APPLIED SCIENCES AND ENGINEERING

Analog memristive synapse based on topotactic phase transition for high-performance neuromorphic computing and neural network pruning

Xing Mou[1], Jianshi Tang[1,2]*, Yingjie Lyu[3], Qingtian Zhang[2], Siyao Yang[1], Feng Xu[1], Wei Liu[1], Minghong Xu[1], Yu Zhou[3], Wen Sun[1], Yanan Zhong[1], Bin Gao[1,2], Pu Yu[2,3]*, He Qian[1,2], Huaqiang Wu[1,2]

Inspired by the human brain, nonvolatile memories (NVMs)–based neuromorphic computing emerges as a promising paradigm to build power-efficient computing hardware for artificial intelligence. However, existing NVMs still suffer from physically imperfect device characteristics. In this work, a topotactic phase transition random-access memory (TPT-RAM) with a unique diffusive nonvolatile dual mode based on $SrCoO_x$ is demonstrated. The reversible phase transition of $SrCoO_x$ is well controlled by oxygen ion migrations along the highly ordered oxygen vacancy channels, enabling reproducible analog switching characteristics with reduced variability. Combining density functional theory and kinetic Monte Carlo simulations, the orientation-dependent switching mechanism of TPT-RAM is investigated synergistically. Furthermore, the dual-mode TPT-RAM is used to mimic the selective stabilization of developing synapses and implement neural network pruning, reducing ~84.2% of redundant synapses while improving the image classification accuracy to 99%. Our work points out a new direction to design bioplausible memristive synapses for neuromorphic computing.

INTRODUCTION

The growth of computing power in digital hardware, including central processing unit and graphics processing unit, has driven the rapid development of artificial intelligence. This, in turn, raises higher and higher demand on the hardware performance, even exceeding the pace of Moore's law. One of the key bottlenecks arises from the physical separation of memory and computing units in the widely adopted von Neumann architecture, which leads to a grand challenge of memory wall problem. Inspired by neurobiological systems, neuromorphic computing has emerged as a promising computing paradigm with the feature of massively parallel computation in memory to break the so-called von Neumann bottleneck (*1*, *2*). Various nonvolatile memories (NVMs), such as resistive random-access memory (RRAM) (*3*, *4*) and phase-change memory (PCM) (*5*), have been extensively studied as artificial synapses and neurons to build prototype artificial intelligence chips (*6*–*8*). Different from digital memory applications, here, reproducible analog switching characteristics (e.g., multilevel conductance states, weight update linearity and symmetry, and low variability) are desired to meet the requirement of high computing accuracy and energy efficiency (*9*, *10*).

Unfortunately, those existing emerging NVMs still suffer from nonideal device characteristics (fig. S1), which are one of the main challenges for the hardware implementation of large-scale neuromorphic computing systems. For example, conventional filament-type RRAM relies on the random oxygen vacancy (V_o) migration in the amorphous switching oxides, leading to intrinsically large device variations, while the absence of local Joule heating effect in interface-type RRAM usually results in poor retention and low speed (*11*). For PCM, it typically shows asymmetric switching due to the abrupt quench process in the crystalline-to-amorphous phase transition and also suffers from the conductance drift issue (*12*). It is noted that those imperfect device characteristics originate from their intrinsic working mechanisms and hence are difficult to be eliminated by simply optimizing their device structures (*13*, *14*). In addition, so far, these devices are mainly limited to mimic the functionalities of an individual neuron or synapse (synaptic plasticity, neuronal firing, etc.), while the biomimicry of many important network-level properties, such as neural network pruning that is critical for cognitive learning in biology, has not been explored yet. Therefore, for future high-performance neuromorphic computing, innovations in materials and devices with new working mechanisms are highly desired to yield more controllable analog switching characteristics and further construct more bioplausible neural networks (*15*, *16*).

In search of new materials and structures for low-variability analog switching memristors, here, we propose a novel synapse, namely, topotactic phase transition RAM (TPT-RAM), using brownmillerite (BM) oxides [such as $SrCoO_{2.5}$ (SCO) (*17*, *18*) and $SrFeO_{2.5}$ (*19*–*21*)] as the resistive switching oxide. We chose SCO as an exemplary material whose unique crystal structure formed by alternating stacks of oxygen octahedra and oxygen tetrahedra provides the favorable conditions to achieve uniform analog switching: (i) The highly ordered one-dimensional oxygen vacancy channels (OVCs) provide predefined freeway for the migration of oxygen ions to induce phase transition and resistive switching (*22*). Compared with other methods intended to confine the ion migration, such as metal doping (*23*) and dislocation engineering (*15*), the highly ordered and atomically precise OVCs in BM oxides are more uniform and easier to manipulate without additional ex situ processes (*24*). (ii) The multivalent cobalt ions change reversibly between BM and perovskite (PV) structures on the basis of the adjustment of oxygen stoichiometry without losing the intrinsic lattice architecture (*25*–*27*), which

[1]School of Integrated Circuits, Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China. [2]Beijing Innovation Center for Future Chips (ICFC), Tsinghua University, Beijing, China. [3]State Key Laboratory of Low-Dimensional Quantum Physics and Department of Physics, Tsinghua University, Beijing 100084, China.
*Corresponding author. Email: jtang@tsinghua.edu.cn (J.T.); yupu@tsinghua.edu.cn (P.Y.)

can yield gradual switching. (iii) High-quality, stable BM oxide as the resistive switching layer ensures excellent retention at multilevel conductance states and also enhanced endurance.

In this work, to implement low-power neuromorphic computing, we designed and fabricated SCO-based TPT-RAM with tunable OVCs as memristive synapses. Such TPT-RAM provided an excellent platform to thoroughly study the topotactic phase transition–associated switching mechanism by correlating electrical and structural characterizations with comprehensive atomic-device modeling and simulations (*28*–*30*), which, however, is difficult to perform for conventional RRAM with amorphous oxide like HfO_2 (*31*). Experimentally, we demonstrated that the high-speed and uniform analog TPT-RAM can be achieved by manipulating OVCs through the top and bottom electrodes (BEs). Furthermore, inspired by the selective stabilization of developing synapses in biological neural networks, we implemented the online training of a sparse neural network through automatic pruning, realizing a substantial reduction of both network size and power consumption.

RESULTS

Orientation-dependent switching characteristics of TPT-RAM

Figure 1 (A and B) illustrates the schematic of TPT-RAM synapses with tunable OVCs and their phase transition mechanism. To start, 35-/17-nm-thick single-crystalline BM-SCO/$SrRuO_3$ (SRO) thin films were epitaxially grown on both (001)- and (110)-oriented $SrTiO_3$ (STO) substrates by pulsed laser deposition (PLD) method. The alternating stacking of oxygen octahedra and tetrahedra in the BM-SCO results in highly ordered OVCs. Because of the epitaxial strain and crystalline symmetry, the OVCs orients within the plane for films grown on SCO $(001)_{pc}$ and have a large out-of-plane component for the films grown on SCO $(110)_{pc}$ (*32*–*34*). The metallic SRO was used as the BE, enabling an atomically clean epitaxial interface without misfit dislocations (fig. S2) (*35*). After the film growth, a 20-nm-thick Al_2O_3 capping layer was deposited on top to protect the SCO layer. Then, the TPT-RAM device area was defined by etching contact vias in Al_2O_3 with sizes ranging from $(4\ \mu m)^2$ to $(100\ \mu m)^2$. Last, the top electrodes (TEs) were made by sputtering of 100-nm-thick Au. Both Au and SRO could form an ohmic contact with SCO (*36*), which is crucial to eliminate the effect of interfacial barriers and obtain symmetric *I*-*V* characteristics. The different crystal structures of these two SCO thin films were further verified by both x-ray diffraction (XRD) (Fig. 1C) and aberration-corrected scanning transmission electron microscopy (Fig. 1D), where the oxygen tetrahedral layers (ordered OVCs) are labeled with pink arrows.

