ (marked by purple color), (iii) $V_i^m = V_o^n = V_a$ (marked by green color), (iv) $V_i^{m'} = 0$ and $V_o^n = V_a$ (marked by orange color) with $V_a \neq 0$.

## 3.4. General Correlative Learning Algorithms in Synstor Circuits

Arbitrary analog input signals, $V_s^m$, from the m[th] input channel can be preprocessed to generate input potential pulses at the m[th] input electrode of an M×N crossbar synstor circuit as shown in Figure 3 with $V_i^m = f(V_s^m)$. The output pulses from the synstor circuit, $V_f^n$, can be processed to generate the potential pulses on the n[th] output electrode of the synstor circuit, i.e. $V_o^n = g(V_f^n)$. The synstor circuit can implement universal correlative learning algorithms in machine learning based on Equation 2

$$\dot{w}^{nm} = \alpha\, V_i^m V_o^n = \alpha\, f(V_s^m)\, g(V_f^n) \qquad 8$$

In this work, the speech signals were converted to input pulses with their firing rates proportional to MFCCs[44] of the speech signals, and $\vec{V}_f$ pulses triggered $\vec{V}_o$ pulses to implement the "winner-take-all" learning algorithm (Equation 11). To implement spike-timing-dependent plasticity (STDP) learning algorithm, the input signals are converted to the positive and negative $\vec{V}_i$ pulses with their average firing rates, $\vec{r}_i^{+}(t)$ and $\vec{r}_i^{-}(t)$, equal to each other, i.e. $\int(\vec{r}_i^{+} - \vec{r}_i^{-})\,\mathrm{dt} = 0$. The firing rate of positive $\vec{V}_o$ pulses, $\vec{r}_o^{+}(t) = \rho^{+}\vec{r}_f(t-\tau)\, e^{-\tau/\tau^{+}}\ for\ \tau < 0$, and the firing rate of negative $\vec{V}_o$ pulses, $\vec{r}_o^{-}(t) = \rho^{-}\vec{r}_f(t-\tau)e^{\tau/\tau^{-}}\ for\ \tau > 0$ with the time constants $\tau^{+}, \tau^{-} > 0$, and parameters $\rho^{+}, \rho^{-} > 0$. From Equation 8,

$$\dot{w}^{nm} = \begin{cases} \beta^{-} r_i^m(t)\, r_f^n(t-\tau)e^{-\tau/\tau^{+}} & for\ \tau > 0 \\ -\beta^{+}\, r_i^m(t)\, r_f^n(t-\tau)e^{\tau/\tau^{-}} & for\ \tau < 0 \end{cases} \qquad 9$$

where $\vec{r}_i(t)$ denotes the firing rates of the $\vec{V}_i$ pulses, $\vec{r}_f(t)$ denotes the firing rates of the $\vec{V}_f$ pulses, $\beta^- = |\alpha|\, \rho^- V_o^a V_i^a t_d^2 > 0$, $\beta^+ = |\alpha|\, V_o^a V_i^a t_d^2 \rho^+ < 0$, $V_i^a$ denotes the magnitude of the $\vec{V}_i$ pulses, $V_o^a$ denotes the magnitude of the $\vec{V}_o$ pulse with $V_i^a = V_o^a$, and $t_d$ denotes the duration of the pulses.

# 4. Speech Recognition Experiment

## 4.1. Experimental and Results

The concurrent signal inference and learning were demonstrated in a $4 \times 2$ crossbar circuit (Figure 26a) composed of 8 synstors and 2 integrate-and-fire "neuron" circuits with the functions according to the Hodgkin–Huxley neuron model[43] (Figure 24). Original speech signals consisted of unlabeled "yes" and "no" utterances were pre-processed to generate the wave of voltage pulses, $V_i^m(t)$, input to the crossbar synstor circuit (Figure 26a). The input pulses had an amplitude of 1.75 V or -1.75 V, a duration of 10 ns, and firing rates, $r_i^m$, proportional to mel frequency cepstral coefficients (MFCCs)[44] of the speech signals at the different frequency ranges (Figure 26c).

The synstor circuit processed the $-1.75\,V$ $V_i^m(t)$ pulses and triggered currents by following $I^n(t) = \sum_m \kappa^{nm} \circledast (w^{nm} V_i^m)$ (Equation 1) under $V_o^n = 0$, which flowed into the integrate-and-fire "neuron" circuits to trigger 10 ns-wide $\pm 1.75\,V$ back-propagating pulses, $V_o^n(t)$, on the output electrodes (Figure 26d), and forwarding-propagating $1.0\,V$ output pulses, $V_f^n(t)$ (Figure 36e). The firing rates of the $V_f^n$ pulses, $r_f^n$, increased monotonically as a nonlinear sigmoid function of $I^n$ (Figure 24b). When a $V_f^n$ pulse was triggered from the $n^{th}$ "neuron" as a "winner" (i.e. $r_f^n > r_f^{n'}$, $n \neq n'$), a series of 10 ns-wide $-1.75\,V$ $V_o$ pulses was triggered on the $n^{th}$ output electrode of the "winner", and then a series of 10 ns-wide $1.75\,V$ $V_o$ pulses was triggered on the $n'^{th}$ output electrode of the "loser" (Figure 26d, e). To reduce the currents in the

circuit, no voltage pulses with opposite polarity were applied on the input and output electrodes simultaneously.

In the parallel unsupervised learning process, when the waves of $V_i^m$ and $V_o^n$ pulses encountered each other in the synstors (Figure 26f and g), the conductance matrix $[w^{nm}]_{NM}$ was modified by following Equation 2 and a "winner-take-all" learning algorithm[45],

$$\dot{w}^{nm} = \alpha\, V_i^m V_o^n \;\propto (r_f^n - r_f^{n'}) r_i^m \qquad 10$$

resulting in the change of $r_f^n$, $\delta r_f^n \propto \begin{cases} (\bar{r}_{f,Y}^n - \bar{r}_{f,Y}^{n'}) & \text{for "yes"} \\ (\bar{r}_{f,N}^n - \bar{r}_{f,N}^{n'}) & \text{for "no"} \end{cases}$, with $\bar{r}_{f,Y}^n$ and $\bar{r}_{f,N}^n$ as the average firing rates of the $V_f^n$ pulses triggered by "yes" word and "no" words, respectively. Before learning started (i.e. $V_o = 0$ when $t < 16\ s$), the synstors had unspecified random conductances, the "yes" and "no" words triggered $V_f$ pulses from the two "neurons", which were not distinguished or orthogonal. After the learning started ($t > 16\ s$), the "yes" words triggered asymmetric $V_f$ pulses from the two "neurons" with $\bar{r}_{f,Y}^2 > \bar{r}_{f,Y}^1$, thus "neuron" 2 was the "winner" with $\delta r_{f,Y}^2 \propto (\bar{r}_{f,Y}^2 - \bar{r}_{f,Y}^1) > 0$, and $r_{f,Y}^2$ increased and stabilized at $\hat{r}_{f,Y}^2 \approx 16\ Hz$; $\delta r_{f,Y}^1 \propto (\bar{r}_{f,Y}^1 - \bar{r}_{f,Y}^2) < 0$, and $r_{f,Y}^1$ decreased to $\hat{r}_{f,Y}^1 = 0$. The "no" words triggered asymmetric $V_f$ pulses from the two "neurons" with $\bar{r}_{f,N}^1 > \bar{r}_{f,N}^2$, thus the "neuron" 1 was the "winner" with $\delta r_{f,N}^1 \propto (\bar{r}_{f,N}^1 - \bar{r}_{f,N}^2) > 0$, and $r_{f,N}^1$ increased and stabilized at $\hat{r}_{f,N}^1 \approx 14\ Hz$; $\delta r_{f,N}^2 \propto (\bar{r}_{f,N}^2 - \bar{r}_{f,N}^1) < 0$, and $r_{f,N}^2$ decreased to $\hat{r}_{f,N}^2 = 0$ (Figure 26e). After the circuit processed and learned from 2-3 unlabeled speech signals, the "yes" and "no" speech signals were stably mapped to two distinguishable orthogonal waves of output pulses with average firing rates $\vec{\hat{r}}_{f,Y} \approx \begin{bmatrix} 0 \\ 16 \end{bmatrix} Hz$ and $\vec{\hat{r}}_{f,N} \approx \begin{bmatrix} 14 \\ 0 \end{bmatrix} Hz$.

In comparison with computers, the $4 \times 2$ synstor circuit concurrently executed the signal inference (Equation 1) and learning (Equation 2) algorithms with an equivalent computational speed of ~$2.4 \times 10^9$ $FLOPS$ (Equation 13), a power consumption of ~15 nW (Equation 14), and an equivalent computational energy efficiency of ~$1.6 \times 10^{17}$ $FLOPS/W$ (Equation15).

We have demonstrated a synaptic resistor (synstor) with analog convolutional signal processing, correlative learning, and nonvolatile memory functions. The device is composed of a p-type semiconducting CNT network which formed Schottky contacts with input and output Al electrodes as a resistor, and a recessed $TiO_2$ charge storage layer embedded in a $HfO_2$ dielectric layer sandwiched between an Al reference electrode and the CNT network as a capacitor. For inference, a synstor processes a series of voltage pulses, $V_i(t)$, on its input electrode by charging the capacitor during the pulses, and discharging the capacitor after the pulses, and triggering a current via the CNT resistor, $I(t) = \kappa \circledast (w\, V_i)$ (Equation 1 and Equation 3) on its grounded output electrode ($V_o = 0$) as a convolution of $V_i(t)$ and the product of its DC conductance, $w$, and a kernel function $\kappa(t)$. When a series of paired $V_i$ and $V_o$ voltage pulses with the same amplitude and duration (i.e. $V_i = V_o$) are applied on the synstor simultaneously, $w$ is modified by following the Hebbian learning rule, $\dot{w}^{nm} = \alpha\, V_i^m V_o^n$ (Equation 2), where $\alpha$ is a nonlinear function of the amplitudes and numbers of $V_i$ and $V_o$ pulses (Equations 5 and 6). The paired negative (positive) pulses generate a potential difference between the CNT network and $TiO_2$ layer to increase (decrease) the electronic charge stored in the $TiO_2$ layer, which in turn attracts (repels) the holes in the p-type semiconducting CNT network to increase (decrease) its conductance with $\alpha > 0$ ($\alpha < 0$). Otherwise, when a synstor experiences $V_i$ and $V_o$ pulses under the condition $V_i \cdot V_o = 0$, the $V_i$ or $V_o$ potential mainly drops beyond the $TiO_2$ charge storage layer/Al reference electrode, and the magnitudes of the potential differences between the CNT network and the recessed $TiO_2$ layer are

below the threshold values to modify the charge stored in the $TiO_2$ layer, thus $\dot{w} = 0$ (Equation 2) for nonvolatile memory. A $4 \times 2$ crossbar synstor circuit connected with integrate-and-fire "neuron" circuits was demonstrated for concurrent inference and learning from "yes" and "no" speech signals. The speech signals were converted to a wave of input voltage pulses, $\vec{V}_i$, processed by the synstor circuit in parallel to generate output currents (Equation 1), which in turn triggered waves of forward-propagating output voltage pulses, $\vec{V}_f$, from the integrate-and-fire "neuron" circuits for inference, and back-propagating voltage pulses, $\vec{V}_o$, on the output electrodes of the synstors. During the inference, the conductance matrix $[w^{nm}]_{NM}$ was concurrently modified by following the correlative learning algorithm (Equation 2) in a parallel learning process, leading to the orthogonal waves of output $\vec{V}_f$ pulses to distinguish "yes" and "no" speech signals for inference.

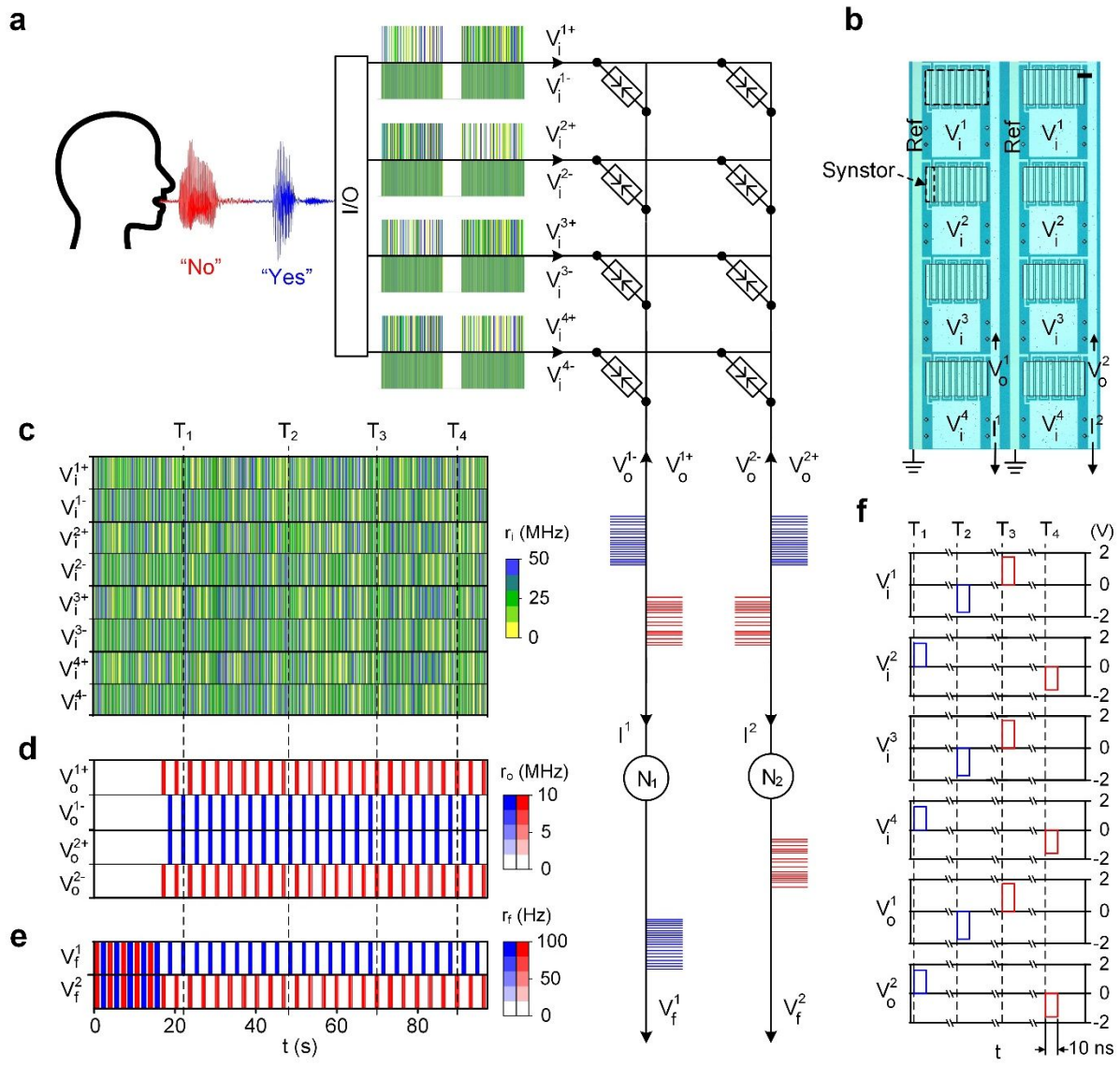


**Figure 26.** Speech recognition circuit diagram and results. a) A 4 × 2 synstor crossbar circuit connected to two integrate-and-fire "neuron" circuits, $N_1$ and $N_2$. The speech signals of "yes" and "no" words are converted to waves of voltage pulses, $V_i^m$ applied to input electrodes to trigger output currents, $I^n$, which in turn trigger back-propagating voltage pulses, $V_o^n$ and forward-propagating voltage pulses, $V_f^1$ from the "neurons". b) An optical image of a synstor chip with input electrodes connected with their contact pads. The scale bar is 200 μm. c) The firing rates of $\pm 1.75\ V$ 10 $ns$-wide pulses ($V_i^{1+}$, $V_i^{1-}$, $V_i^{2+}$, $V_i^{2-}$, $V_i^{3+}$, $V_i^{3-}$, $V_i^{4+}$, and $V_i^{4-}$) on the

four input electrodes. d) The firing rates of $\pm 1.75\ V$ $10\ ns$-wide back-propagating pulses ($V_o^{1+}$, $V_o^{1-}$, $V_o^{2+}$, and $V_o^{2-}$) on the two output electrodes. e) The $1.0\ V$ output pulses generated from the two "neurons" ($V_f^1$ and $V_f^2$) are shown in color gradients versus time. f) A "no" word triggers $1.75\ V$ $V_i^{2+}$ and $V_i^{4+}$ pulses on the 2nd and 4th input electrodes and a $1.75\ V$ $V_o^{2+}$ pulse on the 2nd output electrode concurrently, which decrease $w^{22}$ and $w^{42}$, the conductances of the synstors connected with the electrodes. A "no" word triggers $-1.75\ V$ $V_i^{1-}$ and $V_i^{3-}$ pulses on the 1st and 3rd input electrodes, and a $-1.75\ V$ $V_o^{1-}$ pulse on the 1st output electrode concurrently, which increase $w^{11}$ and $w^{31}$. A "yes" word triggers $1.75\ V$ $V_i^{1+}$ and $V_i^{3+}$ pulses on the 1st and 3rd input electrodes, and a $1.75\ V$ $V_o^{1+}$ on the 1st output electrode concurrently, which decrease $w^{11}$ and $w^{31}$. A "yes" word triggers $-1.75\ V$ $V_i^{2-}$ and $V_i^{4-}$ pulses on the 2nd and 4th input electrodes, and a $-1.75\ V$ $V_o^{2-}$ pulse on the 2nd output electrode concurrently, which increase $w^{22}$ and $w^{42}$. The duration of all the pulses is 10 ns. The dashed lines represent the moments when the paired pulses are triggered. The speech signals and pulses triggered by "yes" and "no" words are displayed in blue and red colors, respectively in d), e), and f).

## 4.2. Analysis

### 4.3.1. Learning Analysis

The mel frequency cepstral coefficients (MFCCs)[44] were extracted from the speech signals. Periodic voltage pulses with an amplitude of $1.75\ V$ or $-1.75\ V$, a duration of 10 ns, and a frequency of $50\ MHz$, were generated from a Tektronix AFG3152C waveform/function generator, and modulated by Si switching circuits (Maxim Integrated, IH5143CPE) and FPGA (National Instruments, cRIO-9063) to generate the waves of $\vec{V}_i(t)$ and $\vec{V}_o(t)$ voltage pulses on the input and

output electrodes of the synstor circuit (Figure 26a). The $\vec{V}_i$ pulses had firing rates, $\vec{r}_i$, proportional to MFCCs of the speech signals (Figure 26c). To reduce the currents in the circuit, no voltage pulses with opposite polarity were applied on the input and output electrodes simultaneously, and the $\vec{V}_i$ and $\vec{V}_o$ pulses encountered in the synstors had the same amplitudes, durations, and phases. The waves of potential pulses, $\vec{V}_i(t)$, input to the crossbar synstor circuit (Figure 26c), and triggered currents by following Equation 1, $I^n(t) = \sum_m (w^{nm}\kappa^{nm}) \circledast V_i^m$ when $V_o^n = 0$ (Equation 1), which triggered a wave of back-propagating pulses, $\vec{V}_o(t)$, with an amplitude of 1.75 V or -1.75 V and a duration of 10 ns on the output electrodes (Figure 26d), and a wave of forward-propagating output pulses, $\vec{V}_f(t)$, with an amplitude of 1 V (Figure 26e) from integrate-and-fire "neuron" circuits (Figure 24). The firing rates of the $\vec{V}_f$ pulses, $\vec{r}_f$, increased monotonically as a nonlinear sigmoid function of $\vec{I}$ (Figure 24b). Following a "winner-take-all" learning algorithm[45], when the $n^{th}$ "neuron" generated more $V_f$ pulses than the $n'^{th}$ "neuron" (i.e. $r_f^n > r_f^{n\prime}$), a series of $-1.75\ V$ $V_o$ pulses was triggered on the $n^{th}$ output electrode, and a series of $1.75\ V$ $V_o$ pulse was triggered on the $n'^{th}$ output electrode (Figure 26d and Figure 26e), therefore $V_o^n = -V_o^a \rho_{o/f}(r_f^n - r_f^{n'})t_d$, where $r_f^n$ and $r_f^{n\prime}$ denote firing rates of the $V_f$ pulses triggered from the $n^{th}$ and $n'^{th}$ "neurons", respectively, $V_o^a$ denotes the magnitude of the $\vec{V}_o$ pulses with $V_o^a = 1.75\ V$, $t_d$ denotes the duration of the pulses with $t_d = 10\ ns$, and $\rho_{o/f}$ denotes the ratio between the firing rates between the $V_o^n$ and $V_f^n$ pulses with $\rho_{o/f} > 0$. The conductance matrix $[w^{nm}]_{NM}$ of the circuit was modified by following Equation 2, $\dot{w}^{nm} = \alpha\ V_i^m V_o^n$. When $V_i = V_o > 0$, $\alpha \leq 0$; when $V_i = V_o < 0$, $\alpha \geq 0$; when $V_o \cdot V_i = 0$, $\alpha = 0$, thus $\dot{w}^{nm} = \alpha\ V_i^m V_o^n = -|\alpha||V_i^m|\ V_o^n = |\alpha||V_i^m|\ V_o^a \rho_{o/f}(r_f^n - r_f^{n'}) = |\alpha|V_i^a V_o^a \rho_{o/f} t_d^2 r_i^m (r_f^n - r_f^{n'})$, thus,

