Forming-less compliance-free multistate memristors as synaptic connections for brain-inspired computing

Ng, Sien; John, Rohit Abraham; Yang, Jing-ting; Mathews, Nripan

2020

Ng, S., John, R. A., Yang, J.-t., & Mathews, N. (2020). Forming-less compliance-free multistate memristors as synaptic connections for brain-inspired computing. ACS Applied Electronic Materials, 2(3), 817-826. doi:10.1021/acsaelm.0c00002

https://hdl.handle.net/10356/140531

https://doi.org/10.1021/acsaelm.0c00002

Forming-Less Compliance-Free Multi-State Memristors as Synaptic Connections for Brain-Inspired Computing

Sien Ng[1], Rohit Abraham John[1], Jing-ting Yang[1], Nripan Mathews[1,2]*

[1]School of Materials Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

[2]Energy Research Institute @ NTU (ERI@N), Nanyang Technological University, Singapore 637553

Corresponding Author: Nripan Mathews (nripan@ntu.edu.sg)

Keywords: forming-less, compliance-free, analog memory window, multi-state memristors, artificial synapses

Abstract:

Hardware realization of artificial neural networks (ANNs) require analog weights to be encoded into the device conductances via blind update and access operations, leveraging Kirchhoff's circuit laws. However, most memristive solutions lag behind in this aspect due to numerous device non-idealities like limited number of addressable states, need for a stringent compliance current control and an electroforming process. By modulating the oxygen vacancy profile of tin oxide switching elements, here we design and evaluate multi-state memristors as synaptic connections for brain-inspired computing. Harnessing the advantages of a forming-less compliance-free operation, our devices display gradual switching transitions across multiple conductance states, sufficing the switching requirements of synaptic connections in an ANN. The soft boundary conditions are analysed systematically, and spike-based plasticity rules, state-dependent spike-timing-dependent-plasticity (STDP) modulations, ternary digital logic and analog updatability schemes are proposed and demonstrated comprehensively to establish the analog programming window of our memristors.

Introduction:

With the advent of artificial intelligence (AI) and machine learning (ML), conventional hardware design based on the von Neumann architecture is progressively facing a bottleneck due to the increased data transfer between the separated processing and memory units[1]. In contrast to the conventional serial processing achieved with today's von-Neumann architectures, the human brain utilizes highly parallel, event-driven, and energy-efficient architectures to achieve computational power of the order of 10^{18} FLOPS (Floating Point Operations per second) at a power consumption of ~20W[2]. Neuromorphic computing- an approach where electronic analog devices and circuits mimic neuro-biological architectures of the human brain, promises to dramatically improve the efficiency of important computational tasks such as perception and decision-making via computation-in-memory[3]. Efficient training of such artificial neural networks (ANNs) hinges on the incremental adjustment of weighted connections between the hidden layers to minimize the cost function of a gradient descent optimization based on back-propagation algorithms[4]. This in turn demands deployment of analog non-volatile memory elements with multiple conductance states and smooth conductance transitions as weighted connections to store and update weights/conductances in a manner congruent with the learning algorithm[5].

Despite the remarkable advancement of standard programmable architectures based on CMOS, innovative neuromorphic hardware circuitry is needed to emulate the scale, connectivity and power effectiveness of biological neural networks[6,7]. Very recently, hysteretic field effect transistors have been reported to extensively emulate the complex signal processing of biological synapses[8–10], but they lag behind in terms of scalability and memory retention. With memory of operational history, excellent non-volatility, nanometer-level scalability, low energy consumption and ultrafast switching speeds, memristive devices are touted as potential building blocks for brain-inspired computing architectures and have been extensively investigated in recent years[11]. Despite their excellent properties, most implementations fall behind in terms of reliable analog switching properties, namely- multiple memory states and non-abrupt state transitions- a prerequisite for efficient training of high-performance neural networks[3].

Specifically in the case of oxide valence change memristors, formation and disruption of highly conductive filaments composed of atomic defects like oxygen vacancies often result in abrupt state transitions and uncontrollably high current levels, requiring stringent compliance current control via additional transistors to prevent device damage[12]. The addition of control transistors makes the architecture area- and power-inefficient and also makes the peripheral circuitry used to address the memory complicated. Additionally, most of these oxide memristors also require a forming step to initiate a stable switching behaviour. This can vary in magnitude from device to device and hence result in undesirable power- and time-consuming steps to address the non-volatile states online during training and inference[13]. Most critically, the abrupt switching physics of oxygen vacancy migration often limits the number of addressable states in memristors to two (1 low resistance state (LRS) and 1 high resistance state (HRS)), limiting their plasticity and storage capacity, and adversely affecting the trainability of ANNs[14]. Therefore, there is a need to explore additional systems to further improve the bit precision, storage capacity and updatability required for energy-efficient in-memory computations.

In an attempt to solve these issues, here we design and evaluate multi-state memristors based on tin oxide as synaptic connections for brain-inspired computing. We utilize an ultra-thin oxide switching layer (SnO_x) with a rich reservoir of oxygen vacancies for stable and repeatable formation of nano conductive filaments, resulting in a forming-less operation. More critically, the oxygen-gettering ability of the top electrode (Al) is harnessed to create a

resistive oxide interfacial gradient that enables gradual switching between conductance states. This also acts as a resistor in series, self-limiting the growth of nano conductive filaments and current flow, overcoming the need for compliance current control in our devices. The deposited thin film has been characterized to be conformal and the thickness is verified by cross-sectional imaging. The fabricated multi-state memristors are first evaluated by DC I-V sweeping tests, demonstrating its operational history and reversible switching memory. As a demonstration of the multi-level switching characteristics, we program the device with different operational pulses to multiple stable conductance states portrayed as weighted synaptic connections for brain-inspired hardware neural networks (HNNs). To further understand the dynamics of the switching behaviour, the weight changes are subjected to a non-linear model, critically identifying the digital and analogue regime in the switching characteristics. Having evaluated the analog programming window of these devices, writing schemes are next optimized to tap on the conductance soft boundaries to study neuronal features such as short- and long-term plasticity. A state-dependent spike-timing-dependent-plasticity (STDP) modulation is established utilizing the soft boundary conditions of our memristors. As a final demonstration for this concept, a ternary digital logic and an analog updatability scheme is proposed and implemented utilizing ramping-profile input pulses.

Results:

DC I-V Characteristics:

Figure 1a, S1 shows the resistive switching operation of the proposed Al/SnO_x/FTO memristor across 100 consecutive DC sweeping cycles (0V$\rightarrow$ +2V$\rightarrow$ 0V$\rightarrow$ -2.5V $\rightarrow$ 0V, step size=10mV). The voltage sweeps are applied across the top Al electrode with the bottom electrode (FTO) always grounded. The device switches (SET) seamlessly from its high resistance state (HRS~1mS) to low resistance state (LRS~11mS) without any electroforming process when the voltage is swept from 0V$\rightarrow$ +2V$\rightarrow$ 0V. In the reverse sweep 0V$\rightarrow$ -2.5V $\rightarrow$ 0V, the device RESETs backs to its initial HRS, again in a gradual manner. Both the SET and RESET processes depict non-abrupt/gradual transitions without the need for compliance current control as evident from Figure 1a, while the high cyclability reflects the excellent stability of the switching process.