To investigate the device switching mechanism, we compared the switching characteristics of SCO $(001)_{pc}$ and SCO $(110)_{pc}$ TPT-RAM. Figure 2A presents the typical forming processes and subsequent consecutive *I*-*V* sweeps for the two SCO devices. Figure 2B plots the statistical distributions of the forming voltage measured from 30 devices to affiliate the oxygen ion migration. The corresponding values in SCO $(001)_{pc}$ devices (5.04 ± 0.07 V) are notably higher than that in SCO $(110)_{pc}$ devices (3.98 ± 0.05 V). It also shows markedly enhanced uniformity in SCO $(110)_{pc}$ devices. Fig. 2C displays the uniformity of TPT-RAM device conductance (measured at a read voltage of 0.2 V) over 500 switching cycles. The results in fig. S3 (A and B) show that the cycle-to-cycle variations (σ/μ) are very low in SCO $(110)_{pc}$ TPT-RAM: only 1.8% for set voltage and 0.9% for reset voltage, while 2.25 and 13.74% variations for the high-resistance state (HRS) and the low-resistance state (LRS), respectively. The cycle-to-cycle uniformity of TPT-RAM can be attributed to the

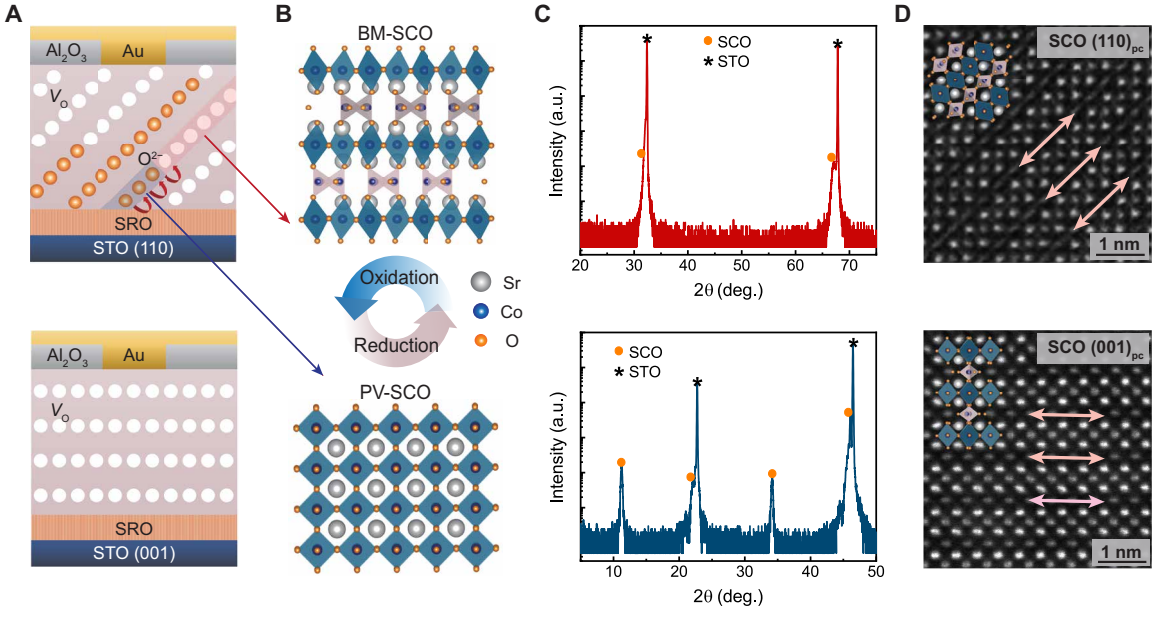

Fig. 1. Design of the SCO-based TPT-RAM synaptic devices with tunable orientations of OVCs. (**A**) Schematics of TPT-RAM based on different crystal orientations of SCO. (**B**) Reversible field–driven topotactic phase transition between BM-SCO and PV-SCO depends on the oxygen stoichiometry. (**C**) XRD θ-2θ patterns of the two epitaxial SCO thin films grown on the SRO-buffered STO (110) substrate (top) and (001) substrate (bottom). (**D**) Atomic-resolution scanning transmission electron microscopy images of SCO $(110)_{pc}$ (top) and SCO $(001)_{pc}$ (bottom) exhibit two different orientations of OVCs, where the oxygen tetrahedral layers are labeled with pink arrows. The insets show the corresponding structure models. a.u., arbitrary units.

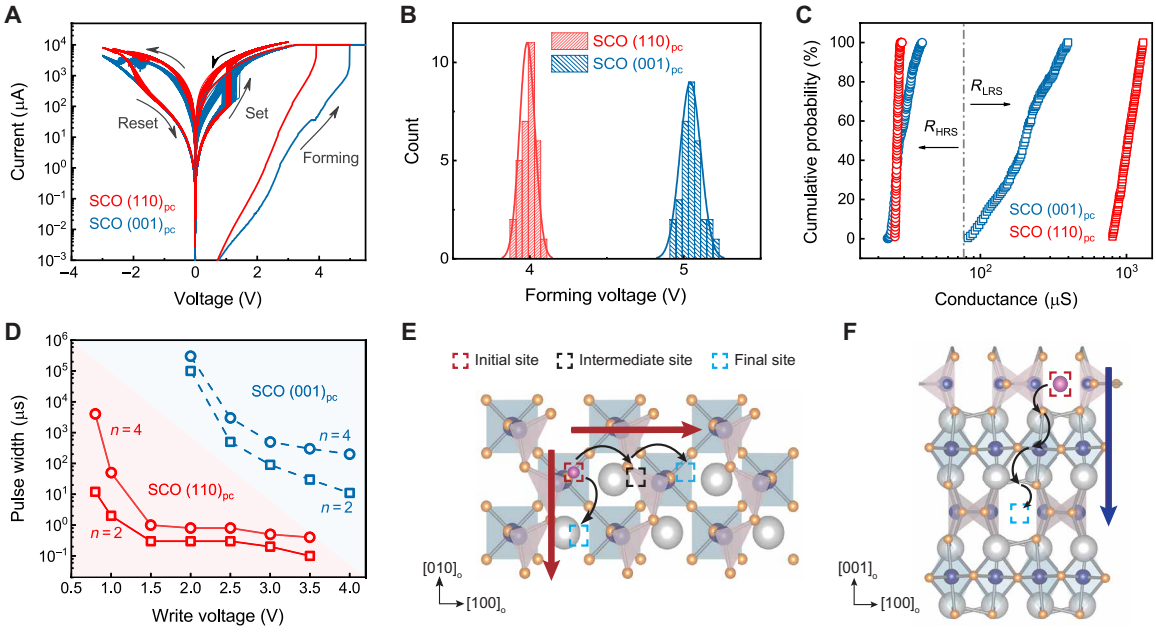

Fig. 2. Effect of the tunable OVCs on the switching characteristics of TPT-RAM. (**A**) Typical DC forming and *I-V* curves for 100 consecutive sweeps. (**B**) Statistical analysis of the forming voltages extracted from 32 different devices of each SCO orientation. (**C**) Cumulative probability plots of the device conductance over 500 *I-V* sweeps with 0.2-V read voltage. (**D**) The required pulse operation condition to program the device from the same initial conductance G_{inital} to the target value G_{target}. The on/off ratio is defined as $n = G_{target}/G_{inital}$. (**E**) The oxygen migration pathways within the oxygen tetrahedral layers and (**F**) across the oxygen tetrahedral layers in SCO.

highly anisotropic ion migration pathways in the anisotropic SCO crystal structure. In addition, fig. S3 (C and D) also shows excellent reproducibility with low device-to-device variation (down to 4.9%) and good batch-to-batch uniformity, which is mainly due to the high-quality epitaxial SCO thin film.