$$\dot{w}^{nm} = \beta(r_f^n - r_f^{n'})r_i^m \qquad 11$$

where $r_i^m(t)$ denotes the firing rates of the $V_i^m$ pulses, $|V_i^m| = V_i^a t_d\, r_i^m$, $\beta = |\alpha|\, V_o^a V_i^a \rho_{\frac{o}{f}} t_d^2 \geq 0$, and $V_i^a$ denotes the magnitude of the $\vec{V}_i$ pulses with $V_i^a = 1.75\, V$. After the synstor circuit processed the "yes" and "no" speech signals, $w^{nm}$ was modified by following Equation 11, $\delta w^{nm} \approx \beta\, T_d \sum_k (r_{f,kY}^n - r_{f,kY}^{n'}) r_{i,kY}^m + \beta\, T_d \sum_{k'} \left(r_{i,k'N}^n - r_{i,k'N}^{n'}\right) r_{i,k'N}^m$, where $r_{i,kY}^m$ and $r_{i,k'N}^m$ denote the average firing rates of the $V_i^m$ pulses triggered by the $k^{\text{th}}$ "yes" word and the $k'^{\text{th}}$ "no" word, respectively; $r_{f,kY}^n$ and $r_{i,k'N}^n$ denote the average firing rates of the $V_f^n$ pulses triggered by the $k^{\text{th}}$ "yes" word and the $k'^{\text{th}}$ "no" word, respectively; and $T_d$ denotes the duration of the speech signals. The change of the $[w^{nm}]_{NM}$ matrix induced the change of $I^n$, which in turn induced the change of $r_f^n$. Based on Equation 1, $I^n = \sum_m (w^{nm}\kappa^{nm}) \circledast V_i^m$, $\frac{\partial I^n}{\partial w^{nm}} = \kappa^{nm} \circledast V_i^m$. In the "neuron" circuit, $r_f^n$ increases monotonically as a sigmoid function of $I^n$, $r_f^n = \frac{r_s}{1+e^{-\chi[I^n - I_0]}}$ (Equation 7), thus $\mu^n = \frac{\partial r_f^n}{\partial I^n} = \chi r_f^n \left[1 - \frac{r_f^n}{r_s^n}\right] \geq 0$. The change of $r_f^n$ due to $\delta w^{nm}$, $\delta r_f^n = \frac{\partial r_f^n}{\partial I^n}\frac{\partial I^n}{\partial w^{nm}}\delta w^{nm} = \mu^n(\kappa^{nm} \circledast V_i^m)\delta w^{nm} = V_i^a t_d \mu^n \delta w^{nm}(\kappa^{nm} \circledast r_i^m)$ with $V_i^m = V_i^a t_d r_i^m$. After the synstor circuit processed the "yes" and "no" speech signals, $\delta r_f^n = V_i^a t_d \mu^n \sum_m (\delta w^{nm}\kappa^{nm}) \circledast r_i^m = V_i^a t_d\, \beta\, T_d \mu^n \{\sum_{m,k}\left[\left(r_{f,kY}^n - r_{f,kY}^{n'}\right) r_{i,kY}^m \kappa^{nm}\right] \circledast r_i^m + \sum_{m,k'}[(r_{i,k'N}^n - r_{i,k'N}^{n'}) r_{i,k'N}^m\, \kappa^{nm}] \circledast r_i^m\} = \sum_{m,k}\left[\left(r_{f,kY}^n - r_{f,kY}^{n'}\right)\gamma_{i,kmY}\right] \circledast r_i^m + \sum_{m,k'}[(r_{i,k'N}^n - r_{i,k'N}^{n'})\gamma_{i,k'mN}] \circledast r_i^m$

$\approx (\bar{r}_{f,Y}^{n} - \bar{r}_{f,Y}^{n'}) \sum_{m,k} \gamma_{i,kmY} \circledast r_i^m + (\bar{r}_{f,N}^{n} - \bar{r}_{f,N}^{n'}) \sum_{m,k'} \gamma_{i,k'mN} \circledast r_i^m$, where $\gamma_{i,kmY} = V_i^a t_d \beta T_d \mu^n r_{i,kY}^m \kappa^{nm} > 0$, and $\gamma_{i,k'mN} = V_i^a t_d \beta T_d \mu^n r_{i,k'N}^m \kappa^{nm} > 0$, and $\bar{r}_{f,Y}^{n}$ and $\bar{r}_{f,N}^{n}$ denote the average firing rates of the $V_f^n$ pulses triggered by "yes" word and "no" words, respectively. The statistical correlations between the signals triggered by the same words were significantly larger than the statistical correlations between signals triggered by the different words, therefore, when $r_i^m$ represented the "yes" word, $(\bar{r}_{f,Y}^{n} - \bar{r}_{f,Y}^{n'}) \sum_{m,k} \gamma_{i,kmY} \circledast r_i^m \gg (\bar{r}_{f,N}^{n} - \bar{r}_{f,N}^{n'}) \sum_{m,k'} \gamma_{i,k'mN} \circledast r_i^m$; when $r_i^m$ represented the "no" word, $(\bar{r}_{f,N}^{n} - \bar{r}_{f,N}^{n'}) \sum_{m,k'} \gamma_{i,k'mN} \circledast r_i^m \gg (\bar{r}_{f,Y}^{n} - \bar{r}_{f,Y}^{n'}) \sum_{m,k} \gamma_{i,kmY} \circledast r_i^m$. Thus,

$$\delta r_f^n \approx \begin{cases} (\bar{r}_{f,Y}^{n} - \bar{r}_{f,Y}^{n'}) \sum_{m,k} \gamma_{i,kmY} \circledast r_i^m & for\ "yes" \\ (\bar{r}_{f,N}^{n} - \bar{r}_{f,N}^{n'}) \sum_{m,k'} \gamma_{i,k'mN} \circledast r_i^m & for\ "no" \end{cases} \quad 12$$

In the experiment, after competitions between the two "neurons" in the initial learning stage ($t < 52\ s$), the "yes" words triggered $V_f$ pulses with $\bar{r}_{f,Y}^{2} > \bar{r}_{f,Y}^{1}$, thus "neuron" 2 was the "winner" with $\delta r_{f,Y}^{2} \propto (\bar{r}_{f,Y}^{2} - \bar{r}_{f,Y}^{1}) > 0$, and $r_{f,Y}^{2}$ increased and stabilized at $\hat{r}_{f,Y}^{2} = 47\ Hz$; $\delta r_{f,Y}^{1} \propto (\bar{r}_{f,Y}^{1} - \bar{r}_{f,Y}^{2}) < 0$, and $r_{f,Y}^{1}$ decreased to $\hat{r}_{f,Y}^{1} = 0$. The "no" words triggered the $V_f$ pulses with $\bar{r}_{f,N}^{1} > \bar{r}_{f,N}^{2}$, thus the "neuron" 1 was the "winner" with $\delta r_{f,N}^{1} \propto (\bar{r}_{f,N}^{1} - \bar{r}_{f,N}^{2}) > 0$, and $r_{f,N}^{1}$ increased and stabilized at $\hat{r}_{f,N}^{1} = 92\ Hz$; $\delta r_{f,N}^{2} \propto (\bar{r}_{f,N}^{2} - \bar{r}_{f,N}^{1}) < 0$, and $r_{f,N}^{2}$ decreased to $\hat{r}_{f,N}^{2} = 0$ (Figure 26e).

## 4.3.2. Speed, Power Consumption, and Energy Efficiency

In comparison with computers, the equivalent computing operations in an $M \times N$ synstor circuit are approximately equal to $3MN$ to implement the signal inference algorithm (Equation 1, $2MN$ for multiplications between $\hat{w}$, $\hat{\kappa}$, and $\vec{V}_i$; $MN$ for accumulations), and $3MN$ to implement the learning algorithm (Equation 2, $2MN$ outer products between $\alpha$, $\vec{V}_i$, and $\vec{V}_o$; $MN$ for $\hat{w}$ modifications). The speed for the circuit to implement $6MN$ parallel inference and learning operations,

$$v_s = 6MNf_s \qquad 13$$

where $f_s$ denotes the circuit operation frequency. With $f_s = 50\ MHz$, the $4 \times 2$ synstor circuit operated at a speed of $2.4 \times 10^9\ \ FLOPS$, and a $2\ k \times 2\ k$ synstor circuit (projected) could operate at a speed of $1.2 \times 10^{15}\ \ FLOPS$ (Figure 28).

When a wave of voltage pulses with an amplitude of $V_p$ is applied on its input or output electrode of an $M \times N$ synstor circuit connected with integrate-and-fire "neuron" circuits, currents are driven through synstors and leakage resistors in "neuron" circuits, and charge or discharge capacitors in electrodes, synstors, and "neuron" circuits. The average power consumption induced by the pulses on the input and output electrodes under $V_i^m \cdot V_o^n = 0$ is approximately equal to $MN\bar{w}V_a^2 t_d \eta_s^- [(1 + |V_L|/V_a)^2 \bar{r}_i^- + \bar{r}_o^-] + MN\bar{w}V_a^2 t_d \eta_s^+ [(1 + |V_L|/V_a)^2 \bar{r}_i^+ + \bar{r}_o^+]$, where $\bar{w}$ denotes the average DC conductance of the synstors, $V_a$ denotes the pulse magnitude, $t_d$ denotes the pulse duration, $\bar{r}_i^-$ and $\bar{r}_i^+$ denote the average firing rates of negative and positive pulses on the input electrodes, $\bar{r}_o^-$ and $\bar{r}_o^+$ denote the average firing rates of pulses on the output electrodes, $\eta_s^-$ and $\eta_s^+$ are unitless coefficients related to the capacitors, resistors, and profiles of negative and positive pulses in the circuit, and $V_L$ denotes the voltage applied on the leakage resistors in the

"neuron" circuits. Voltage pulses with the same amplitude, $V_i^m = V_o^n$, on the input and output electrodes of synstors drive currents through CNTs to charge capacitors between the CNT network and a reference electrode, and the average power consumption induced by the paired pulses is approximately equal to $MN\bar{w}V_a^2 t_d(\bar{r}_p^- \eta_p^- + \bar{r}_p^+ \eta_p^+)$, where $\bar{r}_p^-$ and $\bar{r}_p^+$ denote the average firing rates of the paired negative and positive pulses, and $\eta_p^-$ and $\eta_p^+$ represent unitless coefficients related to the capacitors, resistors, and profiles of the negative and positive pulses in the circuit. The total power consumption of the $M \times N$ synstor circuit,

$$P_s \approx MN\bar{w}V_a^2 D_p + NE_p\bar{r}_f \qquad 14$$

where $D_p = (1 + |V_L|/V_a)^2 t_d(\bar{r}_i^- \eta_s^- + \bar{r}_i^+ \eta_s^+) + t_d(\eta_s^- \bar{r}_o^- + \eta_s^+ \bar{r}_o^+) + t_d(\bar{r}_p^- \eta_p^- + \bar{r}_p^+ \eta_p^+)$ is a unitless coefficient, $E_p$ denotes the average energy consumption to switch the transistors in a "neuron" circuit to trigger an output pulse and modulate back-propagating $V_o$ pulses, and $\bar{r}_f$ denotes the average firing rate of the output $V_f$ pulses from the "neuron" circuit. Based on the experimental results and circuit simulation, $P_s \approx 15$ nW for the $4 \times 2$ crossbar synstor circuit reported in this work with $\bar{w} \approx 1.9$ nS for 9 synstors at each cross-point, $V_a = 1.75\ V$, $V_L = -0.25\ V$, $t_d = 10\ ns$, $\bar{r}_i^- \approx 20.5\ MHz$, $\bar{r}_i^+ \approx 0.41\ MHz$, $\bar{r}_o^- = \bar{r}_o^+ \approx 75\ KHz$, $\bar{r}_p^- = \bar{r}_p^+ \approx 30.8\ kHz$, $\bar{r}_f \approx 15\ Hz$, $\eta_s^- = 1.14$, $\eta_s^+ = 3.65$, $\eta_p^- = 1.3 \times 10^{-6}$, $\eta_p^+ = 4.3 \times 10^{-6}$, $D_p \approx 0.32$, and $E_p \approx 28\ fJ$. $P_s \approx 7.5$ mW for a projected $2k \times 2k$ crossbar synstor circuit operated at the same conditions (Figure 28).

The energy efficiency of a synstor circuit is equal to its speed (Equation 13) divided by its power consumption 14),

$$F_s = 6f_s/(\bar{w}V_a^2 D_p + E_p \bar{r}_f/M) \quad (15)$$

The energy efficiencies of the $4 \times 2$ synstor circuit and projected $2\,k \times 2\,k$ synstor circuit are approximately equal to $1.6 \times 10^{17}\ \ FLOPS/W$ (Figure 28). Based on Equation 15, the energy efficiency of a synstor circuit increases with decreasing synstor conductance $\bar{w}$ (Figure 27).

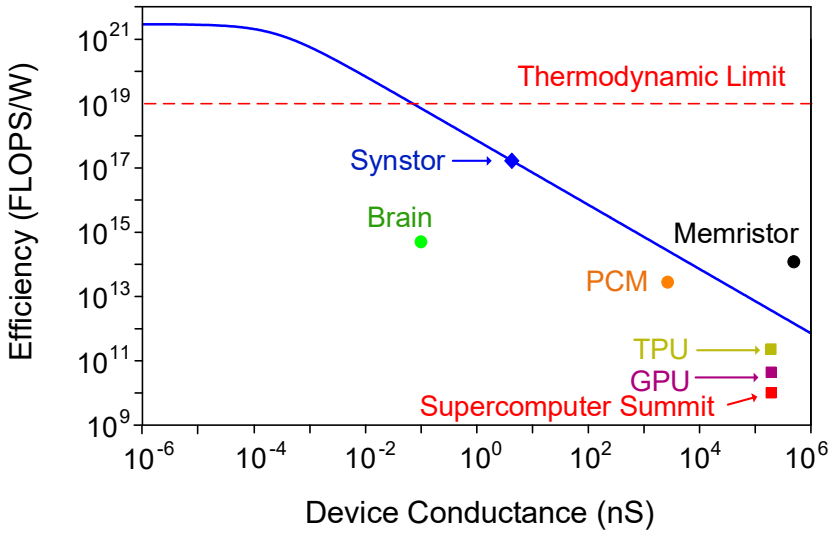


**Figure 27.** Energy efficiencies versus conductance. The energy efficiencies of the Summit supercomputer, Nvidia Volta V100 graphics processing unit (GPU), Google tensor processing unit (TPU), UMass/HP memristor circuits (inference only, training and learning algorithm computation excluded), IBM phase change memory (PCM) circuit (training only, learning algorithm computation excluded), human brain, $4 \times 2$ crossbar synstor circuit (blue diamond), and projected $4 \times 2$ crossbar synstor circuit (blue line) are displayed versus the average conductance of the Si transistors, memristors, PCM, and synstors in the circuits in the unit of nano siemens (nS). The thermodynamic limit of the computing energy efficiency in digital computing circuits is shown by the red dashed line.

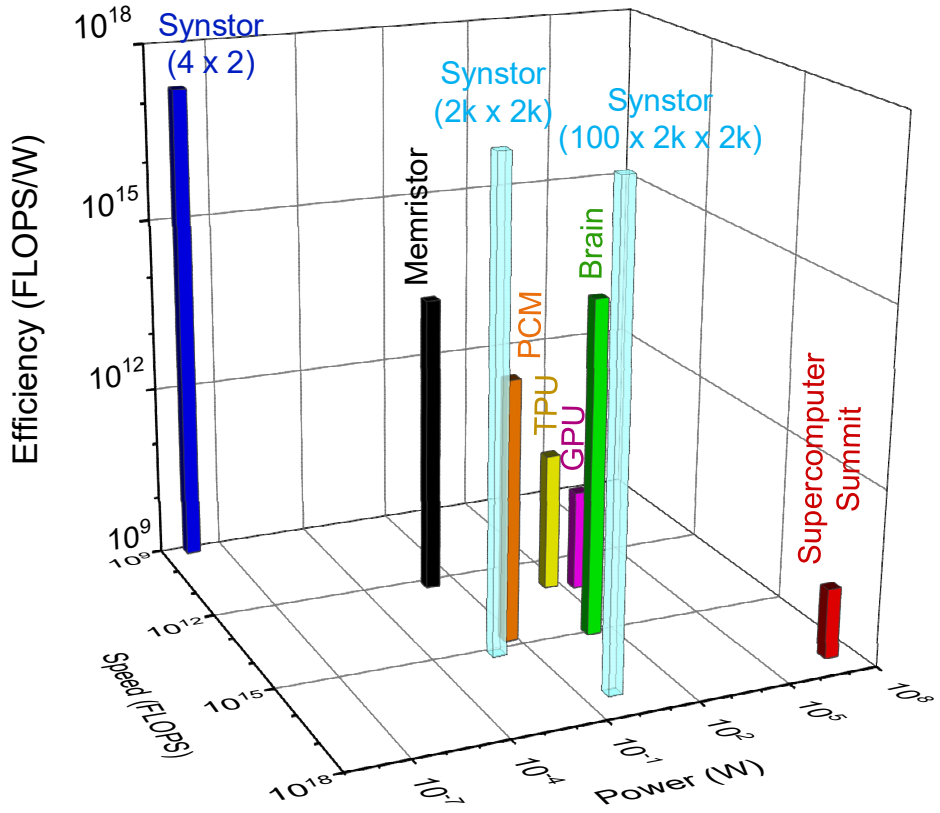


**Figure 28.** Power consumptions, computing speeds, and energy efficiencies of the Summit supercomputer, Nvidia GPU, Google TPU, UMass/HP memristor circuits (inference only, training and learning algorithm computation excluded), IBM PCM circuit (training only, learning algorithm computation excluded), human brain, $4 \times 2$ crossbar synstor circuit (deep blue), and projected $2\ k \times 2\ k$ and $100 \times 2\ k \times 2\ k$ crossbar synstor circuits (sky blue).

## 4.3.3. Latency of Synstor Circuit

The latency of the synstor circuit, $t_L$, is defined by the time needed for the input pulse wave, $\vec{V}_i$, to propagate through the synstor circuit to trigger a wave of output pulses from a "neuron"

circuit. Based on Equation 1, the average current flowing on an output electrode toward the "neuron", $\bar{I} \approx M\bar{w}(V_a - \bar{V}_{IF})D_{pi}$, where $M$ denotes the number of the synstors connected with the "neuron", $\bar{w}$ denotes the average conductance of synstors connected with the "neuron", $V_a$ denotes the amplitude of the $\vec{V}_i$ voltage pulses, $\bar{V}_{IF}$ denotes the average voltage on the capacitor in the "neuron" circuit, and $D_{pi}$ denotes a unitless coefficient related to the capacitors, resistors, and pulse profiles in the circuit. The current flows toward a capacitor with capacitance $C_{IF}$ in the integrate-and-fire "neuron" circuit to increase its voltage, $V_{IF} = \bar{I}t/C_{IF} = M\bar{w}D_{pi}(V_p - \bar{V}_{IF})\ t/C_{IF}$. When $V_{IF}$ reaches a threshold value, $V_{IF}^t$, a wave of output pulse is triggered from the "neuron" circuit, therefore its latency,

$$t_L = \frac{V_{IF}^t\ C_{IF}}{M\bar{w}D_{pi}(V_a - \bar{V}_{IF})} \qquad 16$$

Based on the experimental results, circuit simulation, and Equation 16, $t_L$ in the $4 \times 2$ synstor circuit ($M = 4$) and projected $2\ k \times 2\ k$ synstor circuit ($M = 2 \times 10^3$) is plotted versus $\bar{w}$ in Figure 29 with $C_{IF} = 1\ nF$, $V_a = 1.75\ V$, $\bar{V}_{IF} \approx 0.12\ V$, $D_{pi} = 0.31$, and $V_{IF}^t = 0.1\ V$. $t_L$ increases with decreasing $\bar{w}$ and $M$, and increases with increasing $C_{IF}$. In a series of digital computations, the total latency of serial computations in a transistor circuit is equal to the sum of the latency of individual computation, which need to be reduced to extremely small values by increasing $\bar{w}$. In a parallel analog computation, $t_L$ in a synstor circuit can be reduced by increasing $M$ (Figure 29). In Figure 29 the latencies of the Summit supercomputer, Nvidia Volta V100 graphics processing unit (GPU), Google tensor processing unit (TPU), human brain, $4 \times 2$

crossbar synstor circuit (blue diamond), and projected $2k \times 2k$ crossbar synstor circuit (green line) are displayed versus the conductance of the synstors, transistors, and synapses.