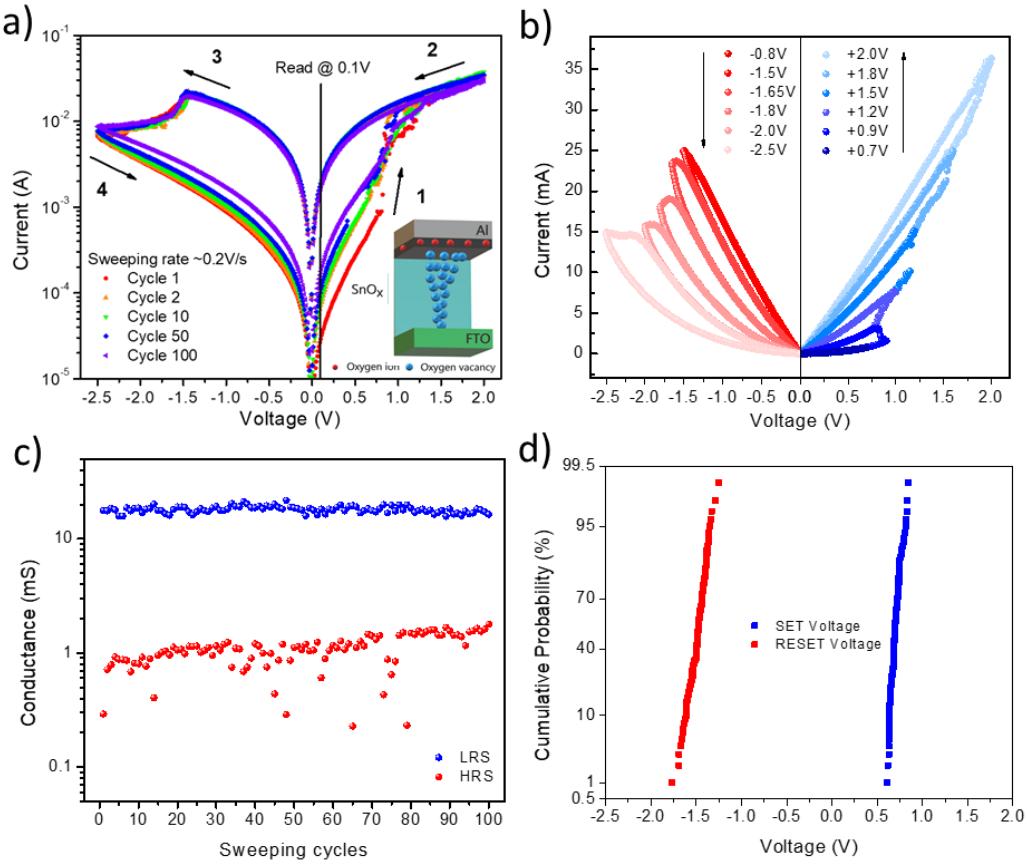

Figure 1. Memristor operation exhibiting gradual switching characteristics. (a) DC I-V switching characteristics of the Al/SnO_x/FTO memristor over 100 consecutive cycles. (b) Voltage-dependent switching with consecutive positive (blue) and negative voltages (red), showing gradual switching transitions across multiple conductance states. (c) Endurance characteristics of our devices over 100 sweeping cycles. The conductance window is read at 0.1V. (d) Cumulative probability distribution function of the SET and RESET voltages across 100 cycles.

Staying within this voltage window, the SET and RESET transitions are gradual and highly tunable. Noticeably within this window, non-abruptness of the conductance transitions is independent of the voltage amplitude. To demonstrate the highly-modulatable non-abrupt SET process, the device is subject to voltage sweeps varying from 0V→ +0.7 to +2V→ 0V; while 0V→ -0.8 to -2.5V → 0V portrays the gradual RESET process (Figure 1b). In all these cases, absence of a forming process avoids undesirable power- and time-consuming steps to address the non-volatile states, increasing the efficiency of the HNN during training and inference[13]. And absence of a compliance current control bypasses the need for current-limiting transistors, increasing its area- and power-efficiency[15].

Stability of the individual states, retention and dynamic range becomes cardinal evaluation parameters in the case of such multilevel memories, since the continuously modulated

conductance states lie close to each other or in other words the on-off ratio between the intermediate conductance levels is small[16]. A high dynamic range would ensure improved device reliability, increased number of distinct addressable states would improve the memory storage capabilities, while excellent retention properties would ensure stable device operation. Figures 1c-d depicts the excellent endurance characteristics and cumulative probability distribution function of SET and RESET voltages of our devices programmed in the window +2.0V (potentiation)/-2.5V (depression). The devices maintain a good on-off ratio of >10 and depict a narrow distribution of the SET-RESET voltages over 100 sweeping cycles, reiterating the stability of the switching process.

Working Principle:

Migration of oxygen vacancies under an applied electrical field is commonly recognized as the switching mechanism in oxide-based valence change memristors[17,18]. But the comprehensive switching mechanism still remains unsure and could be classified into interface or filamentary depending on the device structure and electrical characteristics exhibited by the device[19]. The gradual switching behavior in both the SET and RESET regions (Figures 1a-b, S2a-b) indicate an interfacial switching mechanism due to the formation of an interlayer at the electrode-oxide junction in our devices[20]. Absence of an electroforming process and compliance current control strengthens this inference. Logarithmic I–V curve plots and linear fittings of the SET process reveals an ohmic behavior of the HRS at low voltages and a space charge limited current (SCLC) conduction at higher voltages. The LRS is dominated by an ohmic conduction (Figure S2a-b), in agreement with other interfacial memristive devices. On the other hand, abrupt transitions in the pulsing characteristics (discussed below in detail) also indicate that filamentary mechanism plays a pivotal role as well[21]. A dynamic conductance analysis similar to Y. F. Chang et al.[22] was conducted to understand the contribution of an internal filament to the self-compliance current behavior. In Figure S2c-d, the dynamic conductance is tabulated by taking the derivative of current with respect to the voltage step (0.01V) during a sweeping cycle. During a SET sweeping cycle (Figure S2c), the dynamic conductance increases monotonously before abrupt transition, signaling the onset of a SET process. At the SET transition, the dynamic conductance reaches a peak before stabilizing to a higher conductance state. With the increase in sweeping voltage each cycle, an increase of SET threshold is observed; such self-limiting phenomenon is crucial in the compliance-free operations. Similarly, in the RESET sweeping cycle (Figure S2d), the dynamic conductance reaches a peak before depressing to a