Furthermore, we developed a pulse test scheme to evaluate the operation speed of these two different SCO devices (fig. S4). Each device was started from a similar low-conductance state ($G_{initial}$ = 40 μS), and then a series of pulses with fixed amplitude (0.8 to 4 V) but different widths (100 ns to 10 ms) were continuously applied. The device conductance value was read out after each operation pulse, until it reached the target conductance G_{target} (80 and 160 μS, corresponding to a conductance on/off ratio of $n = G_{target}/G_{initial}$ = 2 and 4, respectively). After that, the device was reset to the initial low-conductance state, and then the pulse amplitude was changed to program the device again. In this way, by adding up the pulse widths applied to program the device from its initial state to the target state, we can estimate the pulse operation conditions needed to reach the target state. This test method can avoid the effect of different initial resistance states and test as many pulse conditions as possible. The results in Fig. 2D show that the conductance change of SCO $(110)_{pc}$ was much easier than SCO $(001)_{pc}$, where lower operation voltage (down to 0.8 V) and faster speed (up to 100 ns, which was limited by our measurement equipment) were achieved in the SCO $(110)_{pc}$ TPT-RAM device as compared to the values (~2.0 V and ~10 μs, respectively) for SCO $(001)_{pc}$.

To better understand the electrical characteristics of TPT-RAM devices, we established a synergistic modeling approach to study the underlying switching mechanism (Fig. 2, E and F, and figs. S5 to S8). First, density functional theory (DFT) calculations were performed to evaluate the migration barriers along different directions within SCO at the atomic scale (fig. S5) (*22*). Figure 2E presents two different migration pathways for oxygen ions within the oxygen tetrahedral layers. Take two adjacent positions as the initial and final states, respectively, and the migration barrier along the OVCs is estimated to be 0.56 eV. In comparison, the lowest migration barrier perpendicular to the OVCs is 0.97 eV, where the oxygen at site *X* moves to site *Y*, while another oxygen atom at *Y* hops to site *Z*. Besides, Fig. 2F presents that the migration barrier from one tetrahedral layer to another through the octahedral layer is 1.84 eV, where there are three oxygen ions involved. The calculated values summarized in fig. S6 suggest that the oxygen ions favor the migration along the orientation of OVCs.

Besides, we performed kinetic Monte Carlo (KMC) simulations to capture the complete physical processes of the resistive switching (fig. S7) (*37*–*39*). In the SCO $(110)_{pc}$ device, the OVCs aligned with the TE-BE electric field direction provide preferable migration paths for oxygen ions so that the BM-to-PV phase transition occurred easily and multiple filaments were formed. In comparison, oxygen ions in the SCO $(001)_{pc}$ device were driven by the electric field perpendicular to the OVCs, so there was a lower probability for ions moving along the electric field direction, leading to detrimental random phase transitions (fig. S8 and movie S1). Therefore, the SCO $(110)_{pc}$ device was easier to form conducting filaments and hence could operate at a higher speed and lower voltage as shown in Fig. 2D. It is also consistent with the observed smaller variation (σ/μ = 0.9%) in the write voltage for SCO $(110)_{pc}$ devices (fig. S3). These results suggest that the formation of conducting filaments depends on the oriented oxygen ion migrations along the OVCs in the SCO $(110)_{pc}$. Moreover, the atomic-scale spatial uniformity shows a significant improvement compared to the conventional amorphous oxide–based RRAM devices, in which the random V_o formations and

movements would induce a larger switching variation (*38*). In addition, the weak dependence of LRS on the device area in SCO $(110)_{pc}$ in fig. S9A suggests that the switching mechanism is filamentary type, which is consistent with the KMC simulation results. The local electric field was enhanced at the tips of filaments after forming, which resulted in a localized region of filament formation within the device area. On the contrary, there is a clear area dependence of the device conductance for SCO $(001)_{pc}$, which suggests that the phase transition is not localized with a small region as in the case of SCO $(110)_{pc}$ (fig. S9B) (*18*).

Analog switching characteristics of TPT-RAM

On the basis of the above results, we used SCO $(110)_{pc}$–based TPT-RAM to further investigate its high-performance analog switching characteristics for neuromorphic computing. The long-term potentiation (LTP) and long-term depression (LTD) characteristics were tested with identical set and reset pulses (Fig. 3A). Each cycle consists of 50 set and 50 reset pulses with the pulse width of 1 μs followed by the read pulses of 0.2 V and 1 μs. The result confirms that the analog switching of TPT-RAM can be realized with identical low-voltage pulses (1 V), and it also exhibited nearly ideal linear switching with nonlinearity factors of ν = 0.20 and 1.29 for LTP and LTD, respectively (fig. S10A). Both values are much lower than those reported for the typical filament-type RRAM (*9*). Besides, the cycle-to-cycle pulse programming tests in Fig. 3B also show reproducible analog switching characteristics with both identical pulses and amplitude-increasing pulses. It is shown that the latter programming scheme could achieve better linearity (fig. S10) and larger on/off ratio (*40*).

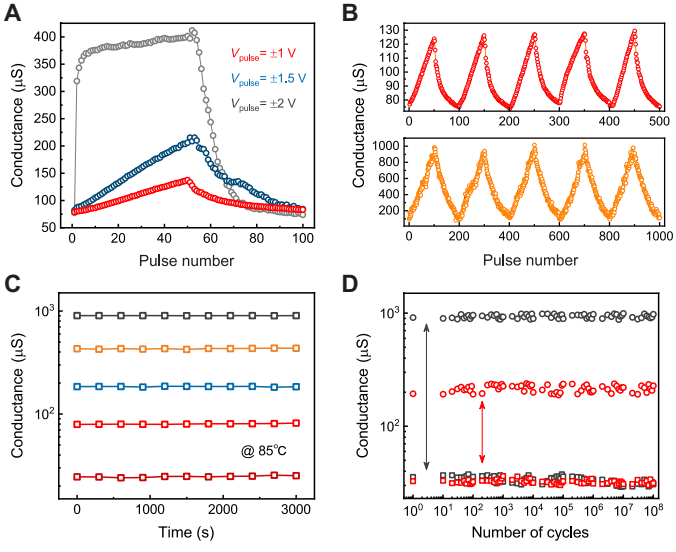

Fig. 3. Analog switching characteristics of SCO $(110)_{pc}$–based TPT-RAM for neuromorphic computing. (**A**) LTP and LTD with identical set and reset pulses. The write pulse train consists of 50 set pulses (amplitudes of 1, 1.5, and 2 V, the width of 1 μs) followed by 50 reset pulses (amplitudes of −1, −1.5, and −2 V, the width of 1 μs). The device conductance was measured by a read pulse (amplitude of 0.2 V, the width of 1 μs) after each write pulse. (**B**) Reproducible and uniform analog switching behavior over multiple cycles: under identical pulses with an amplitude of 1 V, a width of 1 μs (top) and increasing pulse amplitudes with 1 to 2.89 V, width of 100 ns (bottom). (**C**) Retention test over a long period of 3000 s at 85°C for five conductance levels. (**D**) Device endurance measurement with different resistive switching windows (fixed HRS and varied LRS). For a given switching window, here, we use pulses with amplitudes of 1.8 to 3.2 V and a width of 1 μs to program the device to the desired HRS and LRS in each cycle, which were recorded at a read voltage of 0.2 V. The device remains stable after 10^8 cycles in both cases, demonstrating excellent reliability.