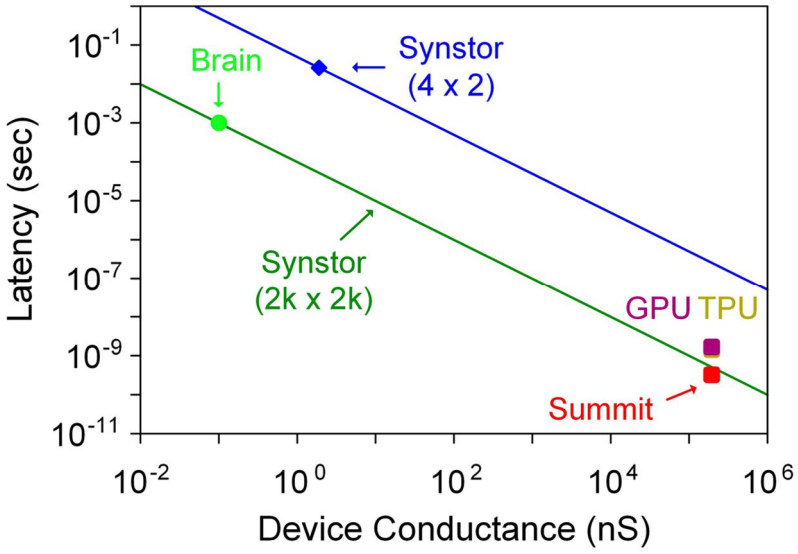


**Figure 29.** Latency comparisons as a function of device conductance.

## 4.4. Speech Pre-processing

Original speech signals consisting of unlabeled "yes" and "no" utterances were converted to their mel frequency cepstral coefficients (MFCCs)[44] at different frequency ranges by a standard MATLAB speech pre-processing program (https://www.mathworks.com/matlabcentral/fileexchange/32849). In the program, the speech signal was first processed using a first order finite impulse response (FIR) filter, and the speech signal was separated into sections in time using a Hamming window. Then, the magnitude spectrum was computed by fast Fourier transform (FFT). Filterbanks of triangular filters uniformly spaced on the mel scale within different frequency ranges were applied to the magnitude spectrum values to produce filterbank energies (FBEs). Log-compressed FBEs were decorrelated using the

discrete cosine transform to produce cepstral coefficients, and then a sinusoidal lifter was used to produce MFCCs at the different frequency ranges. Periodic voltage pulses with an amplitude of $1.75\,V$ or $-1.75\,V$, a duration of 10 ns, and a frequency of $50\,MHz$, were generated from a Tektronix AFG3152C waveform/function generator, and modulated by Si switching circuits (Maxim Integrated, IH5143CPE) and FPGA (National Instruments, cRIO-9063) to generate the waves of $\vec{V}_i(t)$ and $\vec{V}_o(t)$ voltage pulses on the input and output electrodes of the synstor circuit (Figure 26a). The $\vec{V}_i$ pulses had firing rates, $\vec{r}_i$, proportional to MFCCs at the different frequency ranges.

# 5. Circuit Modeling

Note that in Sections 0 and 6, the variable names are different than in Sections 1-4, and new variable names will be introduced in this section.

## 5.1. Signal Processing and Learning

For signal processing (Figure 5a), a synstor processes a series of voltage pulses, $V^x(t)$, on its input electrode by charging the capacitor during the pulses, and discharging the capacitor after the pulses, and triggering a current via the resistor, $I(t) = \kappa * (w\,V^x)$ on its grounded feedback electrode ($z = 0$) as a convolution of $V^x(t)$ and the product of its conductance, $w$, and a kernel function $\kappa(t)$. As shown in Figure 5b, when a series of paired input $V^x(t)$ and feedback $V^z(t)$ voltage pulses with the same amplitude (i.e. $V^x = V^z$) are applied on the synstor simultaneously, $w$ is modified by following the Hebb's learning rule, $\dot{w} = \beta\,V^z V^x$, where $\beta$ is a nonlinear function of the amplitudes and numbers of $V^x$ and $V^z$ pulses. The paired negative

(positive) pulses generate a potential difference between the channel and charge storage layer to increase (decrease) the electronic charge stored in the charge storage layer, which in turn attracts (repels) the holes in the semiconducting channel to increase (decrease) its conductance with $\beta > 0$ ($\beta < 0$). Otherwise, when a synstor experiences $V^x$ and $V^z$ pulses under the condition $V^x V^z = 0$, the $V^x$ or $V^z$ potential mainly drops beyond the charge storage layer and reference electrode, and the magnitudes of the potential differences between the channel and the recessed charge storage layer are below the threshold values to modify the charge stored in the charge storage layer, thus $\dot{w} = 0$ for nonvolatile memory.

## 5.2. Self-Programming

The learning algorithm, $\dot{\boldsymbol{w}} = \beta\, \boldsymbol{V^z} \otimes \boldsymbol{V^x}$ (Equation 2), implemented in a synstor circuit can be viewed as a generic Hebb's learning rule, where $\boldsymbol{w} = w_{ji}$, $\boldsymbol{V^z} = V_j^z$, and $\boldsymbol{V^x} = V_x$. In principle, all major machine learning algorithms, including unsupervised, supervised, and reinforcement learning, can be implemented in synstor circuits based on $\dot{\boldsymbol{w}} = \beta\, \boldsymbol{V^z} \otimes \boldsymbol{V^x}$ by setting $\boldsymbol{V^z} = f(\boldsymbol{w}, \boldsymbol{V^x}, \boldsymbol{V^y})$, where $\boldsymbol{V^y} = V_j^y$ is the vector of output voltage pulses produced by the neuron circuit.[20] The Hebbian or anti-Hebbian learning algorithm can be implemented in the synstor circuit by setting $\boldsymbol{V^z} = \boldsymbol{V^y}$. When an output voltage pulse, $V_j^y(t) = \delta(t - t_j)$, is triggered from the j[th] output neuron circuit at $t = t_j$, a negative (positive) feedback pulse, $V_j^z(t) = \delta(t - t_j)$, is triggered from the j[th] output neuron circuit simultaneously at $t = t_j$. Substituting $\boldsymbol{V^z}$ in Equation 2 by $\boldsymbol{V^y}$ yields $\dot{\boldsymbol{w}} = \beta\, \boldsymbol{V^y} \otimes \boldsymbol{V^x}$ with $\beta > 0$ ($\beta < 0$) for Hebbian (anti-Hebbian) learning.

In a neural network, the matrix of synaptic weights (conductances), $\boldsymbol{w}$, is modified in parallel based on synaptic spike-timing-dependent plasticity (STDP),[46] where $\boldsymbol{V^z}$ is a function of the timing difference between $\boldsymbol{V^x}$ and $\boldsymbol{V^y}$ that represents voltage pulses in the postsynaptic neurons,

$$\boldsymbol{V^z} = \boldsymbol{V^y} * \tilde{\theta} \qquad 17$$

where $\boldsymbol{V^y} * \tilde{\theta} = \int_{-\infty}^{\infty} \tilde{\theta}(t-t')\boldsymbol{V^y}(t')dt'$ with $\tilde{\theta}(t) = \begin{cases} -e^{t/\tau_-}/\tau_- & when\ t < 0 \\ 0 & when\ t = 0 \\ e^{-t/\tau_+}/\tau_+ & when\ t > 0 \end{cases}$, and the time constants $\tau_+ > 0$ and $\tau_- > 0$. $\int_{-\infty}^{\infty} \tilde{\theta}(t)dt = 0$, thus $\int_{-\infty}^{\infty} \boldsymbol{V^z}(t)dt = 0$. By substituting $V^z = \mathbf{V^y} * \tilde{\theta}$ (Equation 16) in Equation 2, $\dot{w}_{ji} = \sum_{t_j} \begin{cases} -\alpha V_i^x(t)\, e^{\frac{t-t_n}{\tau_-}}/\tau_- & when\ t < t_j \\ 0 & when\ t = t_j \\ \alpha V_i^x(t) e^{-\frac{t-t_n}{\tau_+}}/\tau_+ & when\ t > t_j \end{cases}$, where $t_j$ denotes the moment $t_j$ when an output voltage pulse is triggered from the j[th] output "neuron", $\alpha > 0$ for STDP, and $\alpha < 0$ for anti-STDP.

The STDP or anti-STDP learning algorithms can be implemented in the synstor circuit by setting $\boldsymbol{Z} = \boldsymbol{Y} * \tilde{\theta}$ with $\tilde{\theta}(t) = \begin{cases} -e^{t/\tau_-}/\tau_- & when\ t < 0 \\ 0 & when\ t = 0 \\ e^{-t/\tau_+}/\tau_+ & when\ t > 0 \end{cases}$, where $\boldsymbol{Z}$ is the vector of feedback pulse firing rates and $\boldsymbol{Y}$ is the vector of output pulse firing rates. When an output voltage pulse, $V_j^y(t) = \delta(t - t_j)$, is triggered from the j[th] output neuron circuit at $t = t_j$, a train of positive (negative) feedback pulses with a pulse firing rate $\left|Z_j(t-t_j)\right| = e^{(t-t_j)/\tau_-}/\tau_-$ under $t < t_j$, and a train of negative (positive) feedback pulses with a pulse firing rate $\left|Z_j(t-t_j)\right| = e^{-(t-t_j)/\tau_+}/\tau_+$

under $t > t_j$ are triggered from the j[th] output neuron circuit on the output electrode to implement STDP (anti- STDP) algorithms. For concurrent square pulses,

$$\dot{\boldsymbol{w}} = \beta_r \, \boldsymbol{Z} \otimes \boldsymbol{X} \tag{18}$$

where $\beta_r = \beta(V^z t_d)^2$, and $V^z = V^x$. By substituting $\boldsymbol{Z} = \boldsymbol{Y} * \tilde{\theta}$ in Equation 18, $\dot{w}_{ji} = \sum_{t_j} \begin{cases} -\beta_r X_i(t)\, e^{\frac{t-t_j}{\tau_-}}/\tau_- & when\ t < t_j \\ 0 & when\ t = t_j \\ \beta_r X_i(t) e^{-\frac{t-t_n}{\tau_+}}/\tau_+ & when\ t > t_j \end{cases}$, with $\beta > 0$ for STDP learning algorithm, and $\beta < 0$ for anti-STDP learning algorithm. Although more advanced machine learning algorithms can be implemented by setting $\boldsymbol{V^z} = f(\boldsymbol{w}, \boldsymbol{V^x}, \boldsymbol{V^y})$, practically the computations of complex algorithms involving the iterative computation, memory, writing, and reading processes of synstor conductance matrix, $\boldsymbol{w}$, using external transistor-based computing circuits with low speeds and energy efficiencies[47,48] should be avoided in the learning process for synstor circuits.

In the synstor circuit, $\widehat{\boldsymbol{w}}$ is not derived by computing learning algorithms offline based on collected data, but spontaneously transformed in a self-programming process via real-time learning. In the learning process, the feedback pulses, $\boldsymbol{V^z}$, are set as

$$\boldsymbol{V^z} = (\boldsymbol{V^y} - \widehat{\boldsymbol{V}}^{\mathrm{y}}) * \tilde{\theta} \tag{19}$$

where $\mu = 1\ or\ -1$, $\tilde{\theta}(t) = \begin{cases} -\mu\theta_-(t) < 0 & when -\tau_- < t < 0 \\ \mu\theta_+(t) > 0 & when\ \tau_+ > t > 0 \\ 0 & when\ t = 0\ or\ t \geq \tau_+\ or\ t \leq -\tau_- \end{cases}$, the time constants $\tau_+ > 0$ and $\tau_- > 0$, the function $\theta_+(t) > 0$ and $\theta_-(t) > 0$, and $\mu = 1$ or $-1$. The average $\tilde{\theta}$ over learning period $T$, $\langle\tilde{\theta}\rangle = 0$, and the average $\boldsymbol{V^z}$ over learning period $T$, $\langle\boldsymbol{V^z}\rangle = 0$, and $\widetilde{\boldsymbol{V^z}} = \boldsymbol{V^z} - \overline{\boldsymbol{V^z}} = \boldsymbol{V^z}$. $\widehat{\boldsymbol{V^y}}$ can be set in different learning algorithms. For example, $\widehat{\boldsymbol{V^y}} = 0$ in Hebbian and STDP learning. In supervised learning, $\widehat{\boldsymbol{V^y}}$ represented the desired value of $\boldsymbol{V^y}$; in unsupervised learning, $\widehat{\boldsymbol{V^y}}$ represents the converged value of $\boldsymbol{V^y}$; in reinforcement learning, $\widehat{\boldsymbol{V^y}}$ is set as zero. To generate feedback pulses with $V_j^z = \mu(V_j^y - \hat{V}_j^y) * \tilde{\theta}$ with $\mu = 1$, when a $V_j^y$ pulse, but no $\hat{V}_j^y$ pulse, is triggered at the moment $t = t_j$, i.e. $V_j^y(t) - \hat{V}_j^y(t) = \delta(t - t_j)$, a train of positive feedback pulses with a pulse firing rate $\left|Z_j(t - t_j)\right| = \theta_-(t - t_j)$ within the time window $t_j - \tau_- < t < t_j$, and a train of negative feedback pulses with a pulse firing rate $\left|Z_j(t - t_j)\right| = \theta_+(t - t_j)$ within the time window $t_j < t < t_j + \tau_+$ are triggered from the j[th] output neuron circuit on the output electrode. When a $\hat{V}_j^y$ pulse, but no $V_j^y$ pulse, is triggered at the moment $t = t_j'$, i.e. $V_j^y(t) - \hat{V}_j^y(t) = -\delta(t - t_j')$, a train of negative feedback pulses with a pulse firing rate $\left|Z_j(t - t_j')\right| = \theta_-(t - t_j')$ within the time window $t_j' - \tau_- < t < t_j'$, and a train of positive feedback pulses with a pulse firing rate $\left|Z_j(t - t_j')\right| = \theta_+(t - t_j')$ within the time window $t_j' < t < t_j' + \tau_+$ are triggered from the j[th] output neuron circuit on the output electrode. The $\boldsymbol{V^z}$ pulses with the voltage polarities described above with $\mu = 1$ will lead to $\beta < 0$ in the learning process; $\boldsymbol{V^z}$ pulses with their voltage polarities opposite to those described above with $\mu = -1$ will lead to $\beta > 0$ in the learning process.

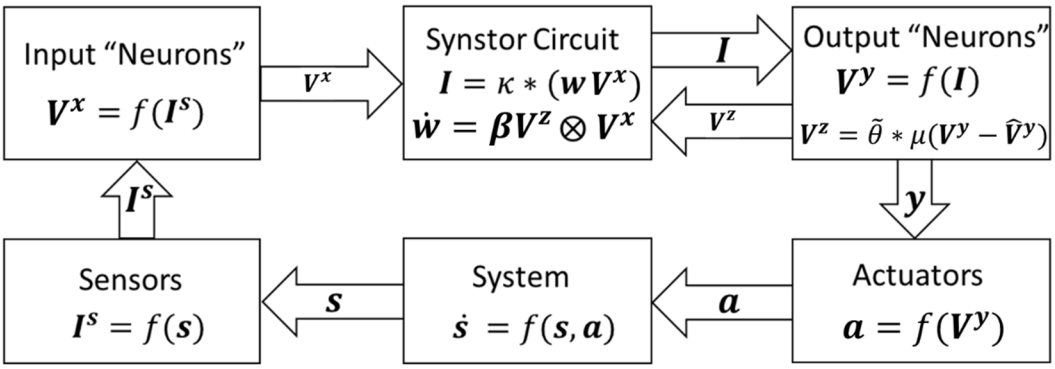


**Figure 30. System schematics for self-programming.** A synstor circuit is integrated with a system or another circuit. Output signals, $\boldsymbol{V^y}$, are transmitted to actuators to trigger actuation signals, $\boldsymbol{a}$, which modify the states of the system, $\boldsymbol{s}$. $\boldsymbol{s}$ is detected by sensors to generate currents, $\boldsymbol{I^s}$, flowing into the input neuron circuits.

## 5.3. Optimization of Objective Function

In a closed-loop synstor circuit as shown in Figure 30, the synstor circuit is connected with an external system or another circuit via sensors and actuators. Output signals $\boldsymbol{V^y}$ from the synstor circuit are transmitted to actuators to trigger actuation signals $\boldsymbol{a}$, which modify the states of the system $\boldsymbol{s}$. $\boldsymbol{s}$ is detected by sensors to generate currents $\boldsymbol{I^s}$, which are transmitted to the input neuron circuits to trigger input signals $\boldsymbol{V^x}$. $\boldsymbol{V^x}$ induces currents $\boldsymbol{I}$ flowing through the synstor circuit and collected by output neuron circuits to generate output voltage pulses $\boldsymbol{V^y}$, and feedback voltage pulses $\boldsymbol{V^z} = \boldsymbol{V^y} * \tilde{\theta}$ (Equation 18 with $\hat{V}^y = 0$). Concurrently $\boldsymbol{w}$ is modified by $\boldsymbol{V^x}$ and $\boldsymbol{V^z}$ by following Equation 2, $\dot{\boldsymbol{w}} = \beta\, \boldsymbol{V^z} \otimes \boldsymbol{V^x}$. The goal of the self-programming process is to modify $\boldsymbol{s}$ toward the desired state of the system, $\hat{\boldsymbol{s}}$, and an objective function can be defined as,

$$F = \frac{1}{2}(\boldsymbol{s} - \hat{\boldsymbol{s}})^2 \qquad 20$$

When $\boldsymbol{s} = \hat{\boldsymbol{s}}$, $F = 0$, $\boldsymbol{V^x} = 0$, and $\dot{\boldsymbol{w}} = \beta\, \boldsymbol{V^z} \otimes \boldsymbol{V^x} = 0$ (Equation 2), $\boldsymbol{w}$ reaches an equilibrium value $\hat{\boldsymbol{w}}$. By substituting $\boldsymbol{V^z}$ in Equation 2 by Equation 19, $\dot{\boldsymbol{w}} = \beta\, (\tilde{\theta} * \boldsymbol{V^y}) \otimes \boldsymbol{V^x}$.

In the system, $\boldsymbol{s} - \hat{\boldsymbol{s}}$ is detected by sensors to generate currents, $\boldsymbol{I^s}$, input to the "neuron" circuits, and triggers input pulses, $\boldsymbol{V^x}$, in the synstor circuit. The firing rate of $\boldsymbol{V^x}$ is a monotonically increasing nonlinear function of $|\boldsymbol{s} - \hat{\boldsymbol{s}}|$. When $\boldsymbol{s} = \hat{\boldsymbol{s}}$, $\boldsymbol{x} = \hat{\boldsymbol{x}} = 0$, thus $F$ can also be expressed as,

$$F = \frac{1}{2}(\boldsymbol{V^x})^2 + O[(\boldsymbol{V^x})^3] \qquad 21$$

where $O[(\boldsymbol{V^x})^3]$ represents higher order terms, $(\boldsymbol{V^x})^k$, with $k \geq 3$.

The input signals $V^{\boldsymbol{x}}$ are modified by the output signals $\boldsymbol{V^y}$ from the synstor circuit in a closed-loop circuit/system. Without loss of generality, $\boldsymbol{V^x}$ can be expressed as a function of $\boldsymbol{V^y}$,

$$\boldsymbol{V^x} = \boldsymbol{g}^{\boldsymbol{V^x/V^y}} * \left(\boldsymbol{V^y} - \hat{\boldsymbol{V}}^{\boldsymbol{y}}\right) + O[(\boldsymbol{V^y} - \hat{\boldsymbol{V}}^{\boldsymbol{y}})^2] \qquad 22$$

where $\boldsymbol{g}^{\boldsymbol{V^x/V^y}}$ represents a $(g_{ij}^{V^x/V^y})_{M\times N}$ matrix with $g_{ij}^{V^x/V^y}$ as a transfer function between $V_i^x$ and $V_j^y$ under $V_j^y = \hat{V}_j^y$, and $\boldsymbol{g}^{\boldsymbol{V^x/V^y}} * \left(\boldsymbol{V^y} - \hat{\boldsymbol{V}}^{\boldsymbol{y}}\right)$ represents the temporal convolution between $\boldsymbol{g}^{\boldsymbol{V^x/V^y}}$ and $\boldsymbol{V^y} - \hat{\boldsymbol{V}}^{\boldsymbol{y}}$ with $\boldsymbol{g}^{\boldsymbol{V^x/V^y}} * \left(\boldsymbol{V^y} - \hat{\boldsymbol{V}}^{\boldsymbol{y}}\right) = \boldsymbol{g}^{\boldsymbol{V^x/V^y}} \int_0^t \left[\boldsymbol{V^y}(t') - \hat{\boldsymbol{V}}^{\boldsymbol{y}}(t')\right] dt'$. When $\boldsymbol{V^y} =$

$\widehat{\boldsymbol{V}}^{\boldsymbol{y}}, \boldsymbol{V}^{\boldsymbol{x}} = 0$, thus $\widehat{\boldsymbol{V}}^{y} = arg\min_{\boldsymbol{V}^{\boldsymbol{y}}} F^{\boldsymbol{s}}$. $O[(\boldsymbol{V}^{\boldsymbol{y}} - \widehat{\boldsymbol{V}}^{\boldsymbol{y}})^2]$ represents higher order terms, $(\boldsymbol{V}^{\boldsymbol{y}} - \widehat{\boldsymbol{V}}^{\boldsymbol{y}})^k$, with $k \geq 2$.