negative dynamic conductance. The intercept at zero dynamic conductance marks the self-compliance current limit and the onset of the RESET process. With the increase in sweeping voltage each cycle, a decrease in maximum dynamic conductance and RESET threshold is observed. Hence, we hypothesize the existence of both interfacial and filamentary switching mechanisms co-existing in our devices (Figure S1). The SnO_x switching layer provides a rich reservoir of oxygen vacancies for stable and repeatable formation of nano conductive filaments. XPS measurements corroborate the proposed creation of oxygen vacancies as shown in Figure S4. For O 1s peaks, the signal at ~529.0eV is attributed to the low binding energy of fully-coordinated lattice oxygen of SnO_2, while the left-shifted signal at ~530.0eV with higher binding energy is attributed to the oxygen vacancies. As evident from Figures S4a and c, a significant increase in the concentration of oxygen vacancies (from 50% to 59%) accompanies the switching transition from the pristine/HRS to LRS, corroborating the proposed hypothesis of the formation and rupture of nano-conductive filaments composed of oxygen vacancies, determining the memristive switching process. Additionally, the oxygen-gettering ability of Al creates a resistive interfacial gradient of vacancies at the Al-SnO_x interface in series with the active switching matrix. On application of a positive voltage at the top electrode, oxygen vacancies at the Al-SnO_x interface migrate through the SnO_x switching layer forming several nano-sized conductive filaments bridging across the device. The rich reservoir of oxygen vacancies makes this process seamless without the need of a forming process. The interfacial resistor becomes more resistive during this process, in turn self-limiting the size of filament and current flow, alleviating the need for compliance current control. These vacancies get filled in a gradual manner by the movement of oxygen ions during the RESET phase, resulting in multiple conductance states and gradual switching between these states. To shed further light on the RESET conduction mechanism, we evaluate the logarithmic I–V curve plots of multiple RESET operations performed on the same device. Transition from an ohmic to a space charge conduction limited (SCLC) conduction (Figure S2) reflects a gradual dissolution of filaments in the analog RESET processes, in accordance with our hypothesized switching mechanism[15].

Multi-level memory- Analysis of the soft boundaries for analog programming:

Ideally, memristors deployed as weighted synaptic connections in a neural network must exhibit gradual transitions across multiple conductance states[23]. The overall efficiency and learning ability of a neural network has been shown to significantly depend on the number of conductance levels, conductance linearity, write noise and multi-level cell characteristics.

Burr et al.[24] recently proposed up to 6-bit resolution between the minimum and maximum conductance with a conductance pair representing a single weight for a deep neural network; while at least a dozen of separate conductance levels was predicted to be necessary for recognition of MNIST handwritten digits by Querlioz et al[25]. But most implementations depict a binary switching logic with 1 non-modulatable LRS and HRS respectively, limiting their plasticity and storage capacity, and adversely affecting the trainability of an ANN. Hence to qualify as weighted synaptic connections in a neural network, the devices are first benchmarked on their degree of plasticity (number of accessible conductance levels and their modulations) as a function of input electrical stimuli.

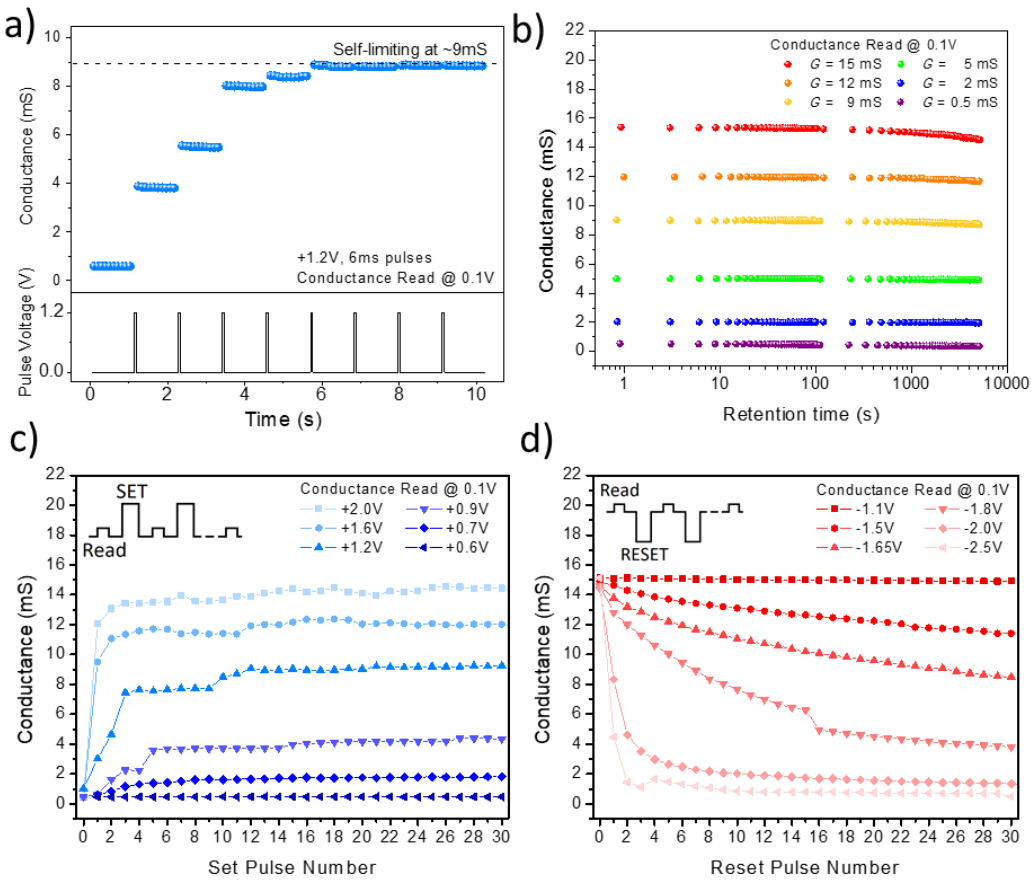

Figure 2. Multi-level memristor operation. (a) Non-volatile weight changes achieved via blind updates depict an asymptotic increase from 0.5mS to 9mS when stimulated with single pulses of amplitude +1.2V and pulse width=6ms. This represents the soft boundary of our devices. (b) Electrical stability and retention characteristics of the multiple conductance states at room temperature. Map of the conductance updates with uniform (c) set and (d) reset pulses.

In congruence with the I-V characteristics shown in Figure 1, the devices depict multiple non-volatile conductance states that asymptotically increase from 0.5mS to 9mS when stimulated with single pulses of amplitude +1.2V and pulse width=6ms (Figure 2a). The

weight changes (stepped increments) are initially large, but decrease quickly with increasing absolute weights, indicating a soft boundary for weight updates. In Figure 2a, the conductance value of 9mS is observed to be a soft boundary and the devices can be pushed beyond this value by optimizing the amplitude and number of repetitions of the input electrical stimuli as shown in Figure 2b. Excellent non-volatility of the individual conductance states, indicated by the large retention times (~$5x10^3$ seconds) and a narrow distribution of the individual weights (Figure 2b), portray their advantages over hysteretic transistor configurations. However, while the ability to write and address multiple non-volatile states seem highly promising, presence of the soft bound behavior enforces a trade-off between the degree of analog modulation and parameters (amplitude, pulse width) of the input writing schemes. All devices could be recovered back to the initial HRS (~0.5mS) after the various pulse measurements.