The reversible phase transition without losing parent crystal structure enables excellent retention and endurance of TPT-RAM. Figure 3C shows that the device has long retention of more than 3000 s at 85°C. Furthermore, using the extracted activation energy from temperature-dependent measurements, the device retention at room temperature can be predicted to be more than 10 years (fig. S11) (*41*). The presence of oriented OVCs in the SCO $(110)_{pc}$–based TPT-RAM enables the device to be programmed readily with a lower migration barrier and also have excellent retention with a higher diffusion barrier (fig. S12). Figure 3D shows reliable multilevel switching with different resistive switching windows over 10^8 pulses. All these results demonstrate that developed TPT-RAM based on SCO $(110)_{pc}$ thin films can serve as a high-performance synaptic device (for comparison with other devices, see fig. S1 and table S1).

The nonvolatile analog switching characteristics presented above were measured after the forming process, before which the device behaved as a diffusive memristor instead (*42*). Figure 4A schematically illustrates the working mechanism under different pulse conditions, where the device initially exhibited a high resistance (>1 megohm) due to the insulating BM phase. A weak excitation below its forming voltage, e.g., 2 V (Fig. 4C), would extract oxygen ions from the SRO layer to the SRO/SCO interface and lower the device resistance. However, it could not facilitate a fully stable phase transition of SCO that requires a sufficiently large amount of oxygen ions (*27*, *43*). Once removing the voltage bias, the chemical potential difference at the interface would induce the diffusion of oxygen ions back to the SRO layer, causing the device to gradually relax back to the initial HRS. A read voltage pulse of 0.2 V was applied before and after the write pulse to record the device conductance state. The device was first programmed to a relatively low resistance state but then spontaneously relaxed back to the initial state within ~1 s (Fig. 4C). As a result, the TPT-RAM device exhibited diffusive memristor behavior under weak excitation. In addition, the endurance test in fig. S13 demonstrates a device endurance of more than 1500 cycles in the diffusive mode.

On the contrary, under strong excitation, e.g., DC forming at about 4 V (Fig. 3A), sufficient oxygen ions were driven into SCO, which triggered a stable phase transition. After that, a small positive voltage could attract more oxygen ions to migrate and induce a stable phase transition from BM-SCO to PV-SCO (Fig. 3A), increasing the device conductance (set process). The reverse phase transition would occur when a negative voltage is applied to drive oxygen ions back to SRO, reducing the device conductance (reset process). As a result, in this nonvolatile mode, relatively low resistance and excellent retention (>3000 s at 85°C) were observed. The electrical test results in the two different modes are shown in Fig. 4 (C and D, respectively). To further clarify the forming process, we tested the device using different numbers of identical pulses (V_{pulse} = 3 V, pulse width = 10 ms) as shown in Fig. 4B, and the transition from diffusive mode to nonvolatile mode can also be observed. Starting from the same initial conductance state, both 10 and 50 pulses could induce incremental conductance change, but it gradually decayed after the pulse operations, suggesting that the device operated in the diffusive mode. However, as the number of consecutive pulses increased

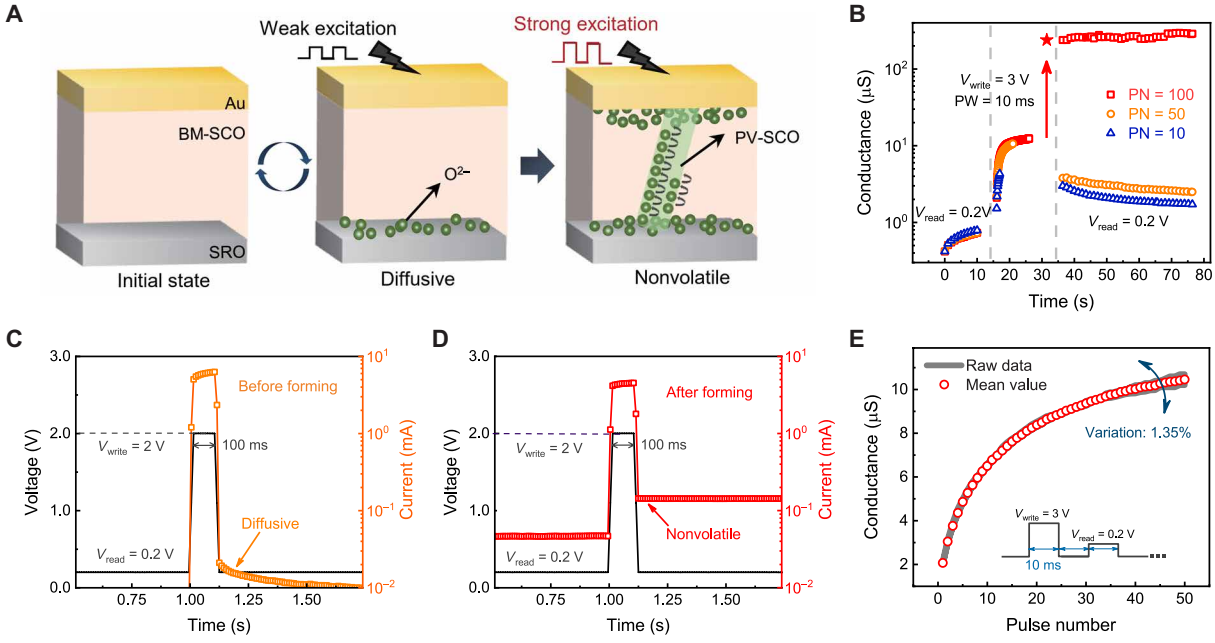

Fig. 4. The diffusive nonvolatile dual mode of the TPT-RAM synapse. (**A**) Schematic of the memristive switching mechanism. In the initial state, a weak excitation cannot induce a stable phase transition, and the device behaves as a diffusive memristor. After a strong excitation, sufficient amount of oxygen ions migrate across the SRO/SCO interface so that conductive filaments are formed, and the device exhibits nonvolatile characteristics. (**B**) The pulse forming process. A small number (≤50) of pulses could only induce diffusive switching behavior, but more pulses (≥100) could electroform the device and turn it into a nonvolatile memristor. (**C**) Pulse test before strong excitation (forming) shows diffusive characteristics under weak pulse. PN, pulse number; PW, pulse width. (**D**) Pulse test after strong excitation (forming) shows nonvolatile memristor characteristics, where the pulse condition is the same as in (C). (**E**)The multicycle pulse test in the diffusive mode also shows good cycle-to-cycle uniformity (low variation of σ/μ = 1.35%) and analog switching characteristics.

to 100, an abrupt conductance change was observed, turning the TPT-RAM device into the nonvolatile mode. These results indicate that the device can be formed by either a single large voltage or an adequate amount of repeated small voltage pulses. Besides, the cycling test with 50 set pulses of 3 V and 10 ms in Fig. 4E shows good uniformity (variation down to σ/μ = 1.35%) and analog switching characteristics (nonlinearity factor of TPT-RAM down to 1.8 as shown in fig. S14) in the diffusive mode as well.