In the closed-loop synstor circuit (**Figure 30**b), feedback voltage signals $\boldsymbol{V}^{\boldsymbol{z}} = \boldsymbol{V}^{\boldsymbol{y}} * \tilde{\theta}$ (Equation 3). $\boldsymbol{w}$ is modified by $\boldsymbol{V}^{\boldsymbol{x}}$ and $\boldsymbol{V}^{\boldsymbol{z}}$ by following Equation 2, $\dot{\boldsymbol{w}} = \beta\, \boldsymbol{V}^{\boldsymbol{z}} \otimes \boldsymbol{V}^{\boldsymbol{x}}$ . By substituting $\boldsymbol{V}^{\boldsymbol{z}}$ in Equation 2 by Equation 18, $\dot{\boldsymbol{w}} = \beta\, (\tilde{\theta} * \boldsymbol{V}^{\boldsymbol{y}}) \otimes \boldsymbol{V}^{\boldsymbol{x}}$. By substituting $\boldsymbol{V}^{\boldsymbol{y}}$ in $\dot{\boldsymbol{w}}$ $\dot{\boldsymbol{w}} = \beta(\tilde{\theta} * \widehat{\boldsymbol{V}}^{\boldsymbol{y}}) \otimes \boldsymbol{V}^{\boldsymbol{x}} + \beta\{\tilde{\theta} * \{\boldsymbol{g}^{\frac{\boldsymbol{V}^{\boldsymbol{y}}}{I}} \circ \{\kappa * [(\boldsymbol{w} - \widehat{\boldsymbol{w}})\boldsymbol{V}^{\boldsymbol{x}}]\}\}\} \otimes \boldsymbol{V}^{\boldsymbol{x}}$. The average $\dot{w}_{ji}$ over learning period $T$, $\langle \dot{w}_{ji} \rangle = \beta\langle(\tilde{\theta} * \hat{V}_j^{y})V_i^{x}\rangle + \sum_{i'} \beta g_j^{V^y/I}\langle\{\tilde{\theta} * \kappa * [(w_{ji'} - \hat{w}_{ji'})V_{i'}^{x}]\}V_i^{x}\rangle +$ $O[(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2] = \sum_{i'} \beta g_j^{V^y/I}\langle\{\tilde{\theta} * \kappa * [(w_{ji'} - \hat{w}_{ji'})V_{i'}^{x}]\}V_i^{x}\rangle + O[(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2] = -\eta_{ji}(\langle w_{ji}\rangle - \langle\hat{w}_{ji}\rangle) + O[\langle(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2\rangle]$. Since $\langle\tilde{\theta}\rangle = 0$, then $\langle\tilde{\theta} * \widehat{\boldsymbol{V}}^{\boldsymbol{y}}\rangle = 0$. $\tilde{\theta} * \widehat{\boldsymbol{V}}^{\boldsymbol{y}}$ and $\boldsymbol{V}^{\boldsymbol{x}}$ are statistically independent, thus $\langle(\tilde{\theta} * \hat{V}_j^{y})V_i^{x}\rangle = 0$. Since $\sum_{i'} \alpha g_j^{V^y/I}\langle\{\tilde{\theta} * \kappa * [(w_{ji'} - \hat{w}_{ji'})V_{i'}^{x}]\}V_i^{x}\rangle = -\eta_{ji}(\langle w_{ji}\rangle - \langle\hat{w}_{ji}\rangle)$ then $\dot{\boldsymbol{w}}$ can be written as,

$$\langle\dot{\boldsymbol{w}}\rangle = -\boldsymbol{\eta} \circ (\langle\boldsymbol{w}\rangle - \langle\widehat{\boldsymbol{w}}\rangle) + \delta\boldsymbol{w} \qquad 23$$

where $\boldsymbol{\eta} \circ (\langle\boldsymbol{w}\rangle - \langle\widehat{\boldsymbol{w}}\rangle)$ denotes the Hadamard product between $\boldsymbol{\eta}$ and $\langle\boldsymbol{w}\rangle - \langle\widehat{\boldsymbol{w}}\rangle$, and $\boldsymbol{\eta}$ denotes a $(\eta_{ji})_{N\times M}$ matrix with $\eta_{ji} = -\alpha g_n^{V^y/I}\langle\tilde{\theta} * \kappa\rangle\langle V_i^x, V_i^x\rangle \geq 0$, $\alpha > 0$, $g_j^{V^y/I} \geq 0$, $\langle\tilde{\theta} * \kappa\rangle = -\int_0^{\tau_-} \kappa(t')\theta_-(t')dt' < 0$, and the variance of $V_i^x$, $\langle V_i^x, V_i^x\rangle \geq 0$. $\delta\boldsymbol{w}$ denotes a $(\delta w_{ji}^{\boldsymbol{y}})_{N\times M}$ matrix with $\delta\boldsymbol{w} = O[(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2]$.

In the self-programming process, the change of $\boldsymbol{w}$ leads to the change of output signals $\boldsymbol{V}^{\boldsymbol{y}}$, which modifies $\boldsymbol{V}^{\boldsymbol{x}}$ and the objective function $F$. Substituting $\boldsymbol{V}^{\boldsymbol{x}}$ yields $\langle F\rangle = \frac{1}{2}\langle[\boldsymbol{g}^{\boldsymbol{V}^{\boldsymbol{x}}/\boldsymbol{V}^{\boldsymbol{y}}} * (\boldsymbol{V}^{\boldsymbol{y}} - \widehat{\boldsymbol{V}}^{\boldsymbol{y}})]^2\rangle + O[(\boldsymbol{V}^{\boldsymbol{y}} - \widehat{\boldsymbol{V}}^{\boldsymbol{y}})^3]$ with $\langle F\rangle$ the average of $F$ over learning period $T$.

Substituting $\boldsymbol{V^y} - \widehat{\boldsymbol{V}}^{\boldsymbol{y}}$ yields $\langle F\rangle = \frac{1}{2}\langle\{\boldsymbol{g}^{\boldsymbol{V^x}/\boldsymbol{V^y}} * \{\boldsymbol{g}^{\boldsymbol{V^y}/\boldsymbol{I}} \circ \{\kappa * [(\boldsymbol{w} - \widehat{\boldsymbol{w}})\boldsymbol{V^x}]\}\}^2\rangle +$ $O[(\boldsymbol{w} - \widehat{\boldsymbol{w}})^3] = \frac{1}{2}\boldsymbol{g}^{F/w} \circ \langle(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2\rangle + O[(\boldsymbol{w} - \widehat{\boldsymbol{w}})^3]$ , where $\boldsymbol{g}^{F/w} \circ \langle(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2\rangle$ denotes the Hadamard product between $\boldsymbol{g}^{F/w}$ and $\langle(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2\rangle$, $\boldsymbol{g}^{F/w} = \langle\{\boldsymbol{g}^{\frac{\boldsymbol{V^x}}{\boldsymbol{y}}} * [\boldsymbol{g}^{\frac{\boldsymbol{V^y}}{\boldsymbol{I}}} \circ (\kappa * \boldsymbol{V^x})]\}^2\rangle \geq 0$, and $\langle(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2\rangle$ denotes the average of $(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2$ over learning period $T$. The change rate of the objective function $\langle\dot{F}\rangle = \langle\boldsymbol{g}^{F/w} \circ (\boldsymbol{w} - \widehat{\boldsymbol{w}}) \circ (\dot{\boldsymbol{w}} - \dot{\widehat{\boldsymbol{w}}})\rangle + O\left[(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2 \circ (\dot{\boldsymbol{w}} - \dot{\widehat{\boldsymbol{w}}})\right]$. Substituting $\dot{\boldsymbol{w}}$ in $\dot{F}$ yields, $\langle\dot{F}\rangle = -\boldsymbol{g}^{\frac{F}{w}} \circ \langle\boldsymbol{\eta} \circ (\boldsymbol{w} - \widehat{\boldsymbol{w}})^2\rangle - \boldsymbol{g}^{\frac{F}{w}} \circ \boldsymbol{\eta} \circ \langle(\boldsymbol{w} - \widehat{\boldsymbol{w}}) \circ \dot{\widehat{\boldsymbol{w}}}\rangle + O[(\boldsymbol{w} - \widehat{\boldsymbol{w}})^3] =$ $-\frac{1}{2}\eta\boldsymbol{g}^{F/w} \circ \langle(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2\rangle - \boldsymbol{g}^{F^s/w} \circ \langle(\boldsymbol{w} - \widehat{\boldsymbol{w}}) \circ \dot{\widehat{\boldsymbol{w}}}\rangle + O[(\boldsymbol{w} - \widehat{\boldsymbol{w}})^3]$, where $\frac{1}{2}\boldsymbol{g}^{F/w} \circ$ $\langle(\boldsymbol{w} - \widehat{\boldsymbol{w}})^2\rangle = \langle F\rangle$, and $\eta = \sum_{i,j} 2\langle\eta_{ji}\rangle / MN = -2\sum_{i,j} \beta g_j^{V^y/I}\langle\tilde{\theta} * \kappa\rangle\langle V_i^x, V_i^x\rangle / MN \geq 0$, thus

$$\langle\dot{F}\rangle = -\eta\langle F\rangle + \delta F \qquad 24$$

where $\delta F = -\boldsymbol{g}^{F/w} \circ \langle(\boldsymbol{w} - \widehat{\boldsymbol{w}}) \circ \dot{\widehat{\boldsymbol{w}}}\rangle + O[(\boldsymbol{w} - \widehat{\boldsymbol{w}})^3]$ and $\boldsymbol{g}^{F/w} = \langle\{\boldsymbol{g}^{\frac{\boldsymbol{V^x}}{\boldsymbol{V^y}}} * [\boldsymbol{g}^{\frac{\boldsymbol{V^y}}{\boldsymbol{I}}} \circ (\kappa * \boldsymbol{V^x})]\}^2\rangle$.

By substituting $V^{\boldsymbol{y}}$ in $\dot{\boldsymbol{w}}$ as a function of $\boldsymbol{w} - \widehat{\boldsymbol{w}}$, the dynamic change of $\boldsymbol{w}$ in the learning process can also be expressed by $\langle\dot{\boldsymbol{w}}\rangle = -\boldsymbol{\beta} \circ (\langle\boldsymbol{w}\rangle - \langle\widehat{\boldsymbol{w}}\rangle) + \delta\boldsymbol{w}$. In the self-programming processes, the change of $\boldsymbol{w}$ leads to the change of output signals $\boldsymbol{V^y}$, which modifies $\boldsymbol{s}$ and the objective function $F$. The dynamic change of $F$ in the self-programming process can also be described by $\langle\dot{F}\rangle = -\eta\langle F\rangle + \delta F$.

In a self-programming process of a closed-loop or opened-loop synstor circuit, when $\boldsymbol{w}$ is modified to $\widehat{\boldsymbol{w}}$, the objective function $F(\widehat{\boldsymbol{w}}) = 0$; when $\boldsymbol{w} \neq 0$, $F(\boldsymbol{w}) > 0$. Based on Equation 6, $\langle\dot{F}\rangle = -\eta\langle F\rangle + \delta F$, when $\eta\langle F\rangle > \delta F$, $\langle\dot{F}\rangle < 0$; when $\eta\langle F\rangle = \delta F$, $\langle\dot{F}\rangle = 0$. $\langle F\rangle$ also represents a Lyapunov function for $\boldsymbol{w}$. When $\langle\dot{F}\rangle < 0$, $\langle F\rangle$ is asymptotically decreased toward its optimal

value $\hat{F} = 0$ when $\langle \boldsymbol{w} \rangle$ is modified toward $\langle \widehat{\boldsymbol{w}} \rangle$ in the self-programming process; when $\langle \dot{F} \rangle = 0$, $\langle F \rangle$ reaches its dynamic equilibrium value $F_e = \delta F/\eta$ under $\langle \boldsymbol{w} \rangle = \langle \widehat{\boldsymbol{w}} \rangle$. $\delta F$ also functions as a perturbation term, which prevents $\langle F \rangle$ from reaching its optimal value $\hat{F} = 0$. $F$, $\delta F$, and $\eta$ may all change dynamically in a self-programming process.

## 5.4. Integrate-and-Fire Circuit Simulation

An integrate-and-fire circuit has the basic functions of the Hodgkin–Huxley neuron model.[43] The collective current, $I$, from synstors or sensors flows toward a capacitor $C_{IF}$, increasing the voltage, $V_C$, on the capacitor. A leakage current, $I_L$, flows out from the capacitor simultaneously, decreasing $V_C$, and $V_C = \int (I - I_L) dt \,/ C_{IF}$. When $V_C$ reaches a threshold value $V_{th}^{IF}$, an output pulse is triggered from the circuit, $V_C$ is reset back to zero, and the capacitor $C_{IF}$ restarts the integration of the current. The simulated input integrate-and-fire neuron is set with $C_{IF} = 50$ pF, $I_L = 0.5\ nA$ or $I_L = 1\ nA$ and a threshold voltage $V_{th}^{IF} = 0.3\ V$, the simulated output integrate-and-fire neuron circuit is set with $C_{IF} = 4$ nF, $I_L = 0$, and $V_{th}^{IF} = 0.3\ V$, and the firing rate, $f_{IF}$, of the pulses output from the circuits is shown versus $I$ input to the circuits in Figure 31. $f_{IF}$ is a nonlinear function of $I$. When $I < I_L$, $V_C$ cannot reach the threshold value $V_{th}^{IF}$, therefore there is no output pulse from the circuit with $f_{IF} = 0$; when $I_L < I < I_{sat}^{IF}$, $f_{IF}$ increases linearly with $I$; when $I > I_{sat}^{IF}$, $f_{IF}$ is saturated to a maximal value due to the time limit to reset $V_C$ to zero or the maximal frequency of input pulses. As shown in Figure 31, when $I_L$ is increased from $0.5\ nA$ to $1\ nA$, the firing rate $f_{IF}$ is decreased in the linear $f_{IF} - I$ region, and the threshold current $I$ to generate output pulses increases with increasing $I_L$. When $C_{IF}$ is increased from 0.5 pF to

4 nF, the firing rate $f_{IF}$ is decreased in the linear $f_{IF} - I$ region, and the slope of $f_{IF} - I$ curve in the linear $f_{IF} - I$ region decrease with increasing $C_{IF}$.

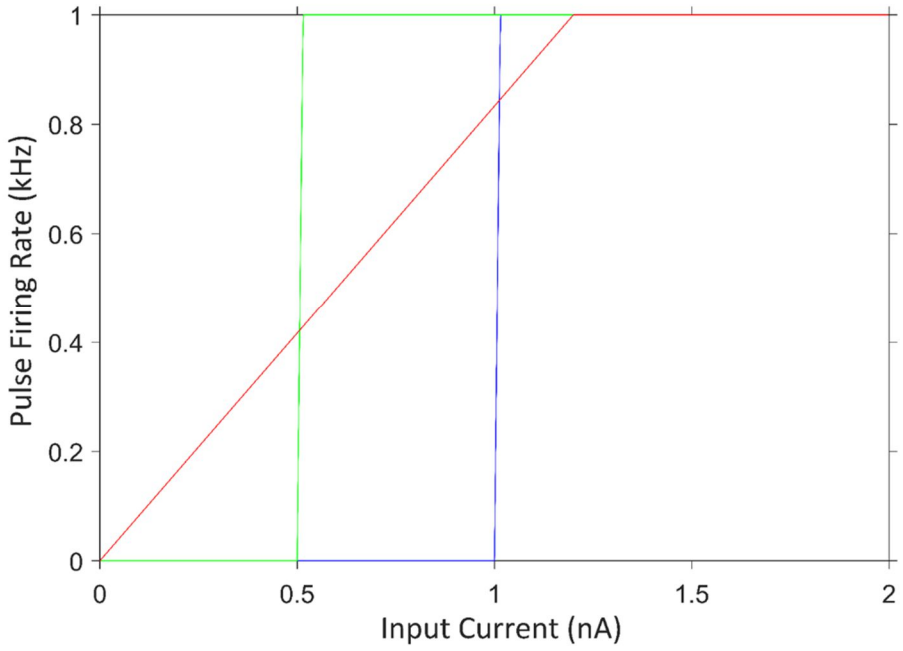


**Figure 31.** The firing rate of the pulses output from a simulated integrate-and-fire neuron circuit is plotted versus the current input to the circuit with a 50 pF capacitor and 0.5 $nA$ leakage current (green line), a 50 pF capacitor and 1 $nA$ leakage current (blue line), and a 4 nF capacitor and 0 $nA$ leakage current (red line).

## 5.5. Simulation of Learning Process

A closed-loop synstor circuit connected with an external system is simulated by a MATLAB software, and is shown in Figure 32a as an example of the self-programming process. The goal of the self-programming process is to modify state $s$ with an arbitrary unit in a system (Figure 32b) toward the desired state, $\hat{s} = 0$. $s$ is detected by sensors to generate a sensory current $\boldsymbol{I^s} = \begin{pmatrix} s \\ 0 \end{pmatrix}$ when $s \geq 0$, and $\boldsymbol{I^s} = \begin{pmatrix} 0 \\ |s| \end{pmatrix}$ when $s < 0$. $\boldsymbol{I^s}$ triggers input pulses $\boldsymbol{V^x}$ with amplitudes

of $\pm 1\,V$ and a duration of $2.5\,\mu s$ (Figure 32c) from input integrate-and-fire neuron circuits (Section 5.4 and Figure 31) with a capacitance $C_{IF} = 50$ pF, a leakage current $I_L = 1\,nA$, and a threshold voltage $V_{th}^{IF} = 0.3\,V$. $\boldsymbol{V^x}$ induces currents, $\boldsymbol{I}$, flowing from synstors into the output neuron circuits by following Equation 1, $\boldsymbol{I = w\,V^x}$, which flows into output integrate-and-fire neuron circuits to generate feedback pulses, $\boldsymbol{V^z}$ (Figure 32d), with amplitudes of $\pm 1\,V$ and a duration of $2.5\,\mu s$, and output pulses, $\boldsymbol{V^y}$ (Figure 32e), with an amplitude of $1\,V$ and a duration of $2.5\,\mu s$. The output integrate-and-fire neuron circuits (Section 5.4 and Figure 31) have a capacitance $C_{IF} = 4$ nF, a leakage current $I_L = 0$, and a threshold voltage $V_{th}^{IF} = 0.3\,V$. Following Equation 17, $\boldsymbol{V^z} = \tilde{\theta} * \boldsymbol{V^y}$ with $\tilde{\theta}(t) = \begin{cases} \delta(t+\tau_+) \\ -\delta(t-\tau_-) \end{cases}$, $\tau_+ = 0$, and $\tau_- = 2.5\,\mu s$. When a $\boldsymbol{V^y}$ pulse is triggered, a $1\,V$ $\boldsymbol{V^z}$ pulse is triggered simultaneously, and a $-1\,V$ $\boldsymbol{V^z}$ pulse is triggered at $2.5\,\mu s$ after the $\boldsymbol{V^y}$ pulse is triggered. $\boldsymbol{V^y}$ pulses triggers actuation pulses $a$ (Figure 32f) with an amplitude of $1\,V$ and a duration of $2.5\,\mu s$ by following $a = \boldsymbol{V_1^y} - \boldsymbol{V_2^y}$, and the system state $s$ is modified by $a$ with $\dot{s} = g^{sa}a + n^s$, where $g^{sa}$ represents a modification coefficient, and $n^s$ represents the random perturbation from environment. When $|s| \geq 12$, it is assumed that the system reaches its boundary, is out of the control, and the self-programming process fails. When the system is modified, the objective function of the system, $F = \frac{1}{2}s^2$ (Figure 32g), is also modified accordingly. The synstor conductances are set to random values with $0 \leq w_{ji} \leq 20\,nS$ before the self-programming process, and the synstor conductances $w_{ji}$ (Figure 32h) is modified by $\boldsymbol{V^x}$ and $\boldsymbol{V^z}$ by following Equation 2, $\dot{\boldsymbol{w}} = \beta\,\boldsymbol{V^z} \otimes \boldsymbol{V^x}$ within the range of $0 \leq w_{ji} \leq 20\,nS$ in the self-programming process. In the simulation shown in Figure 32, $\beta = 3\,nS\,V^{-2}\,s^{-1}$, $g^{sa} = 1\,a.u.$, and $-0.25 \leq n^s \leq 0.25\,a.u.$