To evaluate the bounded nature of the cumulative conductance change, weight/conductance traces produced by sequences of 30 identical pulses/blind updates are recorded as a function of the number of pulses as shown in Figures 2c-d. Amplitude of the blind updates are varied from +0.6 to +2V to map out the conductance changes or weight updates during potentiation, while variations from -1.1 to -2.5V are mapped for the weight updates during depression. The conductance dynamics evolves from a region of no substantial conductance/weight change to regions of gradual and abrupt switching (from dark to clear color tones) as evident from Figures 2c-d. For example, while input potentiating pulses of +0.6V amplitude do not produce substantial weight changes in our devices, pulse amplitudes of +0.7 to +0.9V result in gradual weight updates. Increasing the pulse amplitude further (>+0.9V) result in more or less abrupt switching, demarcating the soft boundary for analog programming in the potentiation window (Figure 2c). Similarly, in the depression phase, the memory window for analog programming is softly bound to amplitudes in the range -1.5 to -1.8V. The devices depict non-updatability below this range, and an abrupt switching behavior above this range as depicted in Figure 2d. These abrupt transitions in the pulsing characteristics (Figure 2), especially in the potentiation window (Figure 2c) and an area-independent switching behavior (Figure S2e), hints at a filamentary mechanism playing a pivotal role as well as explained above[21]. It can also be noticed that the conductivity window expands with the increase in strength of the programming conditions, but at the expense of the smoothness of the weight update. The achievement of conductance saturation for a sufficiently large amount of pulses is another important characteristic prevalent to all conductance curves. As the rate

of saturation approaches asymptotically, the impact generated by a single weight update operation becomes lower. This saturation and non-linear behavior suggest that conductance modulation in such memristors is highly dependent on the current state of the device and reflects the limited linear window for truly analog weight updates, beyond which the devices deviate from linearity.

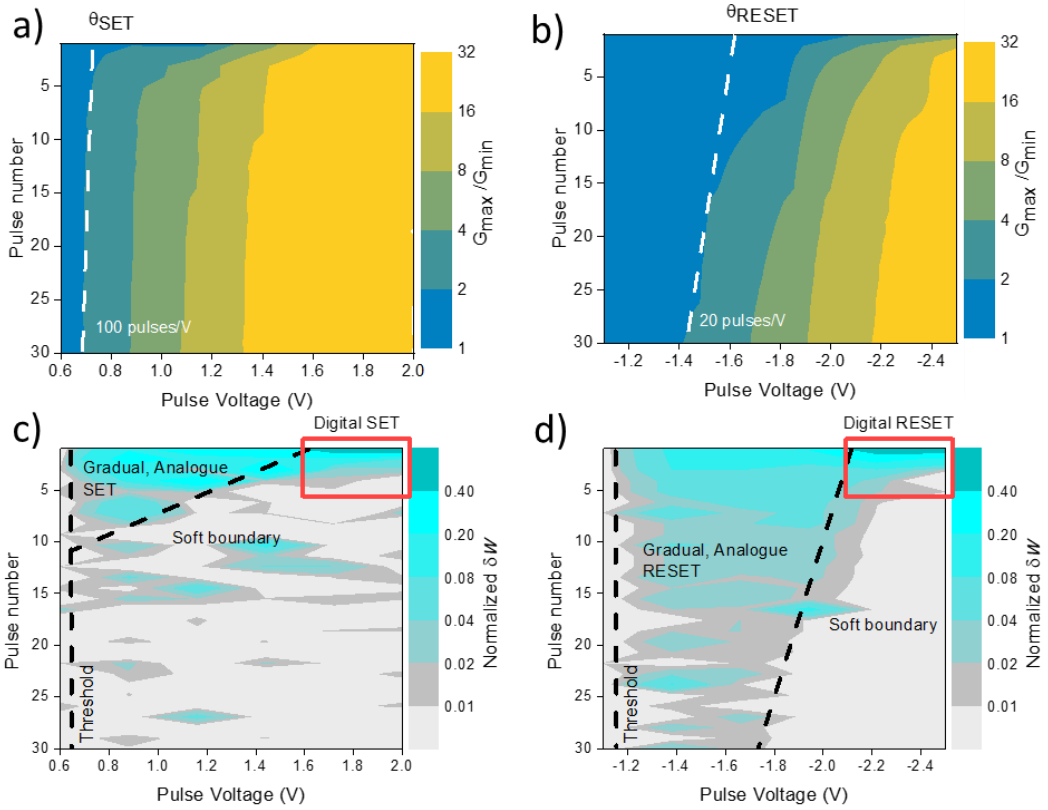

Figure 3. State-dependent modulation and analog programmability window. The conductance window G_{max}/G_{min} with respect to the number and amplitude of the (a) set and (b) reset pulses. The rate of update visualized with normalized weight changes with respect to the number and amplitude of the (c) set and (d) reset pulses.

To illustrate the memory window for analog programmability further, conductance ratios (G_{max}/G_{min}) are plotted as a function of the pulse programming parameters (amplitude and number of repetitions)[26] (Figure 3). A ratio of 1 indicates the inability to update weights in a non-volatile manner, while higher ratios reflect non-volatile conductance changes or weight updates with a retention time > $5x10^3$ seconds. The white dashed line marked in Figures 3a-b (corresponding to a ratio of 2 in conductance) indicates the onset of non-volatile conductance change or weight update and hence represents the threshold for resistive switching. Programming voltages above this threshold result in higher conductance ratios as shown in Figures 3a-b. To assess the differences in switching kinetics in our devices, slope of the

voltage–time (pulse number) relation/function is calculated for both the potentiation and depression phases. The steeper slope of the potentiation phase (100 pulses/V) highlights the faster switching kinetics of the potentiation process as supposed to 20 pulses/V of the depression phase. In general, the origin of such asymmetries resides in the fundamentally distinct switching kinetics of the procedures responsible for potentiation and depression, i.e. the formation and dissolution of filaments. More importantly, this indicates a state-dependent modulation of weight updatability in such memristive devices, often overlooked in most investigations.

To account for this state-dependent weight modulation, we adopt a multiplicative update scheme featuring weight-dependent rules in congruence with the model proposed by Fusi and Abbott[27]. For synaptic weights denoted between 0 and 1, the following generalized soft bound equations are adopted for incremental ($\delta w_+(w)$) and decremental ($\delta w_-(w)$) weight changes in potentiation and depression events, respectively.

$$\delta w_+(w) = \alpha(1 - w)^{\gamma} \quad (1)$$

$$\delta w_-(w) = -\alpha w^{\gamma} \quad (2)$$

where α is a multiplicative parameter that indicates the magnitude of weight modification induced by a plasticity event, and γ reflects the state dependency of the weight update.