Implementation of neural network pruning

Furthermore, we would like to highlight that the unique diffusive nonvolatile dual mode of TPT-RAM memristive synapse as revealed in Fig. 4 can be used to achieve better emulation of neurobiological functions beyond synaptic weight representation (*12*). In the past decade, there have been extensive studies on the emulation of synaptic behaviors in biology, such as synaptic plasticity, short-term memory, and long-term memory, using emerging memristive devices (*42*, *44*). However, so far, they are mainly limited to device-level biomimicry, while the emulation of many important network-level properties has not been explored yet (*2*), such as neural network pruning that is critical in human brain development (*45*). As a result, there is still a huge gap between biological and artificial neural networks, and more attention should be paid to systematically co-optimize the architectural designs and bioplausible hardware properties (*2*). In addition, the deployment of deep neural networks in resource-limited applications (e.g., portable electronics and Internet of Things) is largely limited by the high power consumption and lack of real-time processing capability. Neural network pruning is considered an effective pathway to reduce network complexity and avoid overfitting (*46*). However, this usually results in irregular network connections and may require extra efforts to represent sparse locations, thus incurring additional hardware overhead and computational cost (*47*). Therefore, neural network pruning has not been demonstrated using memristive synapses yet.

In this work, we used the developed TPT-RAM to implement the automatic neural network pruning in the network training with deep learning algorithm (*46*). Such automatic pruning process is inspired by the synapse-developing process in the human brain (*45*), as illustrated in Fig. 5A. In the human brain, the number of synaptic connections reaches its maximum in early childhood, and then active synapses are selectively stabilized, while redundant synapses that are rarely used are gradually eliminated. This natural synaptic pruning process is essential for refining the neural network and increasing network efficiency. Here, a differential pair of two TPT-RAM devices was used to represent one synaptic weight. Initially, all the synaptic devices worked in the diffusive mode. During training, the devices that were frequently updated would be electroformed and turned into the nonvolatile state eventually (representing stabilized synaptic connections), while the others would naturally decay to off state. In this way, the neural network could enhance those important synaptic weight connections and trim down other irrelevant synaptic connections and eventually became a sparse network after the training. The simulation results in Fig. 5 (B and C) are compared to the baselines trained with conventional nonvolatile synapses in both multilayer perceptron (MLP) and convolutional neural network (CNN). Here, the device cycle-to-cycle variations

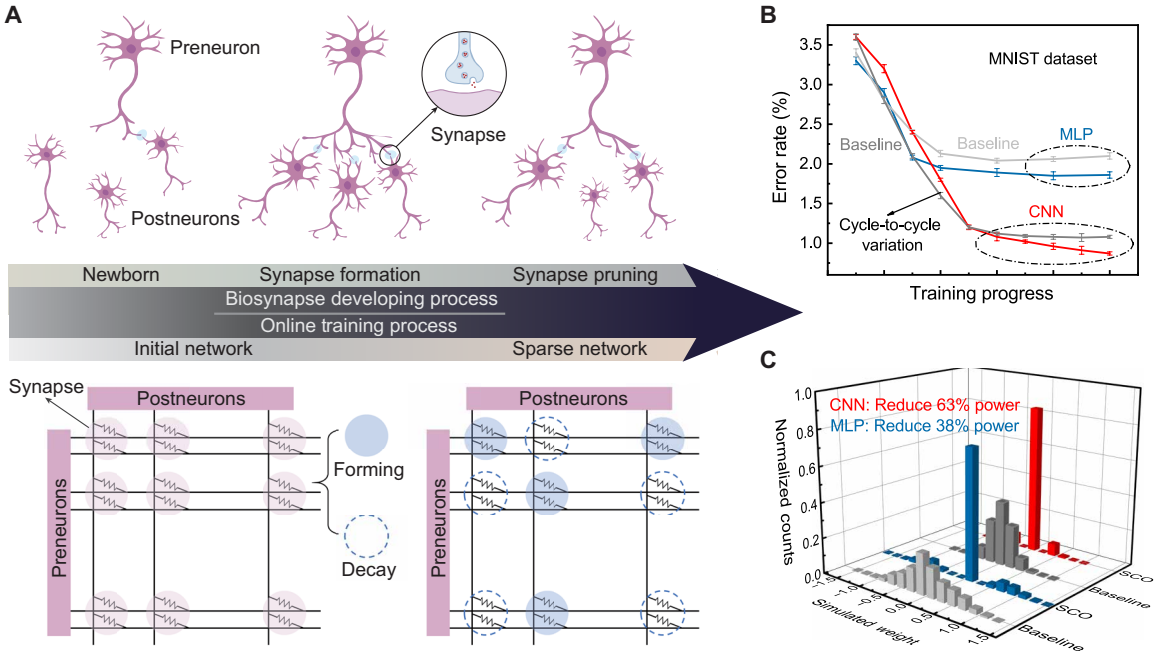

Fig. 5. Implementation of bioplausible neural network pruning. (**A**) The online training with SCO-based TPT-RAM synapses is similar to the selective stabilization of synapse development in the human brain. To start, all the synapses in the neural network are in the diffusive mode. During training, if one synaptic device is frequently updated and the conductance reaches a certain threshold, then it is electroformed into the nonvolatile mode (which implies that the synaptic connection is stabilized). Otherwise, the device conductance spontaneously decays back to almost zero, and the connection is trimmed. (**B**) The recognition error rate is reduced in the pruned neural networks (blue and red) by avoiding overfitting. The highest accuracy achieved is about 99% for the Mixed National Institute of Standards and Technology database (MNIST) recognition. (**C**) The simulated synaptic weight distribution after training. The pruning rates are 73.8 and 84.2% for the MLP and CNN, respectively. In addition, the synaptic power consumption of the network is reduced by 38 and 63% for the pruned MLP and CNN, respectively.

extracted from experiments were taken into consideration in simulation. The results reveal that training with the diffusive nonvolatile dual-mode feature, the pruned network could reduce up to 84.2% of synapses and also save up to 63% in power while improving the accuracy to ~99% for the Mixed National Institute of Standards and Technology database (MNIST) recognition task by avoiding overfitting. These computational advantages brought by TPT-RAM may become more prominent when training a larger neural network with automatic pruning.

DISCUSSION

To sum up, we demonstrated TPT-RAM as a new type of memristive synapse relying on the topotactic phase transition in SCO. The unique oxygen migrations along the highly ordered OVCs led to excellent analog switching characteristics with a much reduced cycle-to-cycle variability of ~0.9% and a device-to-device variability of ~4.9%, the low operation voltage of 0.8 V, and a fast speed below 100 ns. DFT calculations and KMC simulations further confirmed the resistive switching mechanism consistency with the measured device electrical characteristics. These results demonstrated the significance of controlling the ion migration paths to improve the uniformity of RRAM, which is beneficial to guide the optimization of future neuromorphic devices. For future integration with silicon transistors to build functional synaptic arrays based on TPT-RAM, new techniques such as remote epitaxy and sacrificial layer–assisted film transfer could be adopted (*48*, *49*). Furthermore, the SCO-based synapse exhibited a unique diffusive nonvolatile dual mode, which was used to mimic the developing synapses of the human brain and implement neural network pruning during the online training, reducing up to 82.5% redundant synapses and improving the MNIST recognition accuracy to 99%. Our work points out a new direction to design and explore bioplausible analog switching memristive synapses for high-performance neuromorphic computing.