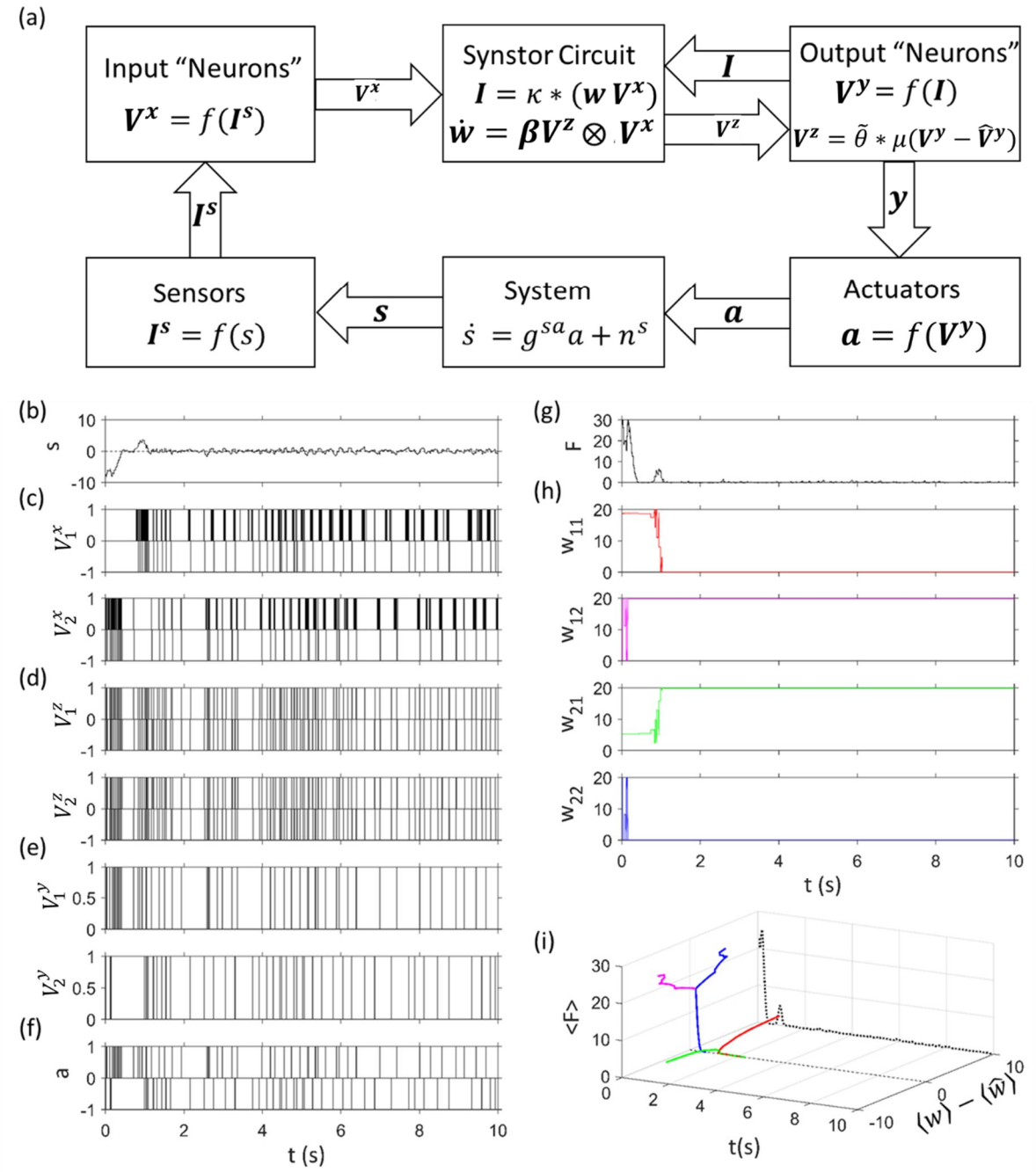


**Figure 32.** Simulation of closed-loop system with SNIC. a) A simulated closed-loop system including a synstor circuit and an external system. In the simulated synstor circuit and system, b) $s$, the state of the system (arbitrary unit), c) $V_1^x$ and $V_2^x$, the input pulses on the input electrodes 1 and 2 (Unit: $V$), d) $V_1^z$ and $V_2^z$, the feedback pulses on the output electrodes 1 and 2 (Unit: $V$), e) $V_1^y$ and $V_2^y$, the output pulses from the output neuron circuits 1 and 2 (Unit: $V$), f) $a$, the

actuation pulses to the system (Unit: $V$), g) $F$, the objective function of the system (arbitrary unit), h) $w_{11}$, $w_{12}$, $w_{21}$, and $w_{22}$, the conductances of the synstors in the circuit (Unit: $nS$), are shown versus time $t$. i) The average objective function, $\langle F \rangle$, is shown as a function of $t$ and the difference between the average synstor conductances and their optimal conductances, $\langle w_{11} \rangle - \langle \widehat{w}_{11} \rangle$ (red solid line), $\langle w_{12} \rangle - \langle \widehat{w}_{12} \rangle$ (magenta solid line), $\langle w_{21} \rangle - \langle \widehat{w}_{21} \rangle$ (green solid line), and $\langle w_{22} \rangle - \langle \widehat{w}_{22} \rangle$ (blue solid line). The black dotted lines show $\langle F \rangle$ changing versus time $t$ in the background.

The dynamic self-programming process shown in Figure 32 is analyzed, and the change of the average objective function $\langle F \rangle$ is shown versus time $t$ and the average synstor conductances $\langle w_{ji} \rangle$ over a moving time window $T$ in Figure 32i. The change of $w_{ji}$ leads to the change of $F$, which is also influenced by the random environment noise $n^s$. When the change of $w_{ji}$ leads to the change of $F$ in a learning period, the covariance between $F$ and $w_{ji}$, $\langle \tilde{F}, \tilde{w}_{ji} \rangle \neq 0$; when $w_{nm}$ does not change beyond the learning period, $\langle \tilde{F}, \tilde{w}_{ji} \rangle = 0$. Learning periods within which $\langle \tilde{F}, \tilde{w}_{ji} \rangle \neq 0$, and the minimal length of moving time window $T$ which satisfies $\langle \dot{F} \rangle \leq 0$ or $\langle \dot{F} \rangle \geq 0$ within each learning period are identified in the analysis. Within a learning period, if $\langle \dot{F} \rangle < 0$, $\langle \widehat{w}_{ji} \rangle$ is identified as $\langle w_{ji} \rangle$ at the end of a learning period when $\langle \tilde{F}, \tilde{w}_{ji} \rangle = 0$, $\langle \dot{F} \rangle = 0$, and $\langle F \rangle$ approaches its minimal equilibrium value $F_e$ ($F_e$ may not be equal to zero due to the environment noise or other synaptic conductances $w_{j'i'}$). After $\langle \widehat{w}_{ji} \rangle$ is identified, $\langle w_{ji} \rangle - \langle \widehat{w}_{ji} \rangle$ and $\langle F \rangle$ within the learning period are shown versus time $t$ in Figure 32i. Within each learning period, $\langle F \rangle$ decreases asymptotically versus time $t$ with $\langle \dot{F} \rangle < 0$, and $\langle w_{ji} \rangle$ is gradually modified toward $\langle \widehat{w}_{ji} \rangle$; at the end of the learning period, $\langle w_{ji} \rangle = \langle \widehat{w}_{ji} \rangle$, $\langle \dot{F} \rangle = 0$, and $\langle F \rangle$ reaches an equilibrium value $F_e$. $\langle F \rangle$ represents a Lyapunov function within each learning period. The system changes dynamically by

the random environment noise. When the system is perturbed, and $F$ is increased from $F_e$ above a threshold value, $\widehat{\boldsymbol{w}}$ is also shifted accordingly, and a learning period will spontaneously be triggered in the self-programming process to modify $\boldsymbol{w}$ toward $\widehat{\boldsymbol{w}}$, and $F$ toward $F_e$ in the dynamically changing environment.

The self-programming processes in the synstor circuit are analyzed statistically by setting the initial synstor conductance matrix $\boldsymbol{w}$ to random values within the range of $0 \leq w_{ji} \leq 20\ nS$, and the system state $s$ within the range of $4 \leq |s| \leq 8\ a.u.$ in multiple simulations. $s$ is also influenced by random environment noise with $0 \leq |n^s| \leq 0.25\ a.u.$ $\boldsymbol{w}$ is modified by following the learning rule, and $\langle \widehat{\boldsymbol{w}} \rangle$ is extrapolated in self-programming processes. The objective function $F$ is modified as the function of $\boldsymbol{w}$ and the random environment noise, and $\langle F \rangle$ is shown versus $\langle w_{ji} \rangle - \langle \widehat{w}_{ji} \rangle$ in each self-programming process of $w_{ji}$ in the multiple simulations in Figure 32. When $w_{ji}$ is set to an initial value which triggers actuations to increase $F$, $w_{ji}$ is modified toward $\widehat{w}_{ji}$ to reduce $\langle F \rangle$ asymptotically, and $\langle F \rangle$ reaches its equilibrium value $F_e$ when $\langle w_{ji} \rangle = \langle \widehat{w}_{ji} \rangle$. When $w_{ji}$ is set to an initial value which triggers actuations to decrease $F$ toward its equilibrium value $F_e$, the initial $w_{ji} \approx \widehat{w}_{ji}$, and $w_{ji}(t) - \widehat{w}_{ji} \approx 0$ is shown as vertical lines in Figure 33. In the circuit, $V_1^x$ input pulses trigger currents $w_{11}V_1^x$ and $w_{21}V_1^x$ via the synstors, and induce $V_1^y$ and $V_2^y$ output pulses, leading to the actuation pulses to increase and decrease $V_1^x$ pulses, respectively; concurrently, the $w_{11}V_1^x$ and $w_{21}V_1^x$ currents also trigger $V_1^z$ and $V_2^z$ output pulses, which encounter $V_1^x$ pulses to decrease $w_{11}$ and increase $w_{21}$, resulting in the reduction of $V_1^x$ and $F$. $V_2^x$ input pulses trigger currents $w_{12}V_2^x$ and $w_{22}V_2^x$ via the synstors, and induce $V_1^y$ and $V_2^y$ output pulses, leading to the actuation pulses to decrease and increase $V_2^x$ pulses, respectively; concurrently, the $w_{12}V_2^x$ and $w_{22}V_2^x$ currents also trigger $V_1^z$ and $V_2^z$ output pulses, which

encounter $V_2^x$ pulses to increase $w_{12}$ and decrease $w_{22}$, resulting in the reduction of $V_2^x$ and $F$. As shown in Figure 33, the synstor conductance matrix $\boldsymbol{w}$ does not need to be pre-programmed, and can be spontaneously modified toward $\widehat{\boldsymbol{w}}$, decreasing $\langle F \rangle$ toward $F_e$ in the 100 simulations of the self-programming processes.

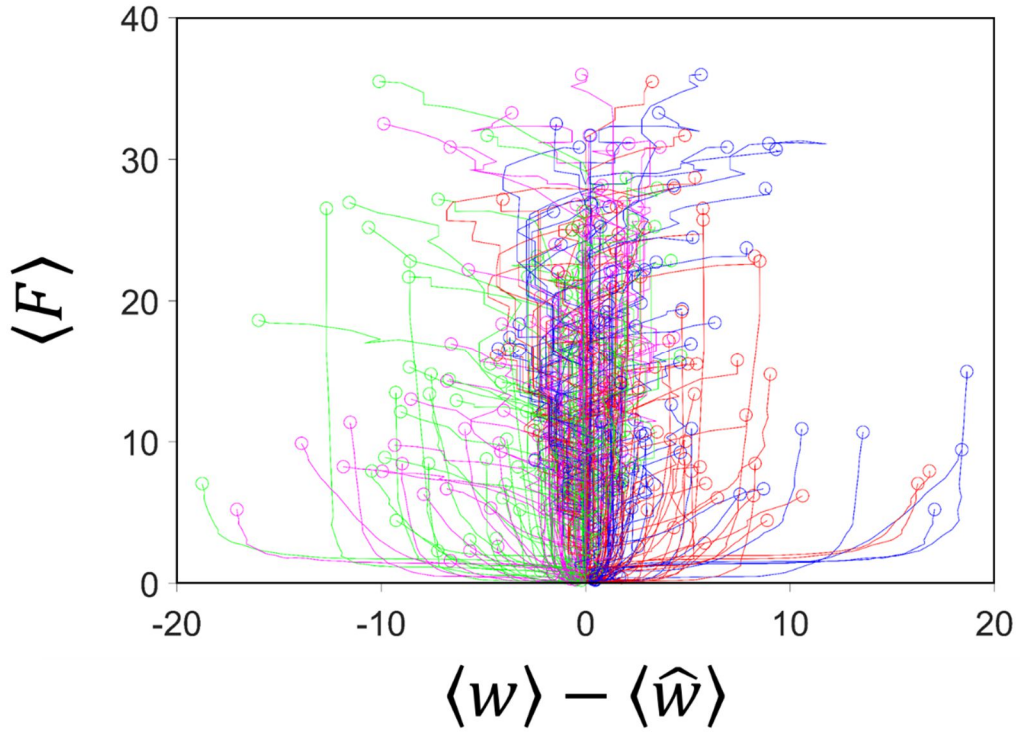


**Figure 33.** Average objective function versus average conductance error. The average objective function, $\langle F \rangle$ (arbitrary unit), is plotted versus the differences between the average synstor conductances and their optimal conductances, $\langle w_{11} \rangle - \langle \widehat{w}_{11} \rangle$ (red), $\langle w_{12} \rangle - \langle \widehat{w}_{12} \rangle$ (magenta), $\langle w_{21} \rangle - \langle \widehat{w}_{21} \rangle$ (green), and $\langle w_{22} \rangle - \langle \widehat{w}_{22} \rangle$ (blue) in the unit of $nS$ with the initial differences marked by circles in multiple self-programming processes.

## 5.6. Stability and Equilibrium of Self-Programming

In a synstor circuit and system, $\langle \dot{F} \rangle = -\eta \langle F \rangle + \delta F < 0$ when $\eta \langle F \rangle > \delta F$, which is the condition for the stability of the circuit and system. Under this condition, its objective function $\langle F \rangle$ is asymptotically decreased when $\langle \boldsymbol{w} \rangle$ is modified toward $\langle \widehat{\boldsymbol{w}} \rangle$ in a self-programming process. The environmental random perturbation causes the increase of $F$, and $\delta F$ due to the perturbation is proportional to the magnitude of the $\eta$ is increased versus the conductance modification coefficient $\alpha$ defined in Equation 2, therefore $\eta \langle F \rangle$ is increased versus $\alpha$, and satisfies the stability condition $\eta \langle F \rangle > \delta F$ when $\alpha$ is increased above a critical value. The closed-loop synstor circuit and system are simulated under the conditions described above with a random perturbation $n^s$ and $\beta = 0.1, 0.8, 1$ or $10\ nS\ V^{-2}s^{-1}$. $\langle F \rangle$ is shown versus the differences between the average synstor conductances and their optimal conductances, $\langle w_{ji} \rangle - \langle \widehat{w}_{ji} \rangle$, in each self-programming process of $w_{ji}$ in the multiple simulations in Figure 34. When $\beta = 0.1\ nS\ V^{-2}s^{-1}$ and the objective function $\langle F \rangle \gtrsim 20$ at the initial stage of the self-programming process (Figure 34a), $\eta \langle F \rangle < \delta F$ and $\langle \dot{F} \rangle > 0$. Under these conditions, $F$ is not stable, and gradually increased to its upper limit value ($\sim 72$). When $\beta = 0.8\ nS\ V^{-2}s^{-1}$, and the objective function $\langle F \rangle \gtrsim 40$ at the initial stage of the self-programming process (Figure 34b), $\eta \langle F \rangle < \delta F$, $\langle \dot{F} \rangle > 0$, $F$ is not stable. When $\beta = 1\ nS\ V^{-2}s^{-1}$ (Figure 34c) and $\beta = 10\ nS\ V^{-2}s^{-1}$ (Figure 34d), $\eta \langle F \rangle > \delta F$, $\langle \dot{F} \rangle < 0$, $\langle F \rangle$ is stable, and gradually decreased toward its optimal value ($\sim 0$) in self-programming processes.

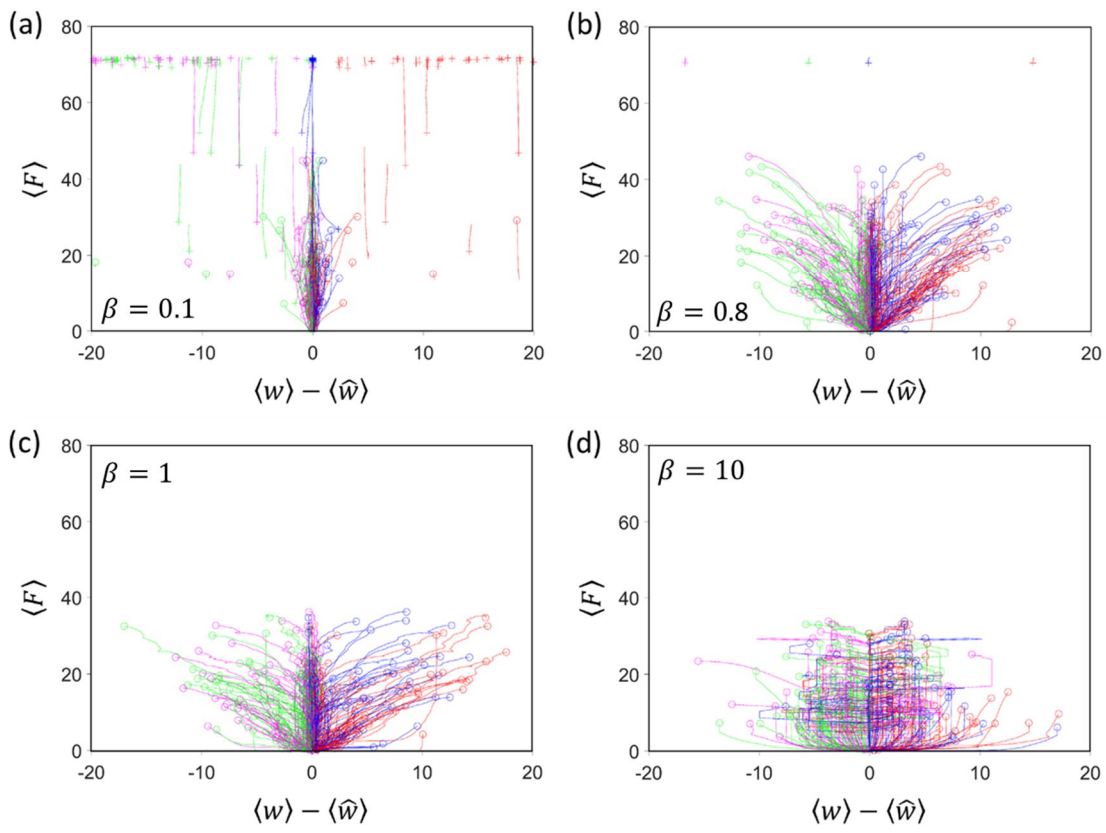


**Figure 34.** Dependence of mean objective function and conductance errors on $\alpha$. The average objective function, $\langle F \rangle$ (arbitrary unit), is plotted versus the differences between the average synstor conductances and their optimal conductances, $\langle w_{11} \rangle - \langle \hat{w}_{11} \rangle$ (red), $\langle w_{12} \rangle - \langle \hat{w}_{12} \rangle$ (magenta), $\langle w_{21} \rangle - \langle \hat{w}_{21} \rangle$ (green), and $\langle w_{22} \rangle - \langle \hat{w}_{22} \rangle$ (blue) in the unit of $nS$ with the initial differences marked by circles for $\langle \dot{F} \rangle < 0$ and crosses for $\langle \dot{F} \rangle > 0$ in multiple self-programming processes under the conductance modification coefficient (a) $\beta = 0.1$, (b) $\beta = 0.8$, (c) $\beta = 1$, and (d) $\beta = 10$.

When $\eta \langle F \rangle > \delta F$, $\langle \dot{F} \rangle < 0$, the circuit and system are table, and $\langle F \rangle$ is asymptotically decreased to an equilibrium value $F_e$ in a self-programming process. Under the equilibrium

condition, $\langle \dot{F} \rangle = -\eta F_e + \delta F = 0$ and $\langle \boldsymbol{w} \rangle = \langle \widehat{\boldsymbol{w}} \rangle$, thus $F_e = \delta F/\eta$. The environmental random perturbation induces $\langle \dot{F} \rangle$, which is proportional to the $\delta F$ is a

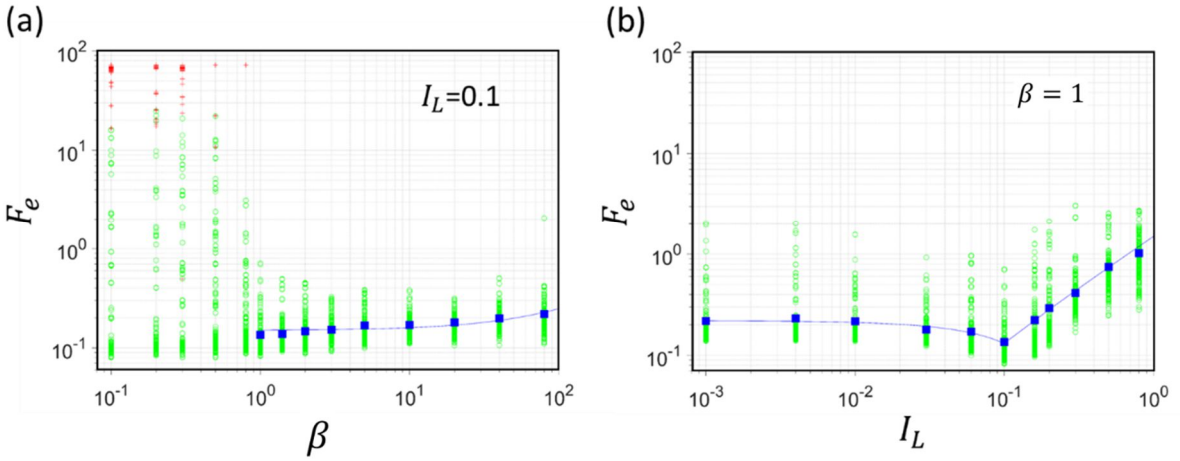


**Figure 35.** Equilibrium objective function versus conductance modification coefficient. The equilibrium objective function, $F_e$ (arbitrary unit), is plotted versus (a) the conductance modification coefficient $\beta$ in the unit of $nS\,V^{-2}s^{-1}$ under $I_L = 0.1\,nA$, and (b) the leakage current in the integrate-and-fire neuron circuit $I_L$ in the unit of $nA$ under $\beta = 1\,nS\,V^{-2}s^{-1}$. The green open circles mark $F_e$, the blue filled squares mark average $F_e$, and the red crosses mark the final $F$ under non-equilibrium condition. The average $F_e$ is fitted by $\langle F_e \rangle = g_{F\beta}\beta + F_e^0$ with $g_{F\beta} \approx 10^{-3}$ (a. u.) and $F_e^0 \approx 0.15$ (a. u.) in (a), and fitted by $\langle F_e \rangle = g_{FI}(I_L - I_L^0)$ with $g_{FI} \approx -0.89$ (a. u.) and $I_L^0 \approx 0.67\,nA$ for $1\,pA \leq I_L \leq 0.1\,nA$ and $g_{FI} \approx 1.5$ (a. u.) and $I_L^0 \approx 0.01\,nA$ for $0.1\,nA \leq I_L \leq 1\,nA$ in (b).