The normalized weight changes $\delta w_+(w)$ and $\delta w_-(w)$ are plotted with respect to the pulse number and amplitude of the events as shown in Figures 3c-d respectively. The first vertical dashed line represents the switching threshold below which weight changes are negligible, while the second dashed line demarcates the soft boundary for analog weight updates. To map this region of analog programmability, the distribution of weight changes over the span of 30 input blind update pulses are plotted and analyzed, as depicted in Figures 3c-d respectively. At input voltages slightly higher than the threshold, the weight changes are incremental and continuous, resulting in large number of addressable states. This window is ideal for the operation of ANNs since it allows efficient mapping of the weighted synaptic connections to multiple conductance states governed by mathematical algorithms. However, these weight updates are softly bound to their absolute values as indicated by the large, sudden digital-like updates beyond this window; restricting their operational window. The parameters α and γ extracted from these distribution plots are shown in Figures S5a-b. Below

the threshold voltage, the flat response produces α very close to 0, indicating only one stable effective state. For blind updates slightly above the threshold, α values are small, indicating an effective analog operation regime. Practically, the system is capable of achieving multiple states with effective steps of 1/α. Beyond this analog operation regime, α approaches a maximum value of 1, indicating only 2 effective states or a digital operation. Similarly, the exponential factor γ of 0 indicates that the weight changes are independent of the current state. An exponent slightly higher than 1 indicates a mild dependency of the current state. An average of $\gamma = 1.2$ and $\gamma = 1.3$ is extracted from the potentiation and depression events respectively as shown in Figures S5a-b. From Figures 3c-d, it is evident that the depression events have a larger window of analog programmability when compared to the potentiation events. This again reflects the fundamentally distinct switching kinetics of the procedures responsible for potentiation and depression in our devices and provides a sound guideline to map out the window of analog programmability from an algorithm perspective.

Memristors as Synapses:

Having mapped out the soft boundaries of analog programmability of our memristors, we next benchmark our devices against the most commonly utilized plasticity rules for spiking neural networks. Classified based on the timescales of operation, short and long-term plasticity rules defining learning and memory are studied by recording excitatory post synaptic currents (EPSCs) in response to pre- and post-synaptic training sequences. A train of pulses (+0.6V, 6ms, number=20) at the top electrode (pre-synaptic terminal) evokes a pulsed current response (called Excitatory post synaptic current (EPSC)) of 64μA (equivalent conductance=0.64mS), which immediately decays back to its resting level of ~0.5mS on removal of the presynaptic pulse (Figure S6a). Fitting the transient current response to the below equation, we derive decay time constant (τ) correlated to the forgetting rate to be ~0.3 seconds.

$$G(t) = G_{steady} + A \exp\left(-\frac{t}{\tau}\right) \qquad (1)$$

G (t)- synaptic weights at time t, G_{steady} - synaptic weights at steady state measured 30s after stimulation pulses and A- a pre-exponential constant.

Tuning the stimulation voltage further results in a controlled modulation of τ from 0 to 0.6 seconds respectively as shown in Figure S6a, indicating the modulatable forgetting rates of our devices. Such short-term memory and plasticity rules have been demonstrated to serve as

working memory and mapping matrices for removing auto-correlations and have been very recently adopted to learn fine temporal structure of event-based signals amidst rate-coded events[28]. While short-term plasticity occurring on a time scale of tens of milliseconds is helpful for temporal filtering and may play a part in speech processing, long-term plasticity occurring on a time scale of several seconds to minutes is considered to be the foundation of experience-dependent neural circuit modification[29]. Stimulation with spikes of higher amplitude and pulse width consolidates the weight changes, leading to precisely controlled non-volatile changes in conductance, emulating long-term plasticity features. For example, increasing the amplitude of input spikes from 0.65 to 0.725V (with other parameters remaining fixed- pulse width=6ms, number of repetitions=40, pulse interval=6ms) consolidates the memory from short- (volatile) to long-term (non-volatile) as depicted in Figure S6b. Similar behaviors are observed with devices pre-programmed to lower (Figure S6a) and higher (Figure S6c) initial conductances as summarized in Figure S6d, reiterating the state-dependency of weight modulations in our devices.

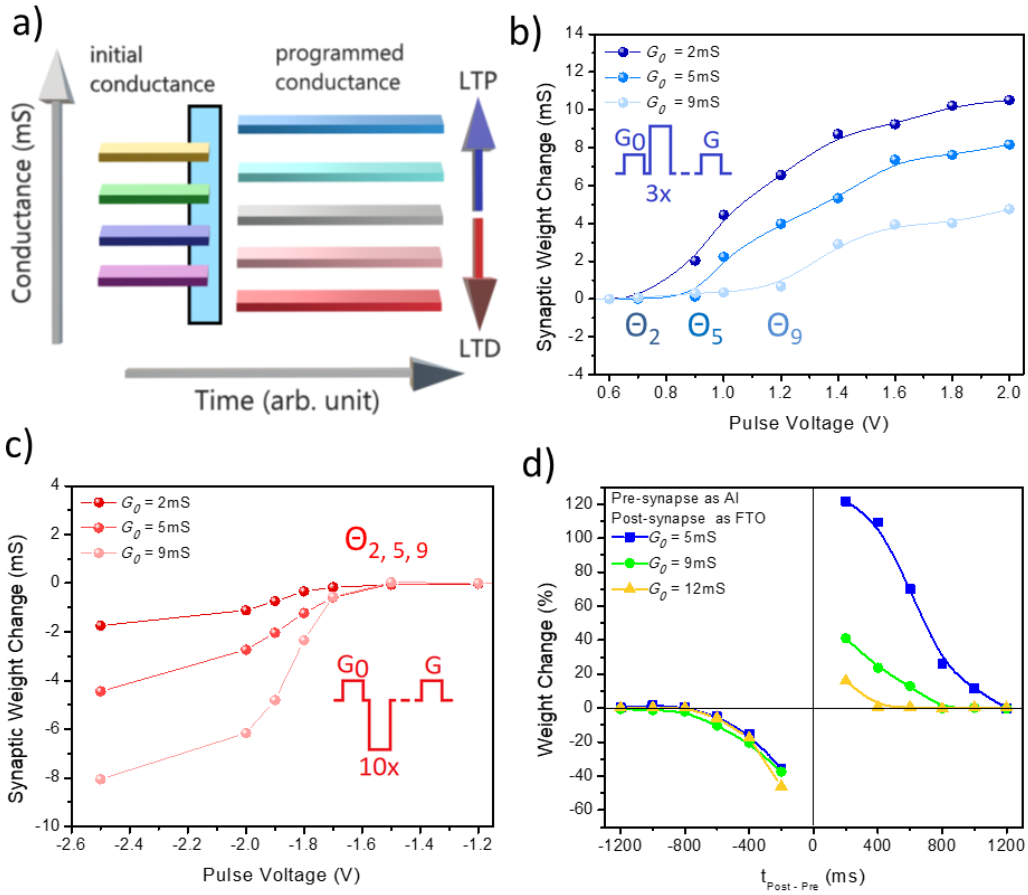

Figure 4. Non-volatile memory operation and its state dependency. (a) Schematic of the experimental flow involving device initialization to various conductance states, followed by the potentiation or depression programming pulses. The weight changes are tabulated with respect to the voltage of pulse trains and initial

conductance state for (b) potentiation operation and (c) depression operation. (d) Spike-timing dependent plasticity with asymmetric Hebbian learning characteristics demonstrated and its state dependency characterized.