MATERIALS AND METHODS

Growth and characterization of the SCO thin films

Thirty-five–nanometer SCO and 17-nm SRO thin films were grown on a STO (001) and STO (110) substrate by using a reflection high-energy electron diffraction–assisted PLD system. The growth conditions were optimized at the temperature of 750°C with an oxygen environment of 100 mtorr. The laser energy (KrF, λ = 248 nm) was fixed at 1.2 J/cm^2 with the repetition rate of 2 Hz. After the film growth, the samples were cooled down to room temperature at the cooling rate of 7°C/min in the oxygen atmosphere of 100 mtorr. Each sample thickness was controlled by the growth time, and the crystalline structures of the thin films were characterized by XRD and reciprocal space mapping. The atomic structures of the SCO films were characterized using an ARM 200CF (JEOL, Tokyo, Japan) transmission electron microscope.

Device fabrication and characterization

A 20-nm isolation dielectric Al_2O_3 layer was formed by atomic layer deposition, and contact vias were open with the size from 4 × 4 to

$100 \times 100\ \mu m^2$, which defined the active device size. The Au (100 nm in thickness) electrodes with the size of $100 \times 100\ \mu m^2$ were deposited on top of the vias by a magnetron sputtering method. The electrical measurements were performed using a semiconductor parameter analyzer (Agilent B1500), a pulse generator (Agilent 81110A), and a switch matrix (Keithley 707).

Neural network simulation

An MLP of $784 \times 100 \times 10$ and a typical CNN LeNet-5 (*50*) were used to demonstrate neural network pruning. LeNet-5, which consists of three convolution layers, two pooling layers, a fully connected layer, and a radial base function (RBF) layer, is simulated in this work. The input is an image of 32×32 pixels. The first convolutional layer (C1) measures $28 \times 28 \times 6$ after convolution with a kernel of $1 \times 5 \times 5 \times 6$. The result is subsampled by a pooling layer (S2) using a 2×2 average pooling with a sliding stride of 2 and then passed through a sigmoid function. The second convolutional layer (C3) measures $10 \times 10 \times 6$ after convolution with a kernel of $6 \times 5 \times 5 \times 16$. Another subsampling layer (S4) subsequently formed in a similar way as C1 and S2. The third convolutional layer (C5) measures $1 \times 1 \times 120$ after convolution with a kernel of $16 \times 5 \times 5 \times 120$. Then, the outputs are fed into the fully connected layer (F6), which has 84 neurons with tanh activation function (F6). Last, the output layer is composed of 10 Euclidean RBF units for each class.

In the network simulations, a differential pair of two SCO $(110)_{pc}$–based TPT-RAM devices was used to represent one synaptic weight. The neural networks were trained with standard backpropagation and stochastic gradient descent algorithms. The batch size was 200. The learning rate was 0.01. The MLP was trained for 10 epochs, and the CNN was trained for 20 epochs. Initially, all the synaptic devices were in the diffusive mode, and the measured curves in Fig. 4B were used for the simulation. The devices exhibited ~100 conductance levels before forming and were programmed with a write pulse of 3 V/10 ms (interval of 10 ms). To mimic the learning process of human, a series of rest stages were used in the training process. The frequency of rest was once an epoch for MLP and once every two epochs for CNN. In the rest stage, there were no learning tasks, and the synapses that were frequently updated would be electroformed and turned into the nonvolatile state eventually (representing stabilized synaptic connections), while other irrelevant synaptic connections that were still in the diffusive mode would be automatically decayed (fig. S13) to realize the pruning function. The remaining stabilized synaptic connections after pruning were simulated with a pulse of 1 V/1 μs (interval of 1 μs) (Fig. 3B) for neural network training. The read voltage was 0.2 V, and the weights were normalized in the range of (−1,1). The cycle-to-cycle variation extracted from the experimental data and a noise model obeying Gaussian distribution were included in the simulation. The classification accuracies of the network were measured on the test dataset after each rest stage. The pruning rates were calculated as the ratio of the zero synaptic weights, specifically, the number of synapses whose weights are less than 1×10^{-6} divide the total number of synapses. The synaptic power consumption of the network is estimated by $P_{synapse} = (V_{input})^2 \times G_{synapse}$.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/29/eabh0648/DC1