The stability can also be improved by optimizing other parameters such as the firing rates and amplitudes of the feedback pulses, $\boldsymbol{V^z}$, the upper limit of synstor conductances, transfer functions between the system and synstor circuit, $g^{\boldsymbol{sa}}$, the gain between the firing rates of output

pulses versus input currents, $g_n^{V^y/I}$, the capacitance of the capacitors, $C_{IF}$, the leakage currents, $I_L$, and the threshold voltages, $V_{th}^{IF}$, and saturated firing rates of output pulses in neuron circuits.

In a synstor circuit and system, $\langle \dot{F} \rangle = -\eta \langle F \rangle + \delta F < 0$ when $\eta \langle F \rangle > \delta F$, which is the condition for the stability of the circuit and system. Under this condition, its objective function $\langle F \rangle$ is asymptotically decreased toward its equilibrium value $F_e = \delta F / \eta$ when $\langle \boldsymbol{w} \rangle$ is modified toward $\langle \widehat{\boldsymbol{w}} \rangle$ in a self-programming process.

The solution of Equation 5 or Equation 7 gives,

$$F(t) = F(0) e^{-\int_0^t \eta(\tau) d\tau} + F^e * e^{-\int_0^t \eta(\tau) d\tau} \qquad 25$$

where $F(0)$ represents the initial value of $F$ based on $\boldsymbol{w}$ at the beginning of the self-programming process, and $F^e * e^{-\int_0^t \eta(\tau) d\tau}$ represents the temporal convolution between $F^e$ and $e^{-\int_0^t \eta(\tau) d\tau}$ with $F^e * e^{-\int_0^t \eta(\tau) d\tau} = \int_0^t F^e(t') \, e^{-\int_0^{t-t'} \eta(\tau) d\tau} dt'$.

# 6. Morphing Wing Learning Experiment

Note that in Section 6, the variable names are the same as in Section 0, with the following exceptions. The objective function is represented as $E$, and the variable $F$ now represents lift force.

## 6.1. Experimental

### 6.1.1. System

We fabricated a crossbar circuit of synstors, each having a transistor structure including Al input, output, and reference electrodes, a semiconducting carbon nanotube (CNT) channel, and a $HfO_2/TiO_2/HfO_2$ charge trap stack, by a process described in Section 2.3 (Figure 8). The reference electrode is grounded, and creates a voltage drop across the oxide layers when voltages are applied simultaneously on the input and output electrodes, causing nonvolatile analog change in the charge concentration of the $TiO_2$ layer by electron hopping through the $HfO_2$ barrier layer. The electric field generated by charges in the $TiO_2$ layer modulates the conductivity of the channel and conductance of the device. To read the conductance state without disrupting the nonvolatile memory, a negative voltage is applied on the input with grounded output, voltage drops across the Schottky barrier at the Al-CNT interface and current flows through the channel to the output (Figure 5). Negligible changes of conductance were measured after read voltage sweeps and pulses (Figure 11). The synstor crossbar was connected externally to integrate-and-fire neuron circuits based on the Hodgkin-Huxley model[49], to form a single layer network SNIC.

The synstor-based SNIC processed the lift force signal, modified the synstor conductances, and adjusted the actuator to modify the shape of the wing by real-time self-programming (Figure 36a). To perform inference, the lift force sensor signal $F$ was processed by input neurons $N_i^x$. Input pulse rates $X_i$ were defined by rectified logistic functions $X(t) = \frac{sgn(E_1)\cdot X_{max}}{1+e^{-a[|E_1|-c]}}$, $X_1 = \max\{X, 0\}$, and $X_2 = -\min\{X, 0\}$ (Figure 37), where $X_{max}$, $a$, and $c$ are parameters, and $E_1 = F - F_{target}$. Input neurons applied voltage pulses signals $V_i^x$ along the input lines to input electrodes of the synstors, generating output currents $I_j$, where $i = 1, 2$ and $j = 1, 2$ (Figure 5). Each synstor can be treated as an RC circuit with the semiconducting CNT channel resistance in series with the capacitance between the channel and reference electrode,

which enables convolutional signal processing. Applying Kirchhoff's law along each column of synstors, the synstor crossbar circuit generated output currents

$$I_j = \sum_i \kappa_{ji}(t) \circledast \left(w_{ji} V_i^x\right) \qquad 26$$

where $\kappa_{ji}(t)$ represents a temporal kernel function (see Section 2.4.4), $w_{ij}$ represents the matrix of synstor weights (conductances) and $\circledast$ represents the temporal convolution.

The summed currents $I_j$ were integrated by integrate-and-fire neuron circuits $N_j^y$ with logistic activation functions[50,51]. The neurons generated output voltage pulses $V_j^y$ with output pulse rates

$$Y_j = \frac{Y_{max}}{1 + e^{-\chi(\bar{I}_j - I_0)}} \qquad 27$$

where $\chi$ and $I_0$ are circuit parameters (Figure 39). A logistic activation function was used to damp oscillations around the optimum of the objective function. Simultaneously, the neurons backpropagated feedback voltage pulses $V_j^z$ (Figure 40a, red pulses) to modify the conductances to analog values via Hebbian learning[52] in real-time. The conductance modification rate was

$$\dot{\mathbf{w}} = \widehat{\boldsymbol{\beta}} \odot \boldsymbol{V^x} \otimes \boldsymbol{V^z} \qquad 28$$

where $\dot{\mathbf{w}} = \dot{w}_{ij}$, $\widehat{\boldsymbol{\beta}} = \beta_{ij}$ denotes the learning coefficient matrix, and $\odot$ denotes the entry-wise or Hadamard product, and $\otimes$ denotes the vector outer product. When an input pulse and a

feedback pulse were applied simultaneously to the input and feedback electrodes of a synstor, $\dot{w}_{ij} \neq 0$, and the conductance was modified (Figure 5). Otherwise, when $V_i^x = 0$ and/or $V_j^z = 0$, then $\dot{w}_{ij} \approx 0$; the conductance remained nonvolatile. This was verified by applying combinations of input and feedback voltage pulse amplitudes and measuring the conductance change (Figure 11). By using the same voltage amplitudes for inference and learning, the synstor circuit did not require varied input and feedback voltage amplitudes, nor a programmed reference voltage to control the output current or tuning voltage, distinguishing it from gate-controlled transistors, 1-transistor 1-memristor (1T1R), or more complex unit cells[53,54].

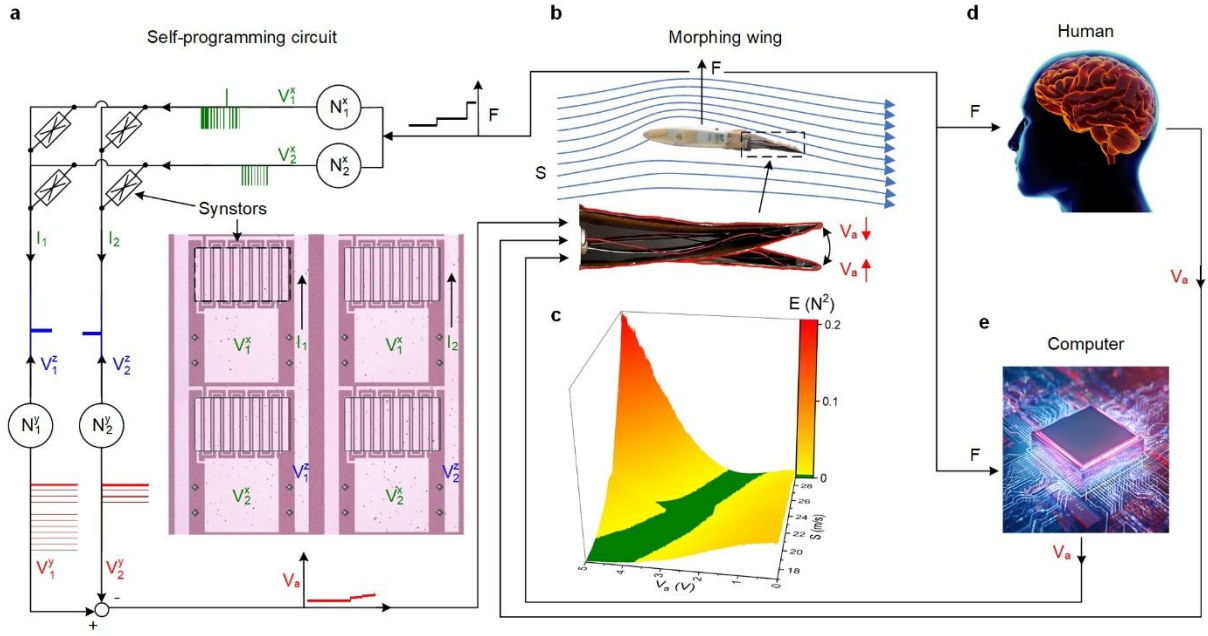


**Figure 36.** Self-programming neuromorphic circuit for morphing wing. a) A schematic of a $2 \times 2$ crossbar circuit of synstors with input and output neurons. Input neurons $N_i^x$ apply voltage pulses $V_i^x$ as a function of the lift force $F$. Output neurons $N_j^z$ generate: feedback pulses $V_j^z$; and output pulses $V_j^y$. The output voltage pulses $V_j^y$ generate an actuation voltage $V_a$. An inset shows a microscope image of a $2 \times 2$ crossbar circuit of synstors. b) A photo of the morphing wing

with lift force $F$, wind speed $S$, and illustrated streamlines. Below, a detail photo of the morphing wing with a piezoelectric actuator, deflected upward at $V_a = 0$, and deflected downward at $V_a = 5\ V$. c) The objective function $E$ plotted as a function of the actuation voltage $V_a$ and the wind speed $S$. d, e, Illustrations of a human and a computer controller, which were tested in control experiments to compare performance with the SNIC.

The morphing wing used a voltage-controlled macrofiber composite (MFC) piezoactuator that modified the wing camber by bending the trailing edge (Figure 36b)[55,56]. It was mounted inside a wind tunnel with a variable rotation speed fan to vary the wind speed. The wind speed $S$ was measured with a pitot tube, and the lift force was measured with a load cell force sensor. An interface changed the actuation control voltage proportional to the difference of output pulse rates $Y_1$ and $Y_2$. The lift force $F$ was dependent on $V_a$ and the wind speed $S$. By measuring $F$ as $V_a$ was swept slowly at various constant $S$ conditions, and the quadratic objective function $E = \frac{1}{2}(E_1)^2$ was calculated (Figure 36c), where $E_1 = F - F_{target}$. The objective of the circuit was to minimize $E$ by modifying the actuation voltage $V_a$ such that $F(V_a, S) = F_{target}$, where $\arg\min_F E(F) = F_{target}$ and $E(F_{target}) = 0$. A region of small $E$ near the optimum of $E = 0$ is displayed with green color.

The task of modifying the actuation voltage $V_a$ in real-time to minimize the lift force error $E_1$ was also performed by 13 different untrained human participants (Figure 36d). The lift force error $E_1$ averaged by a 40 ms time window was displayed to the participants using an NI LabView graphical user interface. The human subjects controlled the actuator by pressing two keyboard keys, randomly mapped to each polarity of change. The magnitude of the actuation voltage change by each keystroke was proportional its duration. Participants wore noise

cancelling headphones to suppress the sound of the wind tunnel fan to prevent inference of the wind speed from the sound. The magnitude of the actuation voltage change by each keystroke was proportional its duration by $|\Delta V_a| = \gamma_h \Delta t$, where $\gamma_h = 31\ mV/ms$ and $\Delta t$ was the duration pressed. The durations output pulses $V_{y,j}$ were equal to the durations of key presses, where $\Delta V_a = \gamma_h^* \int_0^{T_a} \left(V_1^y - V_2^y\right) dt = \gamma_h^* V_a^y (\Delta t_1 - \Delta t_2) = \gamma_h (\Delta t_1 - \Delta t_2)$, $\gamma_h^*$ is a constant, $T_a = 40\ ms$ is the update period, and $V_a^y = 1.5\ V$ is the output pulse amplitude.

Similarly, a computer-based controller (Figure 36e) was used to minimize the lift force error as a comparison with the SNIC and human participants. The computer controller used proportional-integral-derivative (PID)[57] control implemented in National Instruments LabVIEW software (see Section 5.6.3). The controller was tested with various combinations of PID gains $K_p$, $K_i$, and $K_d$ under static wind conditions to find the settings with minimal steady state error in lift force (Figure 38). The PID controller was pre-programmed with these coefficient values for the varied wind speed tests, and its performance was compared with the SNIC and human subjects.

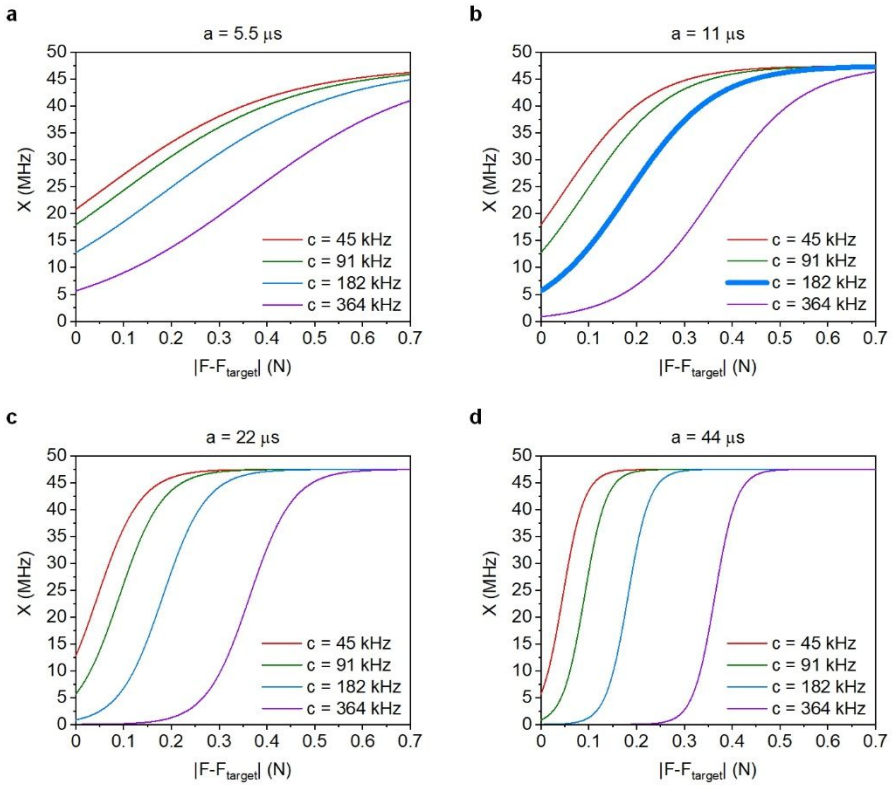


**Figure 37.** Characteristics of artificial input neuron. The input pulse firing rate $X$ applied to the SNIC versus $|F - F_{target}|$, where $X = \frac{X_{max}}{1+e^{-a(|F-F_{target}|-c)}}$ for $c = 45\ kHz$ (red), $c = 91\ kHz$ (green), $c = 182\ kHz$ (blue), and $c = 364\ kHz$ (purple), for a) $a = 5.5\ \mu s$, b) $a = 11\ \mu s$, c, $a = 22\ \mu s$, and d, $d = 44\ \mu s$. The optimal condition, which minimized the mean squared error of lift force $\bar{E}$ during the morphing wing experiments is $c = 182\ kHz$, $a = 11\ \mu s$ (thick blue line).

## 5.6.1. Morphing Wing

The morphing wing was fabricated by the Adaptive Intelligent Multifunctional Structures laboratory at University of Michigan, Ann Arbor[55,56]. The morphing wing is a wing section with a morphing trailing edge, which uses a macrofiber composite (MFC) piezoelectric actuator and a flexure box mechanism to modify the camber of the trailing edge. The actuator has a 3D printed

elastomeric honeycomb skin for tailored stiffness, and the piezoelectric mechanism allows for fast response time. This morphing design has applications in stall recovery during wind gusts, optimizing the lift distribution to increase aerodynamic efficiency, and reducing turbulence. The design is scalable to multiple piezoelectric actuators along the spanwise edge (spanwise morphing trail edge) to achieve continuous shape change. An analog voltage module (National Instruments, NI-9264) generated the analog actuation voltage signal $V_a$ which was amplified by a high-voltage driver (Avid LLC, AVID-EHV-MFC.B2) to a range of (-0.5 to 1.5 kV) to generate axial strain to bend the wing, modifying its camber.

## 5.6.2. Wind Tunnel

The morphing wing was tested in a laminar flow, open circuit Aerolab wind tunnel with a 24x24 inch test section. The wind tunnel fan was controlled by a high voltage driver (ABB, ACS550-01-046A-2 AC). The wind speed $S$ was measured with a pitot tube and air velocity transducer (TSI, 8455). The lift force $F$ was measured using a force-torque multi-axis load cell (JR3, 30E12A-4-I40-EF 40N3.1S) attached to the ¼ inch morphing wing mounting shaft. The wind velocity and lift force voltage signals $(S, F)$ from the sensors were read by a voltage I/O device (National Instruments, PCIe 6353 Multifunction I/O Device), and processed in NI LabVIEW.

The setpoints for the wind speed were generated using a pseudorandom number generator following a Gaussian distribution $N(\mu_S, \sigma_S^2)$, where $\mu_S = 22\ m/s$ and $\sigma_S = 2.8\ m/s$, and were changed with period $N(\mu_{T_s}, \sigma_{T_s}^2)$, where $\mu_{T_s} = 8\ s$ and $\sigma_{T_s} = 1.33\ s$. The fan voltage driver (ABB, ACS550-01-046A-2 AC) used an internal PID controller with fixed coefficients to modify the speed toward the setpoints. All experiments for controlling the morphing wing with varied wind

speed used the same sequence of randomly generated wind speeds and periods to reduce statistical variation of performance due to variations in wind speed signal. The wind speed was initialized at steady state at the first setpoint $S = 26.6\, m/s$. The range of speed was $S = 17.3\, m/s$ to $S = 28.7\, m/s$.

### 5.6.3. PID Controller

The computer-based PID controller used control function $u(t) = K_p e(t) + K_i \int_0^t e(t')\, dt' + K_d \frac{de(t)}{dt}$, where $u = V_a$, and $e = E = F - F_{target}$. To choose PID coefficients the lift force $F$ was first measured under constant wind speed $S = 28.7\, m/s$ as the actuator was controlled by the PID controller with combinations of coefficients $K_p$, $K_i$, and $K_d$ with $\begin{bmatrix} K_p \\ K_i \\ K_d \end{bmatrix} = \begin{bmatrix} 10^{-4\pm k} \\ 10^{-3\pm k} \\ 10^{-2\pm k} \end{bmatrix}$ for $k = 0, 1, 2$. The condition with the minimum mean squared error under steady-state conditions was $K_p = 10^{-5}$, $K_i = 10^{-4}$, and $K_d = 10^{-3}$, and is shown with green color in Figure 38.