Embracing the soft bound limits of analog programmability (Figures 3c-d), we next analyze the short- and long-term plasticity effects as a function of the initial conductance state of the device. The devices are pre-programmed to distinct initial states (0.5, 2 and 5mS) and pulsed current responses are recorded to extract the decay constants of the short-term memory. The devices exhibit a state-dependency in the decay constants as shown in Figure S6. Devices pre-programmed to higher initial conductance states depict longer decay constants and faster consolidation to long-term memory as expected. Non-volatile weight updates crucial for neural networks are next evaluated in detail as a function of the initial conductance states. Devices are pre-programmed to distinct initial states of 2, 5 and 9mS, their conductance changes are mapped as a function of the number of pulses (Figure 4a). Device at the initial state of 2mS exhibit the highest rate of weight change in the potentiation phase when compared to devices pre-programmed to higher initial conductances (5 and 9mS). Figures 4b-c depicts this response clearly by plotting the absolute synaptic weight changes against the input pulse voltage. The depression phase demonstrates an opposite trend with the device pre-programmed to 9mS exhibiting the highest rate of weight change. In conclusion, highly resistive devices or devices with a low initial conductance exhibits larger plasticity window in the potentiation phase, while devices programmed to a high initial conductance exhibits larger plasticity window in the depression phase.

Translating these weight changes to a spiking neural network domain, we next characterize the spike-timing-dependent-plasticity (STDP) properties of our devices. A refinement of Hebb's theory, STDP is regarded as the first law of synaptic plasticity and is the foundation of associative learning[30]. Different types of STDP have been noted in biological synapses and are ascribed to distinct data processing and storage tasks[31]. To characterize this plasticity rule, spike patterns corresponding to Figure S7 are applied to the pre- and post-synaptic terminals as indicated, and the change in conductance (weight) is recorded as a function of the pulse interval between pre- and postsynaptic spikes[9,10,32]. Temporal correlations between the pre- and postsynaptic spikes create voltage-dependent changes in conductance/weight, establishing the STDP rules. Repeated arrival of pre-post or post-pre spike pairs lead to resistance changes above the threshold for non-volatile conductance change in proportion to

the voltage and time-integrated device conductance function ($f(V_{pre}\text{-}V_{post})$, t), where the net voltage on the device at each instant of time (t) is defined by the voltage difference between the pre- and post-spike ($V_{pre}–V_{post}$). Changes in conductance are compared to the initial conductance value to convert the data to percentage weight changes (reading pulses of 0.1V is utilized for this measurement). The device is then allowed to relax back or erased to the initial conductance state (e.g. 5mS) before the next measurement to avoid dependence of previous history. For example, an interval ($t_{post\text{-}pre}$) of +200ms results in a net voltage of $V_{pre\text{-}Vpost}$ = (+0.70) – (-0.80) = 1.5V developed across the device, triggering a permanent increase in the channel conductance or LTP (~121 %). On arrival of presynaptic pulses after postsynaptic pulses, i.e. $t_{post\text{-}pre}$ of -200 ms, the maximum net voltage developed across the device is $V_{pre\text{-}Vpost}$ = (-1.10) – (+1.20) = -2.3V and this results in a decrease in conductance or LTD (~ -35 %). These measurements are repeated for several combinations of spike intervals and the weight changes are plotted as a function of $t_{post\text{-}pre}$ as shown in Figure 4d. Weight changes are predominant at small pulse intervals, and weakens with increase in the interval, reflecting strong temporal correlations between the pre- and post-synaptic spikes. The STDP time windows shown here in milli-seconds and weight changes are comparable to biological values and could be further tuned by modulating the width, number and shape of the input spikes. As expected, the magnitude of weight changes also depicts a strong dependence on the initial device conductance as shown in Figure 4d. In congruence with the above measurements and observations, devices pre-programmed to lower initial conductance states depict a higher plasticity window in the potentiation phase and vice versa in the depression phase (Figure 4d). Mapping the window of analog programmability and understanding of the analogue conductance update under different programming conditions is necessary to effectively develop software algorithms to run memristive HNNs. Bound conductance values and state-dependent modulation are key factors that affect the storage capacity and performance of the neural network, but is very often ignored. This work offers a comprehensive guideline for establishing and optimizing the analog programming window of memristive devices in the perspective of HNNs.

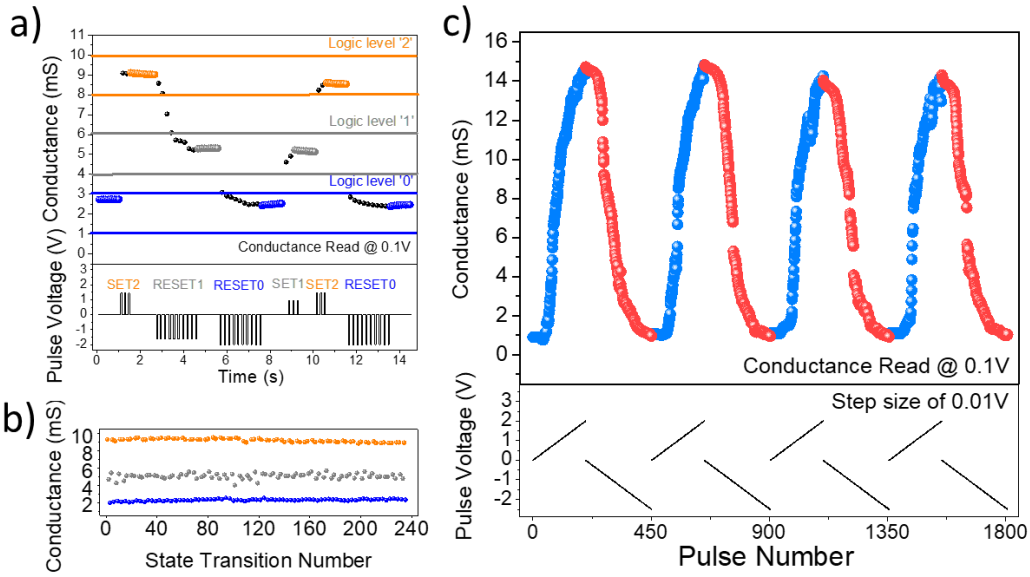

Figure 5. (a) Ternary digital switching between 3 multilevel states using both intermediate and abrupt, setting and resetting pulses. All set operations are in train of 3 pulses and reset operations are in train of 10 pulses. 'SET2' is a setting pulse to state '2' regardless of initial conductance state while 'SET1' is a setting pulse to state '1' from state '0'. Likewise for resetting pulses 'RESET0' and 'RESET1'. (b) Cycling of 234 transitions between '0' , '1' and '2' states. (c) A continuous variation of analog synaptic weights across a memory window of ~10 using incremental pulses.