REFERENCES AND NOTES

1. M. A. Zidan, J. P. Strachan, W. D. Lu, The future of electronics based on memristive systems. *Nat. Electron.* **1**, 22–29 (2018).
2. W. Zhang, B. Gao, J. Tang, P. Yao, S. Yu, M.-F. Chang, H.-J. Yoo, H. Qian, H. Wu, Neuro-inspired computing chips. *Nat. Electron.* **3**, 371–382 (2020).
3. H. Wu, P. Yao, B. Gao, W. Wu, Q. Zhang, W. Zhang, N. Deng, D. Wu, H.-S. P. Wong, S. Yu, H. Qian, Device and circuit optimization of RRAM for neuromorphic computing, in *Proceedings of the 2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2017), pp. 11.5.1–11.5.4.
4. D. Ielmini, Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling. *Semicond. Sci. Technol.* **31**, 063002 (2016).
5. H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, K. E. Goodson, Phase change memory. *Proc. IEEE* **98**, 2201–2227 (2010).
6. Z. Wang, H. Wu, G. W. Burr, C. S. Hwang, K. L. Wang, Q. Xia, J. J. Yang, Resistive switching materials for information processing. *Nat. Rev. Mater.* **5**, 173–195 (2020).
7. P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, H. Qian, Fully hardware-implemented memristor convolutional neural network. *Nature* **577**, 641–646 (2020).
8. X. Li, J. Tang, Q. Zhang, B. Gao, J. J. Yang, S. Song, W. Wu, W. Zhang, P. Yao, N. Deng, L. Deng, Y. Xie, H. Qian, H. Wu, Power-efficient neural network with artificial dendrites. *Nat. Nanotechnol.* **15**, 776–782 (2020).
9. Y. Xi, B. Gao, J. Tang, A. Chen, M. F. Chang, X. S. Hu, J. Van Der Spiegel, H. Qian, H. Wu, In-memory learning with analog resistive switching memory: A review and perspective. *Proc. IEEE* **109**, 14–42 (2021).
10. S. Yu, Neuro-inspired computing with emerging nonvolatile memorys. *Proc. IEEE* **106**, 260–285 (2018).
11. B. Gao, H. Wu, J. Kang, H. Yu, H. Qian, Oxide-based analog synapse: Physical modeling, experimental characterization, and optimization, in *Proceedings of the 2016 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2016), pp. 7.3.1–7.3.4.
12. J. Tang, F. Yuan, X. Shen, Z. Wang, M. Rao, Y. He, Y. Sun, X. Li, W. Zhang, Y. Li, B. Gao, H. Qian, G. Bi, S. Song, J. J. Yang, H. Wu, Bridging biological and artificial neural networks with emerging neuromorphic devices: Fundamentals, progress, and challenges. *Adv. Mater.* **31**, 1902761 (2019).
13. Y. Fang, Z. Yu, Z. Wang, T. Zhang, Y. Yang, Y. Cai, R. Huang, Improvement of HfOx-Based RRAM device variation by inserting ALD TiN buffer layer. *IEEE Electron Dev. Lett.* **39**, 819–822 (2018).
14. R. Jiang, Z. Han, X. Du, Reliability/uniformity improvement induced by an ultrathin TiO_2 insertion in Ti/HfO_2/Pt resistive switching memories. *Microelectron. Reliab.* **63**, 37–41 (2016).
15. S. Choi, S. H. Tan, Z. Li, Y. Kim, C. Choi, P.-Y. Chen, H. Yeon, S. Yu, J. Kim, SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations. *Nat. Mater.* **17**, 335–340 (2018).
16. F. Zhang, H. Zhang, S. Krylyuk, C. A. Milligan, Y. Zhu, D. Y. Zemlyanov, L. A. Bendersky, B. P. Burton, A. V. Davydov, J. Appenzeller, Electric-field induced structural transition in vertical $MoTe_2$- and $Mo_{1-x}W_xTe_2$-based resistive memories. *Nat. Mater.* **18**, 55–61 (2019).
17. O. T. Tambunan, K. J. Parwanta, S. K. Acharya, B. W. Lee, C. U. Jung, Y. S. Kim, B. H. Park, H. Jeong, J.-Y. Park, M. R. Cho, Y. D. Park, W. S. Choi, D. W. Kim, H. Jin, S. Lee, S. J. Song, S.-J. Kang, M. Kim, C. S. Hwang, Resistance switching in epitaxial $SrCoO_x$ thin films. *Appl. Phys. Lett.* **105**, 063507 (2014).
18. H.-B. Li, N. Lu, Q. Zhang, Y. Wang, D. Feng, T. Chen, S. Yang, Z. Duan, Z. Li, Y. Shi, W. Wang, W.-H. Wang, K. Jin, H. Liu, J. Ma, L. Gu, C. Nan, P. Yu, Electric-field control of ferromagnetism through oxygen ion gating. *Nat. Commun.* **8**, 2156 (2017).
19. V. R. Nallagatla, J. Kim, K. Lee, S. C. Chae, C. S. Hwang, C. U. Jung, Complementary resistive switching and synaptic-like memory behavior in an epitaxial $SrFeO_{2.5}$ thin film through oriented oxygen-vacancy channels. *ACS Appl. Mater. Interfaces* **12**, 41740–41748 (2020).
20. V. R. Nallagatla, T. Heisig, C. Baeumer, V. Feyer, M. Jugovac, G. Zamborlini, C. M. Schneider, R. Waser, M. Kim, C. U. Jung, R. Dittmann, Topotactic phase transition driving memristive behavior. *Adv. Mater.* **31**, 1903391 (2019).
21. J. Tian, H. Wu, Z. Fan, Y. Zhang, S. J. Pennycook, D. Zheng, Z. Tan, H. Guo, P. Yu, X. Lu, G. Zhou, X. Gao, J. M. Liu, Nanoscale topotactic phase transformation in SrFeOx epitaxial thin films for high-density resistive switching memory. *Adv. Mater.* **31**, 1903679 (2019).
22. C. Mitra, T. Meyer, H. N. Lee, F. A. Reboredo, Oxygen diffusion pathways in brownmillerite $SrCoO_{2.5}$: Influence of structure and chemical potential. *J. Chem. Phys.* **141**, 084710 (2014).
23. J. H. Yoon, J. H. Han, J. S. Jung, W. Jeon, G. H. Kim, S. J. Song, J. Y. Seok, K. J. Yoon, M. H. Lee, C. S. Hwang, Highly improved uniformity in the resistive switching parameters of TiO_2 thin films by inserting Ru nanodots. *Adv. Mater.* **25**, 1987–1992 (2013).
24. S. K. Acharya, R. V. Nallagatla, O. Togibasa, B. W. Lee, C. Liu, C. U. Jung, B. H. Park, J. Y. Park, Y. Cho, D. W. Kim, J. Jo, D. H. Kwon, M. Kim, C. S. Hwang, S. C. Chae, Epitaxial brownmillerite oxide thin films for reliable switching memory. *ACS Appl. Mater. Interfaces* **8**, 7902–7911 (2016).