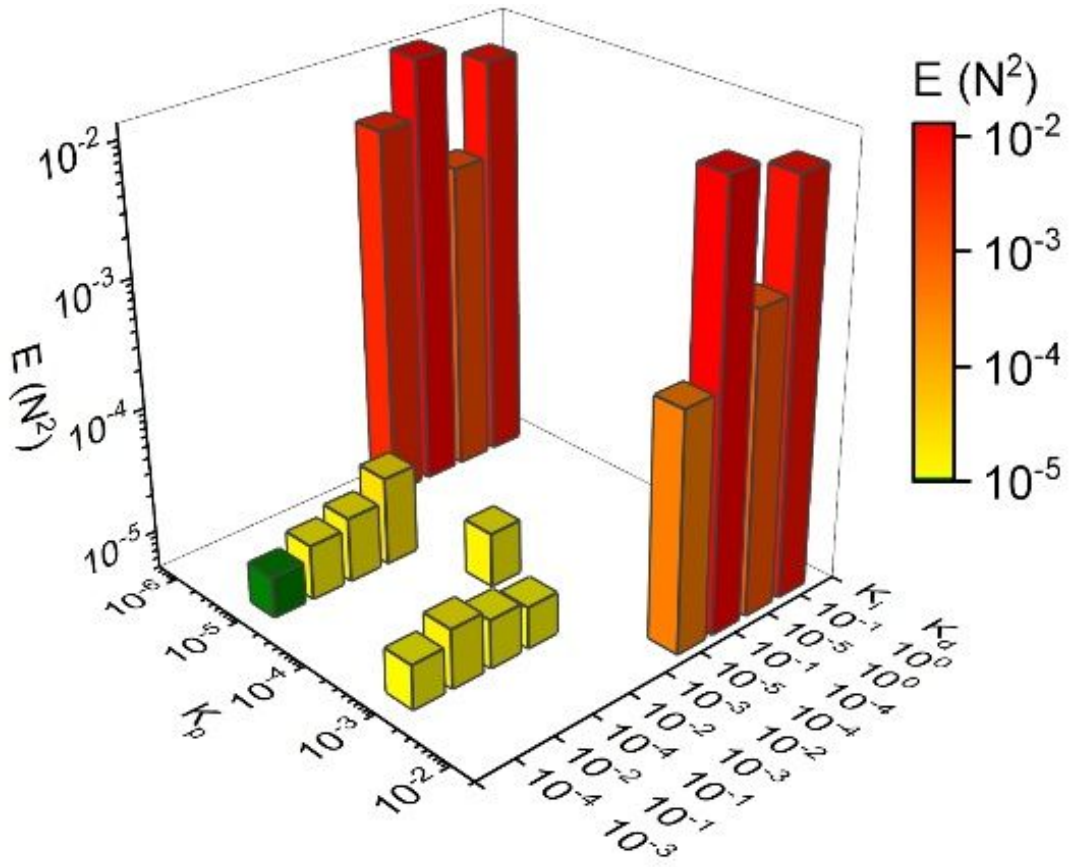


**Figure 38.** Lift force error for computer controller. Lift force error $E$ on a morphing wing in a wind tunnel under constant speed $S \approx 29\ m/s$, versus PID coefficients $K_p$, $K_i$, and $K_d$ of a computer-based PID controller of the actuation voltage $V_a$. The optimal condition with minimal error $E$, and $K_p = 10^{-5}$, $K_i = 10^{-4}$, and $K_d = 10^{-3}$ is shown with green color.

## 5.6.4. Electrical Hardware

Electrical voltage pulse signals $V_i^x$, $V_j^z$, were generated by function generators (Agilent, 33250A 80MHz) with voltage amplitudes $V_+ = 1.5\ V$, $V_- = -1.75\ V$, 10 ns pulse duration, and 50 MHz frequency. This synchronized input and feedback pulses, to prevent phase shift between them. The transmission of voltage pulse signals $V_i^x$, $V_j^z$ from the function generators to the synstor circuit was gated by switches (Maxim, MAX383), which were activated by a digital input/output

voltage module (National Instruments, 9403E Digital Input/Output). Output voltage signals $V_j^y$ were also read by the digital input/output voltage module. Input, feedback, actuation, and reference voltage signals $V_i^x$, $V_j^z$, $V_a$, and $V_r$ were measured by an analog input module (National Instruments, 9205 Analog Input).

A custom spring-loaded pogo-pin socket (Ironwood Electronics) contacted Al pads on the synstor chip to connect to input, feedback, and reference electrodes. A custom printed circuit board (PCB) adapter (Figure 20), fabricated by Ironwood Electronics),connected the pogo-pin socket to a male PGA socket connected to a female PGA socket (3M, 200-6313-9UN-1900) on the custom interface and integrate-and-fire PCB (Express PCB).

## 5.6.5. Integrate-and-Fire Circuit

A synstor was connected to an integrate-and-fire neuron circuit to measure their transfer characteristics (Figure 39). The synstor conductance was increased by applying by $9.5 \times 10^5$ negative coincident input and feedback pulses of amplitude $V^x = V^z = -1.75\,V$. Then, square voltage waves with constant pulse duration $t_d$ and varied pulse rate $X$ were applied to the input electrode of a synstor while the output voltage pulse rate $Y$ of the neuron circuit was recorded (red circles). This process was repeated after tuning to a decreased conductance state by applying $9.5 \times 10^5$ positive coincident pulses of amplitude $V^x = V^z = 1.5\,V$ (blue circles). The voltage pulse rate $Y$ was approximated as a logistic function of the average current $I$, thus $Y = \frac{Y_{max}}{1+e^{-\chi*(\bar{I}-I_0)}}$, where $Y_{max}$, $\chi$ and $I_0$ are circuit parameters. With square input pulses, the average output current of a synstor is $\bar{I} = w\overline{V_x} = wV_{x,a}t_dX$, where $V_{x,a}$ is the voltage amplitude, and $w$, $t_d$, and $X$ are constant during the averaging window. Then, $Y = \frac{Y_{max}}{1+e^{-\chi*(wV_{x,a}t_dX-I_0)}} = \frac{Y_{max}}{1+e^{-a_y*(X-c_y)}}$ where $a_y$,

and $c_y$ are parameters based on the neuron circuit and synstor properties. The parameters $Y_{max}$, $a_y$, and $c_y$ were fit for the high conductance (red line) and low conductance (blue line) cases. In the low conductance case, the output pulse rate $Y$ is identically zero because $\bar{I}$ remains below the threshold current $I_0$ for the range of $X$. In the high conductance case, the output pulse rate $Y$ saturates at $Y \approx 6\ kHz$.

The actuation voltage time derivative is defined as $\dot{V}_a = \alpha(Y_1 - Y_2)$. The input neurons generate input pulses with pulse rate $X(t) = \frac{sgn(E_1)X_{max}}{1+e^{-a[|E_1|-c]}}$, $X_1 = \max\{X, 0\}$, and $X_2 = -\min\{X, 0\}$. Since both the logistic and rectifying functions are nondecreasing as functions of their inputs, $X$ is nondecreasing as a function of lift force error magnitude $\left|F(t) - F_{target}\right|$. Output pulse rates $Y_1$ and $Y_2$ are logistic, and therefore nondecreasing as a function of $X$. The actuation voltage rate was $\dot{V}_a = \alpha(Y_1 - Y_2)$, where $\alpha$ was a constant. Then, the magnitude of actuation voltage change $\left|\dot{V}_a\right|$ is nondecreasing as a function of error magnitude $\left|F(t) - F_{target}\right|$. This property slows the actuation change and learning rate on average as the lift force $F$ approaches its target $\left|F_{target}\right|$. In minimizing a convex objective function a variable step size or learning rate allows the circuit to converge more quickly, while reducing oscillations around the optimum[58].

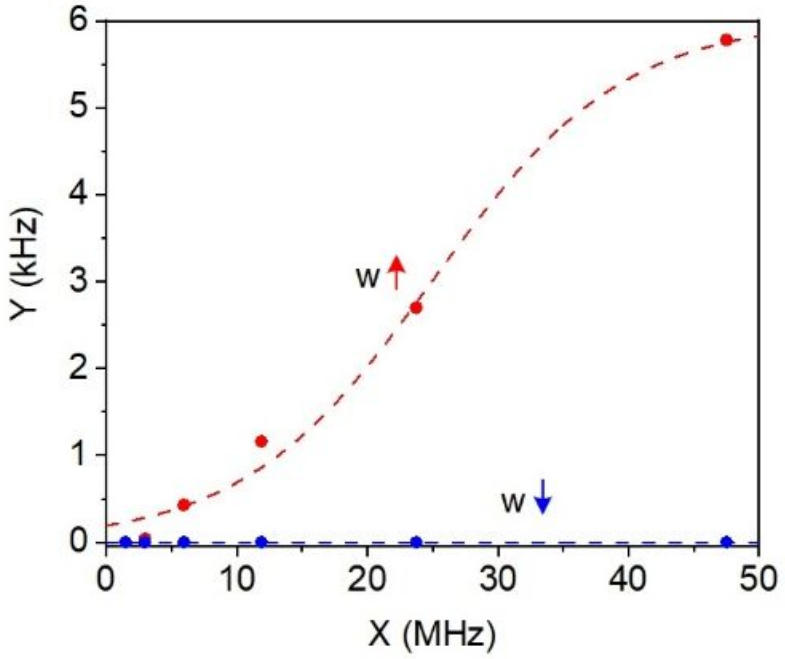


**Figure 39.** Input/output characteristics of a synstor connected to a neuron circuit.

## 5.7. Results

Figure 2a shows results from a 5-minute long experiment from a SNIC under dynamic wind speed conditions with the input neuron settings of $c = 182\ kHz$, $a = 11\ \mu s$ (Figure 37). The setpoints and periods of the wind speed $S$ were varied randomly with normal distributions (Figure 40a, black curve). The lift force $F$ (Figure 40a, green curve) was initialized with large deviation from its target $F_{target} = 0.3\ N$ by initializing the actuation voltage (Figure 40a, red curve) to $V_a = 5\ V$, far from its optimum at the initial speed. Prior to the start of each SNIC learning experiment, the conductances were initialized by applying negative coincident pulses to increase conductances $w_{i,1}$, and positive coincident pulses to decrease conductances $w_{i,2}$. The learning process modified $V_a$ to set $F \approx F_{target}$. As the wind speed varied, the lift force remained stably near its target value by adjusting $V_a$ toward its own dynamic optimum value. Signal processing was performed using input voltage pulses $V_i^x$ (Figure 40a, green pulses). Inputs $X_1$

and $X_2$ were mutually exclusive, and $X_1$ was active when $F$ was larger than the target $F_{target}$, and $X_2$ was active when $F$ was below the target.

Complimentary trains of feedback pulses (Figure 40a, blue pulses, detail plot) of amplitude $V_1^z = 1.5\ V$ for a duration of $20\ ms$ and then $V_2^z = -1.75\ V$ for $20\ ms$ decreased conductances $w_{i,1}$ and increased conductances $w_{i,2}$. When positive feedback pulses were applied, the amplitude of input pulses was $1.5\ V$, and was $-1.75\ V$ otherwise, ensuring that coincident input and feedback pulses had the same amplitude. The input and feedback pulses were generated by the same two positive and negative voltage generators, gated by switches, which prevented phase shifts between them. The coincident pulses increased output pulse rate $Y_2$ relative to $Y_1$, decreasing $V_a$. As $V_a$ decreased, the error $E_1$ decreased, and input pulse rate $X_1$ decreased. This reduced both $Y_1$ and $Y_2$, decreasing the magnitude of $\dot{V}_a$, so that $V_a$ changed more slowly as it approached the optimum, and there was a small overshoot. Feedback pulses were not applied during a period of $\sim 600\ ms$, mainly to account for system delay. Complimentary trains of feedback pulses of amplitude $V_2^z = 1.5\ V$ and then $V_1^z = -1.75\ V$ decreased conductances $w_{i,2}$ and increased conductances $w_{i,1}$. During the $20\ ms$ time windows that feedback pulses were applied, the pulse rates $Z_j$ were logistic functions of the lift force error $E_1$, and updated every $1200\ ms$. This increased $V_a$, but by a smaller amount than magnitude of the previous change $600\ ms$ earlier, because the input pulse rates were lower, since $E_1$ was smaller. This process repeated for the duration of the experiment, as the optimum $V_a$ changed randomly due to wind speed variation.

Figure 40b displays data for a learning experiment with a human controller under similar experimental conditions. The results show that the average errors $\mu_{\bar{E}}$ were larger for the human and PID controller than the synstor circuit average errors (**Table 1**). Figure 40c displays data for

a computer-based PID controller. Under the dynamic speed conditions, the PID controller with gains optimized under static speed conditions, has more precision than the SNIC and human controllers, but less accuracy due to insensitivity to error around the target lift force. In all cases, the initial actuation and the speed settings were identical to those used in the SNIC trials.

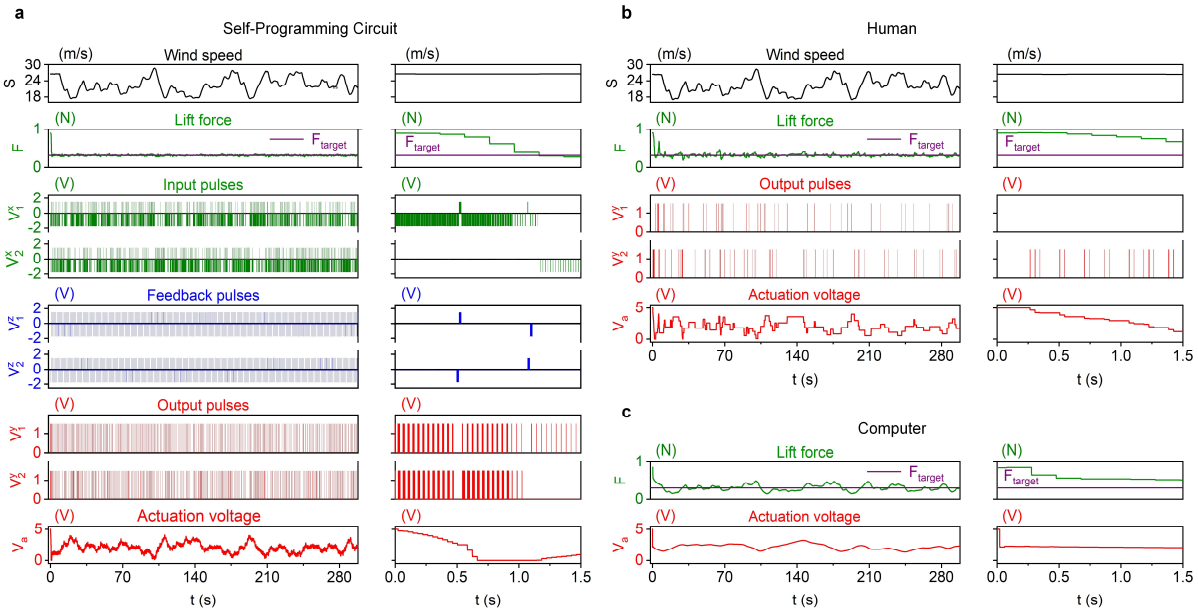


**Figure 40.** Optimization of lift force on a morphing wing. a) The wind speed $S$, lift force $F$ and target $F_{target}$, input voltage pulses $V_i^x$, feedback voltage pulses $V_j^z$, output voltage pulses $V_j^y$, and actuation voltage $V_a$ plotted versus time for one SNIC experiment with the $a, c$ settings that minimize the objective function $E$. Plots of each signal for the duration of the experiment are shown to the left, and insets for the first 1.5 seconds are on the right. b) The wind speed $S$, lift force $F$ and target lift force $F_{target}$, output voltage pulses $V_j^y$, and actuation voltage $V_a$ versus time for one human subject experiment. c) The lift force $F$ and target $F_{target}$ and actuation voltage $V_a$ versus time for a computer controller.

## 5.8. Analysis

Data for the SNIC, Human, and Computer experiments shown in Figure 40 were analyzed to derive the average objective function $\langle E \rangle$, its time derivative $\frac{d\langle E \rangle}{dt}$, conductance $\langle w \rangle$, and its equilibrium value $\langle \hat{w} \rangle$ when $\frac{d^2\langle E \rangle}{dt^2} < 0$ and $\frac{d\langle E \rangle}{dt} = 0$. The average $\langle w \rangle$ is defined as a dimensionless parameter representing the transfer function of the input, lift force error $F - \hat{F}$ to the output $V_a$. The SNIC is initialized with arbitrary device conductances, and modifies them via Hebbian learning to drive the objective function toward its minimum during the dynamic wind speed conditions. The Human similarly has the ability to optimize its conductances. However the logic of the Computer is fixed by its constant PID gains, $K_p$, $K_i$, and $K_d$, and cannot efficiently minimize the objective function.

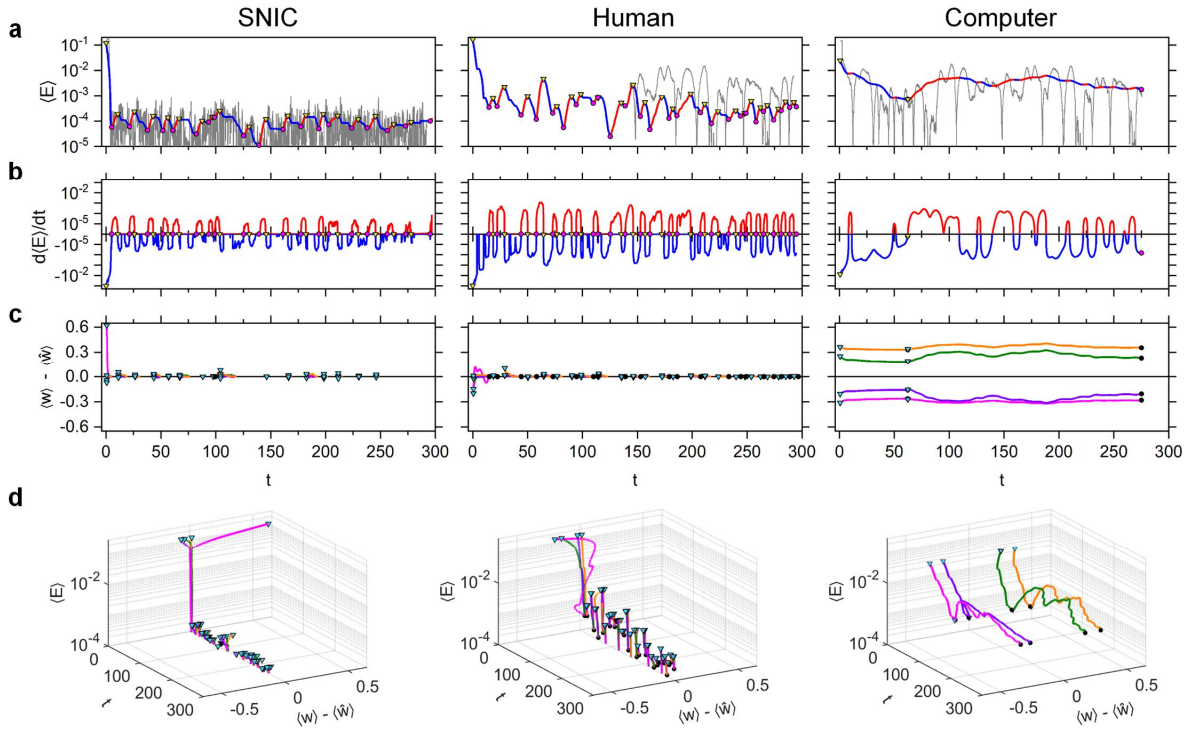


**Figure 41.** The average objective function $\langle E \rangle$ and conductance error $\langle w \rangle - \langle \hat{w} \rangle$ for SNIC, Human, and Computer. **a**) $\langle E \rangle$ versus time, where blue color is $\langle E \rangle$ for $\frac{d\langle E \rangle}{dt} < 0$, red color is $\langle E \rangle$ for $\frac{d\langle E \rangle}{dt} > 0$, yellow triangles are $\frac{d^2\langle E \rangle}{dt^2} > 0$ and $\frac{d\langle E \rangle}{dt} = 0$, pink circles are $\langle E \rangle$ $\frac{d^2\langle E \rangle}{dt^2} < 0$ and $\frac{d\langle E \rangle}{dt} =$

0, and gray is raw $E$. **b**) $\frac{d\langle E\rangle}{dt}$ versus time, with same color scheme as in **a**). **c**) $\langle w_{ij}\rangle - \langle \widehat{w}_{ij}\rangle$ versus time, where green, pink, purple, and orange are $(i,j) = (1,1)$, $(i,j) = (1,2)$, $(i,j) = (2,1)$, $(i,j) = (2,2)$, respectively, blue triangles are starting times of analysis periods, and black circles are end times of analysis periods. **d**) $\langle E\rangle$ versus $\langle w_{ij}\rangle - \langle \widehat{w}_{ij}\rangle$ versus t.

The values for learning rate $\beta$ and equilibrium objective function $E_e$ were derived for the SNIC over a range of input neuron conditions and were compared to those of 13 separate Human trials, with different human participants. The values were derived according to the procedures described in Section 5.6. In Figure 42, values of $E_e$ are plotted versus $\log_2\left(\frac{k}{k_0}\right)$ and $\log_2\left(\frac{I_L}{I_{L0}}\right)$, where $k = a$, $k_0 = \min(a) = 5.5\ \mu s$, $I_L = c$, and $I_{LO} = \min(c) = 45\ kHz$. Values from trials with 13 different human participants are plotted in the bottom right corner. The blue circles demarcate values from different learning periods within different trials of each condition. The mean is demarcated with a bar with red top surface at the mean value of $E_e$ for a condition. The optimum condition was $\log_2\left(\frac{k}{k_0}\right) = 2$, $\log_2\left(\frac{I_L}{I_{L0}}\right) = 1$, which had the minimum mean $E_e$. A trial with these input neuron settings is shown in Figure 40. The results  In Figure 43, values of learning rate $\beta$ are plotted versus $\log_2\left(\frac{k}{k_0}\right)$ and $\log_2\left(\frac{I_L}{I_{L0}}\right)$, along with values from human participants. The optimum condition was $\log_2\left(\frac{k}{k_0}\right) = 2$, $\log_2\left(\frac{I_L}{I_{L0}}\right) = 2$, which had the maximum $\beta$. The second highest $\beta$ value was with the condition $\log_2\left(\frac{k}{k_0}\right) = 2$, $\log_2\left(\frac{I_L}{I_{L0}}\right) = 1$, with a value close to that of the optimal condition. By weighting both the equilibrium value of objective function $E_e$ and the learning rate $\beta$, the condition with $\log_2\left(\frac{k}{k_0}\right) = 2$, $\log_2\left(\frac{I_L}{I_{L0}}\right) = 1$ was the optimal condition.