As a final illustration of the multi-level programmability of our memristive devices, we demonstrate a highly stable ternary digital logic and an analog updatability scheme as shown in Figures 5a-c. Ternary switching is an utility of state-dependent weight changes and demonstrates how the compliance-free multi-state memristor can achieve stable switching between 3 distinctly accurate programmable state without transitioning through any intermediate state. For the ternary digital logic, we split the conductance band into 3 distinct levels denoted by logic '0' (1-3mS), '1' (4-6mS) and '2' (8-10mS) respectively. Blind input update schemes are optimized to seamlessly switch between these conductance states as shown in Figure 5b. For example, state transitions from logic '0' to '1' are SET ("SET1") via pulses of amplitude 1.1V and pulse width 6ms and transitions from logic '1' to '2' are achieved via 1.55V, 6ms pulses ("SET2"). More importantly, direct transitions from logic '0' to '2' are also mapped via 1.55V, 6ms pulses as denoted by "SET2" in Figure 5a, indicating the independence of these transitions from the initial state due to the presence of soft boundaries. These represent the potentiation part of the update schemes. Similarly, direct and indirect depression transitions from logic '2' to '0' are also mapped, as denoted by "RESET1" and "RESET0". To demonstrate the highly stable switching behavior of our

devices, they are cycled continuously between each of these logic states for ~240 cycles. The device switches seamlessly between the 3 logic states without any signs of degradation as shown in Figure 5b. Being compliance-free, the complete ternary memristor is free from any initialization step[33] to the low resistance state before stabilizing to the desired state. (see Figure S8) Compared to the traditional binary logic enabled by memristor systems, ternary logics with greater information-carrying ability can provide greater computational effectiveness with decreased circuit complexity[9,10].

A linear and symmetric weight update behavior of the analog synapse is critical for achieving high learning accuracy in artificial neural networks based on the back-propagation learning rules. To unlock this potential, requires analog weights to be encoded into the device conductances via blind update (write) and access (read) operations, accelerated in parallel via simple Kirchhoff's circuit laws[3]. Hence going beyond the ternary digital logic, we demonstrate analog programmability in our devices across the window 0.5 to 15mS with gradual synaptic weight updates utilizing ramping nonidentical pulses of step size 10mV (Figure 5c). While this strategy complicates the peripheral circuit design and increases the latency and power consumption, the ability to address multiple linearly-distributed conductance states could alleviate the addressing requirements at an algorithm level and enhance the in-memory computing efficiency when compared to traditional binary systems with limited non-linear conductances[9,10].

Our Al/SnO_x/FTO multi-state memristor device demonstrated good endurance of more than 100 cycles, comparable to present efforts in multi-state memristors (see Table S1). The weaker RESET process, probably due to oxide induced interface can also be partially recovered by increasing the number of pulses[34]. By using more optimized triangular-shaped pulses with less capacitative overshoot, we hope that the switching endurance can be further improved[35].

Conclusion:

Here we design and evaluate multi-state memristors based on tin oxide as synaptic connections for brain-inspired computing. The ultra-thin oxide switching layer (SnO_x) provides a rich reservoir of oxygen vacancies for stable and repeatable formation of nano conductive filaments, resulting a forming-less operation. More critically, the oxygen-

gettering ability of the top electrode (Al) creates a resistive oxide interfacial gradient that enables gradual switching between conductance states. This also acts as a resistor in series with the active switching matrix, self-limiting the growth of nano conductive filaments and current flow, alleviating the need for compliance current control in our devices. Both these features facilitate writing and addressing of multiple conductance states in response to pulsed inputs, making them eligible as weighted synaptic connections in brain-inspired hardware neural networks. However the stochastic nature of migration of oxygen vacancies results in a soft boundary for analog weight updates, restricting their memory window for neuromorphic computing. We investigate these soft boundary conditions in detail to evaluate the effective storage capacity and performance of our memristors when deployed as synaptic connections in a neural network. Neuronal features such as short- and long-term plasticity are studied extensively and a state-dependent modulation is established to effectively utilize the soft boundary conditions of our memristors. As a final illustration of the multi-level programmability of our memristive devices, we demonstrate a highly stable ternary digital logic and an analog updatability scheme utilizing ramping profile pulses. The systematic soft-boundary estimation carried out in this work can be extended to all memristive platforms and provides a comprehensive guideline for the optimization of analogue programming in the context of hardware implementation of neural networks.

Experimental Section

*Sample Preparation***:** 65nm SnO_x film is grown on a conductive fluorine-doped tin oxide (FTO) conductive glass by thermal atomic layer deposition (ALD) at 120ºC using Tetrakis Dimethylamino Tin (TDMASn) precursor and H_2O as oxidative reactant. The chamber pressure is maintained at 0.8Torr and the precursor exposure/purge time is optimized to be 100ms/15s for TDMASn and 50ms/15s for water respectively. Growth rate is determined to be approximately 0.1nm/cycle (measured post-deposition using a surface step profiler). No further annealing is done. The 100nm Al top electrodes are deposited by thermal evaporation through a shadow mask with 200μm to 500μm circular patterns.

Measurements: All memristive switching characteristics and pulse operations are carried out under ambient conditions using Keithley 4200 SCS Semiconductor characterization system. Retention characteristics are measured in vacuum ($<10^{-2}$ mbar) at room temperature. The AFM images are captured using Asylum Research Cypher S at AC air non-contact mode. The

XPS spectra are captured using Kratos AXIS Supra (monochromatized aluminium Kα x-ray source). No etching of sample is done in this experiment; a 55µm aperture is used for selective area photoelectrons capture. ESCApe software from Kratos is used for detailed peak fitting of the narrow scan regions. A Shirley background is subtracted prior to peak fitting and a Gaussian*Lorentzian function is used. The Full Wave Half Maximum (FWHM) is manually kept at acceptable ranges below 2eV (1.4 to 1.9eV), the peak finding and the error minimization are done by the software.

Retention characteristics: In order to get intermediate states, we first initialize by setting and resetting the device through a negative DC sweep. The device will be initialized to a conductance of approximately 0.5mS. After which, a train of 30 pulses (Pulse width of 6ms, 40Hz, see Figure 2c) will be applied, bringing the device to its corresponding conductance state. For example, to bring the device to 9mS, a train of +1.2V pulses will be applied. Once the pulse stimuli are removed, the resultant conductance state will be measured. The device is biased at a constant voltage of 0.1V, a reading is taken every 2s for the first 100s and every 100s for subsequent time period.

To obtain the dynamic weight changes for a potentiation series, we apply the procedure as shown:

1) In a series of 30 potentiation pulses, the conductance G_n is measured before and after every applied pulse

2) We label the normalized conductance as w_n (where n = 0 to 30 is the step number in the series)

3) The conductance values are first normalized to fall between 0 and 1 (with the initial normalized conductance $w_0 = 0$, and the final normalized conductance $w_{30} = 1$), therefore $w_n = (G_n – G_0)/ (G_{30} – G_0)$

4) Delta *w* for potentiation series ($\delta w_{+, n}$) is hence the change in normalized weight ($w_n - w_{n-1}$), where n is the step number in the potentiation series.

Similarly, for the depression series, the conductance is normalized with 1 being the initial normalized conductance and 0 being the final normalized conductance. Delta w ($\delta w_{-, n}$) is also the change in normalized weight ($w_{n-1} - w_n$).

Spike-timing dependent plasticity: As a standard procedure in all our long-term weight change calculations, final conductance states are measured 30s after pulsing. This is to avoid transient effects (at the timescale of <10 seconds) that could occur after the pulsing and to standardize the weight change calculation.

Supporting Information

Supporting Information is available from the ACS Publications website or from the author.