25. N. Lu, P. Zhang, Q. Zhang, R. Qiao, Q. He, H.-B. Li, Y. Wang, J. Guo, D. Zhang, Z. Duan, Z. Li, M. Wang, S. Yang, M. Yan, E. Arenholz, S. Zhou, W. Yang, L. Gu, C.-W. Nan, J. Wu, Y. Tokura, P. Yu, Electric-field control of tri-state phase transformation with a selective dual-ion switch. *Nature* **546**, 124–128 (2017).
26. H. Jeen, W. S. Choi, M. D. Biegalski, C. M. Folkman, I. C. Tung, D. D. Fong, J. W. Freeland, D. Shin, H. Ohta, M. F. Chisholm, H. N. Lee, Reversible redox reactions in an epitaxially stabilized $SrCoO_x$ oxygen sponge. *Nat. Mater.* **12**, 1057–1063 (2013).
27. H. Jeen, W. S. Choi, J. W. Freeland, H. Ohta, C. U. Jung, H. N. Lee, Topotactic phase transformation of the brownmillerite $SrCoO_{2.5}$ to the perovskite $SrCoO_{3-\delta}$. *Adv. Mater.* **25**, 3651–3656 (2013).
28. Q. Zhang, X. He, J. Shi, N. Lu, H. Li, Q. Yu, Z. Zhang, L.-Q. Chen, B. Morris, Q. Xu, P. Yu, L. Gu, K. Jin, C.-W. Nan, Atomic-resolution imaging of electrically induced oxygen vacancy migration and phase transformation in $SrCoO_{2.5-\sigma}$. *Nat. Commun.* **8**, 104 (2017).
29. L. Yao, S. Inkinen, S. van Dijken, Direct observation of oxygen vacancy-driven structural and resistive phase transitions in $La_{2/3}Sr_{1/3}MnO_3$. *Nat. Commun.* **8**, 14544 (2017).
30. E. Ferreiro-Vila, S. Blanco-Canosa, I. Lucas del Pozo, H. B. Vasili, C. Magén, A. Ibarra, J. Rubio-Zuazo, G. R. Castro, L. Morellón, F. Rivadulla, Room-temperature AFM electric-field-induced topotactic transformation between perovskite and brownmillerite $SrFeO_x$ with sub-micrometer spatial resolution. *Adv. Funct. Mater.* **29**, 1901984 (2019).
31. W. Sun, B. Gao, M. Chi, Q. Xia, J. J. Yang, H. Qian, H. Wu, Understanding memristive switching via in situ characterization and device modeling. *Nat. Commun.* **10**, 3453 (2019).
32. M. D. Rossell, O. I. Lebedev, G. Van Tendeloo, N. Hayashi, T. Terashima, M. Takano, Structure of epitaxial $Ca_2Fe_2O_5$ films deposited on different perovskite-type substrates. *J. Appl. Phys.* **95**, 5145–5152 (2004).
33. J. Young, J. M. Rondinelli, Crystal structure and electronic properties of bulk and thin film brownmillerite oxides. *Phys. Rev. B Condens. Matter Mater. Phys.* **92**, 174111 (2015).
34. S. Inoue, M. Kawai, N. Ichikawa, H. Kageyama, W. Paulus, Y. Shimakawa, Anisotropic oxygen diffusion at low temperature in perovskite-structure iron oxides. *Nat. Chem.* **2**, 213–217 (2010).
35. J. Jo, V. R. Nallagatlla, S. K. Acharya, Y. Kang, Y. Kim, S. Yoon, S. Lee, H. Baik, S. Han, M. Kim, C. U. Jung, Effects of the heterointerface on the growth characteristics of a brownmillerite $SrFeO_{2.5}$ thin film grown on $SrRuO_3$ and $SrTiO_3$ perovskites. *Sci. Rep.* **10**, 3807 (2020).
36. F. Saib, B. Bellal, M. Trari, Preparation and characterization of the brownmillerite $Sr_2Co_2O_5$ as novel photocatalyst in the hydrogen generation. *Mater. Sci. Semicond. Process.* **63**, 122–126 (2017).
37. F. Pan, V. Subramanian, Kinetic Monte Carlo simulation of resistive switching and filament growth in electrochemical RRAMs, in *Proceedings of the 68th Device Research Conference* (IEEE, 2010), pp. 255–256.
38. B. Gao, H. Wu, W. Wu, X. Wang, P. Yao, Y. Xi, W. Zhang, N. Deng, P. Huang, X. Liu, J. Kang, H. Y. Chen, S. Yu, H. Qian, Modeling disorder effect of the oxygen vacancy distribution in filamentary analog RRAM for neuromorphic computing, in *Proceedings of the 2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2017), pp. 4.4.1–4.4.4.
39. C. Z. Hang, C. Wang, B. Gao, H. Chen, M.-H. Xu, L. Hao, H.-L. Lu, Sub-nanosecond pulse programming and device design strategy for analog resistive switching in HfO_x-based resistive random access memory. *Appl. Phys. Lett.* **114**, 112102 (2019).
40. M. Jerry, P. Y. Chen, J. Zhang, P. Sharma, K. Ni, S. Yu, S. Datta, Ferroelectric FET analog synapse for acceleration of deep neural network training, in *Proceedings of the 2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2017), pp. 6.2.1–6.2.4.
41. M. Zhao, H. Wu, B. Gao, Q. Zhang, W. Wu, S. Wang, Y. Xi, D. Wu, N. Deng, S. Yu, H. Chen, H. Qian, Investigation of statistical retention of filamentary analog RRAM for neuromophic computing, in *Proceedings of the 2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, 2017), pp. 39.4.1–39.4.4.
42. Z. Wang, S. Joshi, S. E. Savel'ev, H. Jiang, R. Midya, P. Lin, M. Hu, N. Ge, J. P. Strachan, Z. Li, Q. Wu, M. Barnell, G. L. Li, H. L. Xin, R. S. Williams, Q. Xia, J. J. Yang, Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nat. Mater.* **16**, 101–108 (2017).
43. L. Wang, Z. Yang, M. E. Bowden, Y. Du, Brownmillerite phase formation and evolution in epitaxial strontium ferrite heterostructures. *Appl. Phys. Lett.* **114**, 231602 (2019).
44. J.-T. Yang, C. Ge, J.-Y. Du, H.-Y. Huang, M. He, C. Wang, H. Bin Lu, G.-Z. Yang, K.-J. Jin, Artificial synapses emulated by an electrolyte-gated tungsten-oxide transistor. *Adv. Mater.* **30**, 1801548 (2018).
45. J.-P. Changeux, A. Danchin, Selective stabilisation of developing synapses as a mechanism for the specification of neuronal networks. *Nature* **264**, 705–712 (1976).
46. S. Han, H. Mao, W. J. Dally, Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding, in *Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME)* (IEEE, 2017), pp. 667–672.
47. S. Anwar, K. Hwang, W. Sung, Structured pruning of deep convolutional neural networks. *ACM J. Emerg. Technol. Comput. Syst.* **13**, 1–18 (2017).
48. Y. Kim, S. S. Cruz, K. Lee, B. O. Alawode, C. Choi, Y. Song, J. M. Johnson, C. Heidelberger, W. Kong, S. Choi, K. Qiao, I. Almansouri, E. A. Fitzgerald, J. Kong, A. M. Kolpak, J. Hwang, J. Kim, Remote epitaxy through graphene enables two-dimensional material-based layer transfer. *Nature* **544**, 340–343 (2017).
49. D. Lu, D. J. Baek, S. S. Hong, L. F. Kourkoutis, Y. Hikita, H. Y. Hwang, Synthesis of freestanding single-crystal perovskite films and heterostructures by etching of sacrificial water-soluble layers. *Nat. Mater.* **15**, 1255–1260 (2016).
50. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
51. QuantumATK version Q-2019.12, Synopsys QuantumATK; www.synopsys.com/silicon/quantumatk.html.
52. H. Yeon, P. Lin, C. Choi, S. H. Tan, Y. Park, D. Lee, J. Lee, F. Xu, B. Gao, H. Wu, H. Qian, Y. Nie, S. Kim, J. Kim, Alloying conducting channels for reliable neuromorphic computing. *Nat. Nanotechnol.* **15**, 574–579 (2020).
53. S. Chen, M. R. Mahmoodi, Y. Shi, C. Mahata, B. Yuan, X. Liang, C. Wen, F. Hui, D. Akinwande, D. B. Strukov, M. Lanza, Wafer-scale integration of two-dimensional materials in high-density memristive crossbar arrays for artificial neural networks. *Nat. Electron.* **3**, 638–645 (2020).
54. R. Yang, H. Li, K. K. H. Smithe, T. R. Kim, K. Okabe, E. Pop, J. A. Fan, H.-P. Wong, Ternary content-addressable memory with MoS_2 transistors for massively parallel data search. *Nat. Electron.* **2**, 108–114 (2019).
55. S. Park, A. Sheri, J. Kim, J. Noh, J. Jang, M. Jeon, B. Lee, B. R. Lee, B. H. Lee, H. Hwang, Neuromorphic speech systems using advanced ReRAM-based synapse, in *Proceedings of the 2013 EEE International Electron Devices Meeting (IEDM)* (IEEE, 2013), pp. 625–628.
56. K. Ding, J. Wang, Y. Zhou, H. Tian, L. Lu, R. Mazzarello, C. Jia, W. Zhang, F. Rao, E. Ma, Phase-change heterostructure enables ultralow noise and drift for memory operation. *Science* **366**, 210–215 (2019).

Acknowledgments

Funding: This work was, in part, supported by the National Natural Science Foundation of China (91964104, 61974081, 51872155, and 52025024) and the National Basic Research Program of China (2016YFA0301004). **Author contributions:** X.M., J.T., Y.L., and P.Y. conceived and designed the experiments. Y.L., Y. Zhou, and P.Y. contributed to the thin-film growth and characterizations. W.S., H.W., and H.Q. contributed to the transmission electron microscopy analysis. X.M. and Y. Zhong performed the experiments and data analysis. S.Y., W.L., F.X., B.G., and M.X. performed the device modeling and simulation. Q.Z. carried out the neural network simulations. X.M. and J.T. wrote the paper. All authors discussed the results and commented on the manuscript. J.T. and P.Y. supervised the project. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 14 February 2021
Accepted 2 June 2021
Published 16 July 2021
10.1126/sciadv.abh0648

Citation: X. Mou, J. Tang, Y. Lyu, Q. Zhang, S. Yang, F. Xu, W. Liu, M. Xu, Y. Zhou, W. Sun, Y. Zhong, B. Gao, P. Yu, H. Qian, H. Wu, Analog memristive synapse based on topotactic phase transition for high-performance neuromorphic computing and neural network pruning. *Sci. Adv.* **7**, eabh0648 (2021).