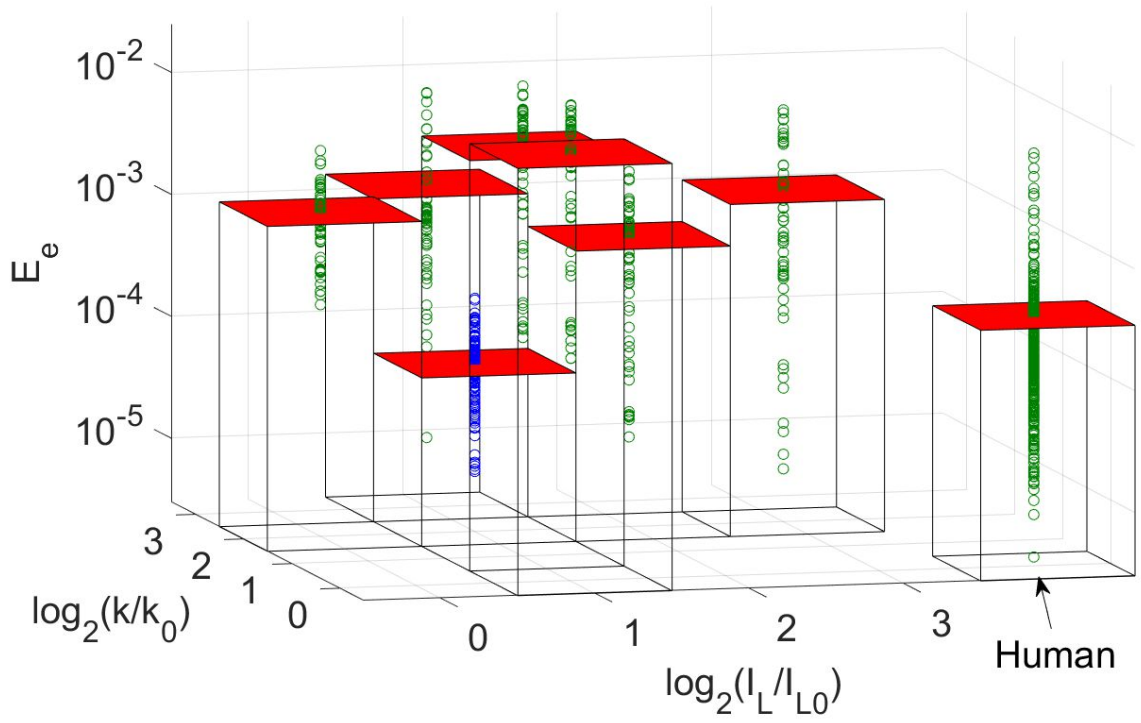


Figure 42. Equilibrium values for objective function $E_e$ with different SNIC conditions and human participants.

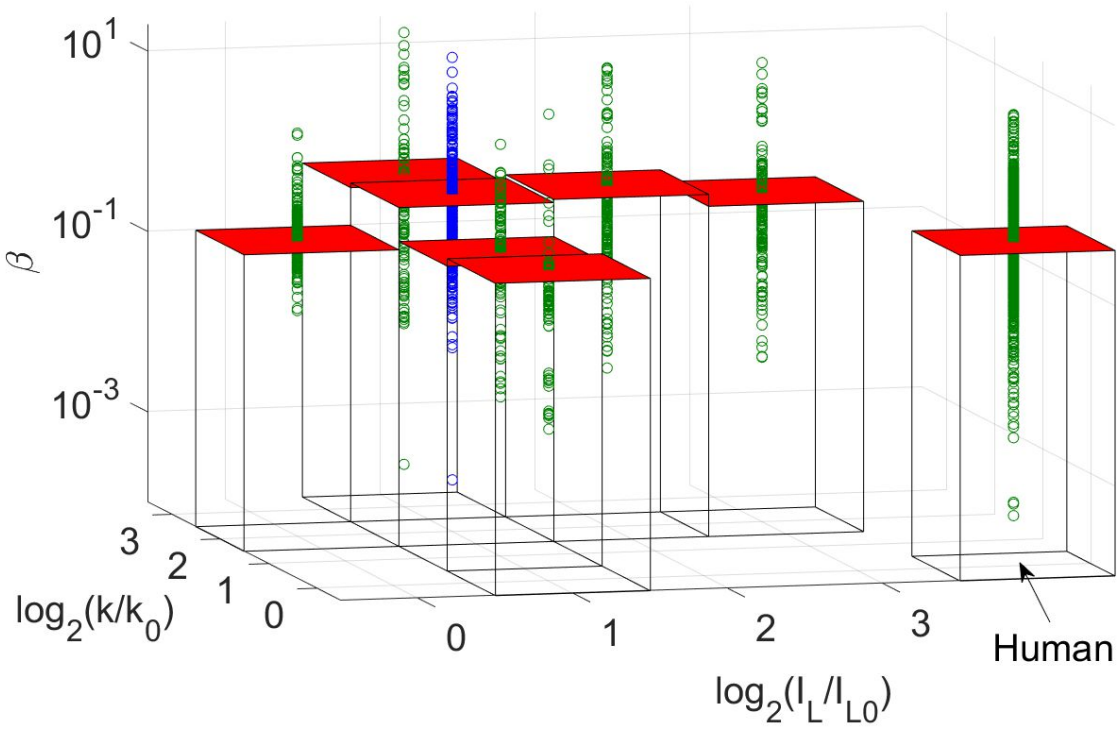


Figure 43. Learning rates $\beta$ with different SNIC conditions and human participants.

## 5.9. Performance Comparison

The performance of the SNIC at controlling the actuation to minimize the average objective function $\bar{E}$ over the duration of 5 minutes was compared with human and computer-based controllers. This value weights both the equilibrium value of objective function $E_e$ and the learning rate $\beta$. Figure 44 shows the results of the SNIC was tested over a range of $a$ and $c$ values (Figure 37), which controlled the sigmoidal shape of the input pulse rates. The optimal condition, $a = 11\ \mu s$ and $c = 182\ kHz$ (Figure 37b) was tested with a sample size of $N_s = 6$, of the optimized computer-controller was also $N_c = 6$, and of the humans was $N_h = 13$ unique participants. The sample mean $\mu_{\bar{E}} = \frac{\sum_{i=1}^{N} \bar{E}}{N}$ was calculated over $N$ experiments, with $\mu_{\bar{E}} =$

$6.11 \times 10^{-4}\ N^2, 1.76 \times 10^{-3}\ N^2$, and $3.07 \times 10^{-3}\ N^2$ for the SNIC, humans, and computer controller, respectively (Fig. 4).

| | $\mu_{\bar{E}}$ ($N^2$) | $\sigma_{\bar{E}}$ ($N^2$) | N |
|---|---|---|---|
| SNIC | $6.11 \times 10^{-4}$ | $7.52 \times 10^{-5}$ | 6 |
| PID | $3.07 \times 10^{-3}$ | $8.92 \times 10^{-5}$ | 6 |
| Human | $1.76 \times 10^{-3}$ | $1.14 \times 10^{-3}$ | 13 |

**Table 1.** Comparison of SNIC, PID, and Human performance for optimizing morphing wing in dynamic wind speed conditions. $\mu_{\bar{E}}$ and $\sigma_{\bar{E}}$ are the sample means and sample deviations of the average objective functions $\bar{E}$ over $N$ trials.

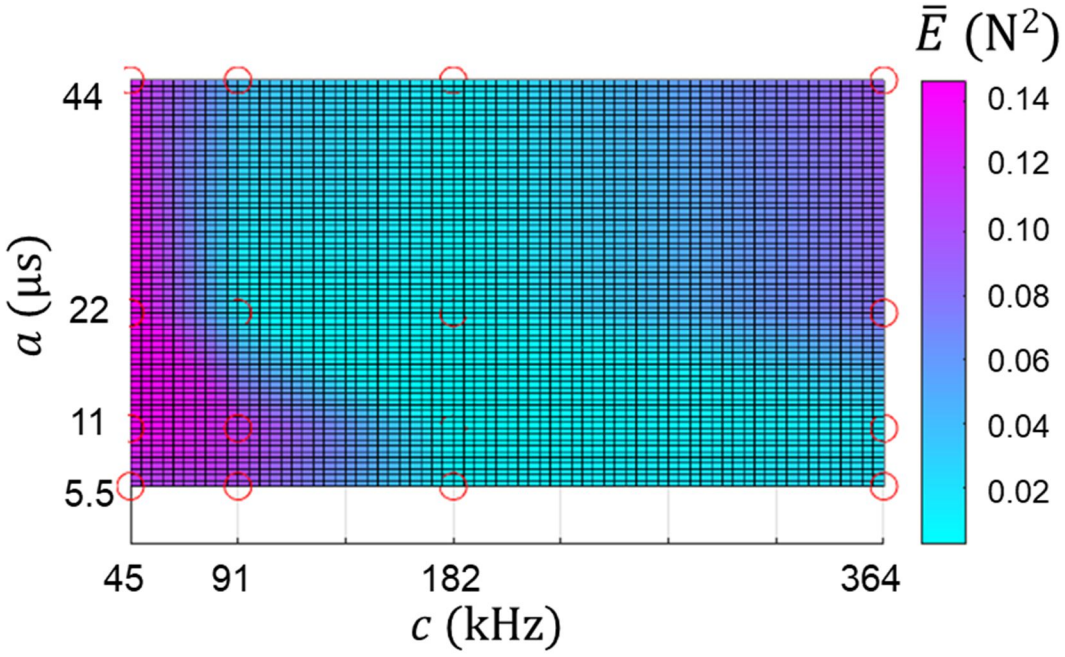


**Figure 44.** Average objective function $\bar{E}$ under various $a$ and $c$ values during 5-minute long morphing wing experiments with SNIC. Red circles indicate test points. The surface is a fitting with linear interpolation between the test points. The optimum condition is $a = 11\ \mu s, c = 182$ kHz.

# 6. Conclusion

The structure and properties of artificial synapse device based on Al-CNT Schottky interfaces and an oxide charge trap memory stack were developed to enable concurrent spatiotemporal inference (Equation 1) and correlative learning (Equation 2) algorithms. A crossbar of synstors was fabricated and connected to artificial neurons to form an artificial neural network. The synstor circuit (SNIC) demonstrated concurrent inference and learning with high energy efficiency by circumventing the fundamental computing limitations in existing electronic circuits such as physically separated logic and memory units, data transmission between memory and logic, the execution of the inference and learning algorithms in serial mode in different circuits, and the signal transmissions between the circuits. The equivalent computing energy efficiency of the $4 \times 2$ synstor circuit is $1.6 \times 10^{17}\ FLOPS/W$ (**Figure 1**), which exceeds the energy efficiencies of digital transistor circuits ($\sim 10^{9} - 10^{11}\ FLOPS/W$),[10-12,18] and of the analog neuromorphic circuits of memristors and PCM ($\sim 10^{5} - 10^{14}\ FLOPS/W$, excluding learning algorithm computations).[29,30,36]

In digital serial mode, transistors operate at high conductance ($\sim 10^{5}\ nS$) in order to enhance computing speed ($\sim 10^{9}\ Hz$); in analog parallel mode, synstors (synapses) operate at low conductance ($\lesssim 2\ nS$), and the computing speed and of an $M \times N$ synstor (synapse) circuit increase with increasing $M$ and $N$, the numbers of parallel input/output electrodes (Equations 13, 14, 15, Figure 27 and Figure 29). With the energy efficiency of the $4 \times 2$ circuit reported in this work, a $2\ k \times 2\ k$ circuit with a power consumption of $\sim 10\ mW$ could have a speed of $\sim 10^{15}\ FLOPS$ (Figure 28), exceeding the speeds of digital transistor circuits such as TPU, GPU, and FPGA ($\sim 10^{13}\ FLOPS$),[10-12] and the analog memory devices ($\sim 10^{12} - 10^{14}\ FLOPS$).[29,30,36]

Based on simulation of nanoscale devices (Section 2.7), synstors could potentially be miniaturized to nanoscale ($\sim 40\ nm$) with a performance density of $\sim 10^{17}\ \ FLOPS/mm^2$.

The self-programming synstor circuit also performed inference and Hebbian learning in situ to optimize a nonlinear morphing wing system in a dynamic environment in real time. The SNIC was integrated into a feedback control system to process a sensor signal as inputs and generate output actuation signals to optimize the system. The synstor circuit minimized the objective function of lift force error on the morphing wing under dynamic wind speed conditions by spontaneously estimating the gradient between error and actuation. The performance to minimize error by the SNIC was recorded to superior to computer-based and human controllers over multiple trials. The synstor crossbar and self-programming ability enable a scalable architecture to process multiple sensor signals and generate multiple actuation signals in real-time multi-dimensional system. The synstor circuit can potentially bypass the "curse of dimensionality" and "von Neumann bottleneck" of transistor computing circuits, leading to a new computing platform with real-time self-programming functionality and general intelligence in complex and erratic environments. There is "plenty of room at the bottom" to miniaturize the synstor size, scale up synstor circuits, optimize their materials and fabrication processes, and improve their energy efficiency, speed, power consumption, and uniformity for concurrent inference and learning from big data in intelligent systems.

# References


1 Turing, A. M. Computing machinery and intelligence. *Mind* **49**, 28 (1950).
2 Waldrop, M. M. More than Moore. *Nature* **530**, 144-147 (2016).
3 Koomey, J., Berard, S., Sanchez, M. & Wong, H. Implications of historical trends in the electrical efficiency of computing. *IEEE Annals of the History of Computing* **33**, 46-54 (2011).
4 Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484-489 (2016).
5 Williams, R. S. What's Next? *Comput Sci Eng* **19**, 7-13 (2017).
6 Zhirnov, V., Cavin, R. & Gammaitoni, L. in *Minimum energy of computing, fundamental considerations, ICT-Energy-Concepts Towards Zero-Power Information and Communication Technology* (InTech, 2014).
7 Semiconductor Industry Association https://www.semiconductors.org/clientuploads/Resources/RITR%20WEB%20version%20FINAL.pdf. 2015).
8 Hey, A. J. G. & Feynman, R. P. *Feynman and computation : exploring the limits of computers*. (Westview Press/Perseus Books, 2002).
9 Backus, J. Can Programming Be Liberated from Von Neumann Style - Functional Style and Its Algebra of Programs. *Commun Acm* **21**, 613-641, doi:Doi 10.1145/359576.359579 (1978).
10 Nvida. http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pd, 2017).
11 Jouppi, N. P. e. a. In-Datacenter Performance Analysis of a Tensor Processing Unit. *44th Annual International Symposium on Computer Architecture (Isca 2017)*, 1-12, doi:10.1145/3079856.3080246 (2017).
12 Nurvitadhi, E. *et al.* in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* 5-14 (ACM, Monterey, California, USA, 2017).
13 Merolla, P. A. *et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668-673, doi:10.1126/science.1254642 (2014).
14 Hasler, J. & Marr, H. B. Finding a roadmap to achieve large neuromorphic hardware systems. *Front Neurosci-Switz* **7**, 118 (2013).
15 Xu, X. *et al.* Scaling for edge inference of deep neural networks. *Nature Electronics* **1**, 216 (2018).
16 Conti, J. *et al.* International energy outlook 2016 with projections to 2040. (USDOE Energy Information Administration (EIA), Washington, DC (United States). Office of Energy Analysis, 2016).
17 Kurzweil, R. *The singularity is near : when humans transcend biology*. (Viking, 2005).
18 Oak Ridge National Laboratory, America's newest and smartest supercomputer, https://www.olcf.ornl.gov/summit/, 2018).
19 Zeki, S. A massively asynchronous, parallel brain. *Philos T R Soc B* **370**, 103-116 (2015).
20 Chen, Z., Haykin, S., Eggermont, J. J. & Becker, S. Correlative Learning: A Basis for Brain and Adaptive Systems. *Correlative Learning: A Basis for Brain and Adaptive Systems*, 1-448, doi:10.1002/9780470171455 (2007).

21 Friston, K. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* **11**, 127 (2010).
22 Gerstner, W. & Kistler, W. M. *Spiking neuron models: Single neurons, populations, plasticity*. (Cambridge university press, 2002).
23 Hebb, D. O. *The organization of behavior; a neuropsychological theory*. (Wiley, 1949).
24 Diorio, C., Hasler, P., Minch, A. & Mead, C. A. A single-transistor silicon synapse. *Ieee T Electron Dev* **43**, 1972-1980, doi:Doi 10.1109/16.543035 (1996).
25 Lai, Q. X. *et al.* Ionic/Electronic Hybrid Materials Integrated in a Synaptic Transistor with Signal Processing and Learning Functions. *Adv Mater* **22**, 2448-2453, doi:10.1002/adma.201000282 (2010).
26 Kim, K., Chen, C. L., Truong, Q., Shen, A. M. & Chen, Y. A Carbon Nanotube Synapse with Dynamic Logic and Learning. *Adv Mater* **25**, 1693-1698 (2013).
27 Gu, X. & Iyer, S. S. Unsupervised Learning Using Charge-Trap Transistors. *Ieee Electr Device L* **38**, 1204-1207 (2017).
28 Prezioso, M. *et al.* Training andoperation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61-64, doi:10.1038/nature14441 (2015).
29 Hu, M. *et al.* Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine. *Adv Mater* **30** (2018).
30 Li, C. *et al.* Analogue signal and image processing with large memristor crossbars. *Nature Electronics* **1**, 52 (2018).
31 Shafiee, A. *et al.* ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. *ACM SIGARCH Computer Architecture News* **44**, 14-26 (2016).
32 Chi, P. *et al.* in *ACM SIGARCH Computer Architecture News*. 27-39 (IEEE Press).
33 Yao, P. *et al.* Face classification using electronic synapses. *Nat Commun* **8** (2017).
34 Wang, Z. *et al.* Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nat Mater* **16**, 101 (2017).
35 Eryilmaz, S. B. *et al.* Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array. *Front Neurosci-Switz* **8** (2014).
36 Ambrogio, S. *et al.* Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60 (2018).
37 Bellman, R. & Rand Corporation. *Dynamic programming*. (Princeton University Press, 1957).
38 Chen, Z. H., Appenzeller, J., Knoch, J., Lin, Y. M. & Avouris, P. The role of metal-nanotube contact in the performance of carbon nanotube field-effect transistors. *Nano Lett* **5**, 1497-1502 (2005).
39 Rinkio, M. *et al.* High-yield of memory elements from carbon nanotube field-effect transistors with atomic layer deposited gate dielectric. *New J Phys* **10** (2008).
40 Sze, S. M. & Ng, K. K. *Physics of semiconductor devices*. (John wiley & sons, 2006).
41 Tans, S. J., Verschueren, A. R. & Dekker, C. Room-temperature transistor based on a single carbon nanotube. *Nature* **393**, 49 (1998).
42 Xu, Z. *et al.* Tuning the Fermi Level of TiO2 Electron Transport Layer through Europium Doping for Highly Efficient Perovskite Solar Cells. *Energy Technol-Ger* **5**, 1820-1826 (2017).
43 Hodgkin, A. L. & Huxley, A. F. A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve. *J Physiol-London* **117**, 500-544 (1952).

44 Sahidullah, M. & Saha, G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Commun* **54**, 543-565 (2012).
45 Maass, W. On the computational power of winner-take-all. *Neural Comput* **12**, 2519-2535 (2000).
46 Dan, Y. & Poo, M.-M. Spike Timing-Dependent Plasticity of Neural Circuits. *Neuron* **44**, 23-30, doi:10.1016/j.neuron.2004.09.007 (2004).
47 Eryilmaz, S. B. *et al.* Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array. *Frontiers in neuroscience* **8**, 205 (2014).
48 Yao, P. *et al.* Face classification using electronic synapses. *Nature Communications* **8**, 15199, doi:10.1038/ncomms15199 (2017).
49 Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology* **117**, 500-544, doi:10.1113/jphysiol.1952.sp004764 (1952).
50 Indiveri, G. *et al.* Neuromorphic Silicon Neuron Circuits. *Frontiers in Neuroscience* **5**, doi:10.3389/fnins.2011.00073 (2011).
51 Danesh, C. D. *et al.* Synaptic Resistors for Concurrent Inference and Learning with High Energy Efficiency. *Advanced Materials* **31**, 1808032, doi:10.1002/adma.201808032 (2019).
52 Hebb, D. O. *The organization of behavior: a neuropsychological theory*. (J. Wiley; Chapman & Hall, 1949).
53 Fuller, E. J. *et al.* Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing. *Science* **364**, 570-574 (2019).
54 Ambrogio, S. *et al.* Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* **558**, 60-67, doi:10.1038/s41586-018-0180-5 (2018).
55 Gamble, L. L., Pankonien, A. M. & Inman, D. J. Stall Recovery of a Morphing Wing via Extended Nonlinear Lifting-Line Theory. *AIAA Journal*, 1-8, doi:10.2514/1.j055042 (2017).
56 Pankonien, A., Inman, Daniel. Experimental testing of spanwisemorphing trailing edge concept. *Proc. SPIE 8688, Active and Passive Smart Structures and Integrated Systems* **868815**, doi:10.1117/12.2009400.short (2013).
57 Åström, K. J. & Hägglund, T. *PID controllers: theory, design, and tuning*. Vol. 2 (Instrument society of America Research Triangle Park, NC, 1995).
58 Boyd, S., Boyd, S. P. & Vandenberghe, L. *Convex optimization*. (Cambridge university press, 2004).