Memristor device fabrication, electrical behaviour of multiple conductance states, XPS fitting of HRS and LRS, STP measurements, STDP waveform.

Acknowledgements

The authors would like to acknowledge the funding from the Tier 2 Grant MOE2016-T2-1-100 and Grant MOE2018-T2-2-083. S.E.N., R.A.J., and N.M. conceived the project and experimental flow. S.E.N. fabricated the devices, performed the experiments, and analyzed the data, with help of R.A.J. and J.T.Y. The authors would like to thank N. Yantara and Y. B. Tay for the SEM images. S.E.N., R.A.J., and N.M. wrote the manuscript with input from all the authors.

References:

(1) Wright, C. D.; Hosseini, P.; Diosdado, J. A. V. Beyond Von-Neumann Computing with Nanoscale Phase-Change Memory Devices. *Adv. Funct. Mater.* **2013**, *23* (18), 2248–2254.

(2) John, R. A.; Ko, J.; Kulkarni, M. R.; Tiwari, N.; Chien, N. A.; Geok, N.; Leong, W. L.; Mathews, N. Flexible Ionic-Electronic Hybrid Oxide Synaptic TFTs with Programmable Dynamic Plasticity for Brain-Inspired Neuromorphic Computing. *Small* **2017**, *13* (32), 1701193.

(3) Burr, G. W.; Shelby, R. M.; Sebastian, A.; Kim, S.; Kim, S.; Sidler, S.; Virwani, K.; Ishii, M.; Narayanan, P.; Fumarola, A. Neuromorphic Computing Using Non-Volatile Memory. *Adv. Phys. X* **2017**, *2* (1), 89–124.

(4) Soudry, D.; Di Castro, D.; Gal, A.; Kolodny, A.; Kvatinsky, S. Memristor-Based Multilayer Neural Networks with Online Gradient Descent Training. *IEEE Trans. neural networks Learn. Syst.* **2015**, *26* (10), 2408–2421.

(5) Jo, S. H.; Chang, T.; Ebong, I.; Bhadviya, B. B.; Mazumder, P.; Lu, W. Nanoscale

Memristor Device as Synapse in Neuromorphic Systems. *Nano Lett.* **2010**, *10* (4), 1297–1301.

(6) Hutchby, J. A.; Bourianoff, G. I.; Zhirnov, V. V; Brewer, J. E. Extending the Road beyond CMOS. *IEEE Circuits Devices Mag.* **2002**, *18* (2), 28–41.

(7) Zhao, W.; Querlioz, D.; Klein, J.-O.; Chabi, D.; Chappert, C. Nanodevice-Based Novel Computing Paradigms and the Neuromorphic Approach. In *2012 IEEE International Symposium on Circuits and Systems*; IEEE, 2012; pp 2509–2512.

(8) Zhu, L. Q.; Wan, C. J.; Guo, L. Q.; Shi, Y.; Wan, Q. Artificial Synapse Network on Inorganic Proton Conductor for Neuromorphic Systems. *Nat. Commun.* **2014**, *5*, 3158.

(9) John, R. A.; Tiwari, N.; Yaoyi, C.; Ankit; Tiwari, N.; Kulkarni, M.; Nirmal, A.; Nguyen, A. C.; Basu, A.; Mathews, N. Ultralow Power Dual-Gated Subthreshold Oxide Neuristors: An Enabler for Higher Order Neuronal Temporal Correlations. *ACS Nano* **2018**, *12* (11), 11263–11273.

(10) John, R. A.; Liu, F.; Chien, N. A.; Kulkarni, M. R.; Zhu, C.; Fu, Q.; Basu, A.; Liu, Z.; Mathews, N. Synergistic Gating of Electro-Iono-Photoactive 2D Chalcogenide Neuristors: Coexistence of Hebbian and Homeostatic Synaptic Metaplasticity. *Adv. Mater.* **2018**, *30* (25), 1800220.

(11) Indiveri, G.; Linares-Barranco, B.; Legenstein, R.; Deligeorgis, G.; Prodromakis, T. Integration of Nanoscale Memristor Synapses in Neuromorphic Computing Architectures. *Nanotechnology* **2013**, *24* (38), 384010.

(12) Lee, T.-W.; Nickel, J. H. Memristor Resistance Modulation for Analog Applications. *IEEE Electron Device Lett.* **2012**, *33* (10), 1456–1458.

(13) Kim, K. M.; Zhang, J.; Graves, C.; Yang, J. J.; Choi, B. J.; Hwang, C. S.; Li, Z.; Williams, R. S. Low-Power, Self-Rectifying, and Forming-Free Memristor with an Asymmetric Programing Voltage for a High-Density Crossbar Application. *Nano Lett.* **2016**, *16* (11), 6724–6732.

(14) Cristiano, G.; Giordano, M.; Ambrogio, S.; Romero, L. P.; Cheng, C.; Narayanan, P.; Tsai, H.; Shelby, R. M.; Burr, G. W. Perspective on Training Fully Connected Networks with Resistive Memories: Device Requirements for Multiple Conductances of Varying Significance. *J. Appl. Phys.* **2018**, *124* (15), 151901.

(15) Abbas, Y.; Jeon, Y.-R.; Sokolov, A. S.; Kim, S.; Ku, B.; Choi, C. Compliance-Free, Digital SET and Analog RESET Synaptic Characteristics of Sub-Tantalum Oxide Based Neuromorphic Device. *Sci. Rep.* **2018**, *8* (1), 1228.

(16) Leydecker, T.; Herder, M.; Pavlica, E.; Bratina, G.; Hecht, S.; Orgiu, E.; Samorì, P.

Flexible Non-Volatile Optical Memory Thin-Film Transistor Device with over 256 Distinct Levels Based on an Organic Bicomponent Blend. *Nat. Nanotechnol.* **2016**, *11* (9), 769–775.

(17) Yang, J. J.; Pickett, M. D.; Li, X.; Ohlberg, D. A. A.; Stewart, D. R.; Williams, R. S. Memristive Switching Mechanism for Metal/Oxide/Metal Nanodevices. *Nat. Nanotechnol.* **2008**, *3* (7), 429.

(18) Chen, J.; Huang, C.; Chiu, C.; Huang, Y.; Wu, W. Switching Kinetic of VCM-based Memristor: Evolution and Positioning of Nanofilament. *Adv. Mater.* **2015**, *27* (34), 5028–5033.

(19) Wang, Z.; Wang, L.; Nagai, M.; Xie, L.; Yi, M.; Huang, W. Nanoionics-Enabled Memristive Devices: Strategies and Materials for Neuromorphic Applications. *Adv. Electron. Mater.* **2017**, 1600510.

(20) Cho, D.-Y.; Luebben, M.; Wiefels, S.; Lee, K.-S.; Valov, I. Interfacial Metal–Oxide Interactions in Resistive Switching Memories. *ACS Appl. Mater. Interfaces* **2017**, *9* (22), 19287–19295.

(21) Huang, R.; Yan, X.; Ye, S.; Kashtiban, R.; Beanland, R.; Morgan, K. A.; Charlton, M. D. B.; de Groot, C. H. K. Compliance-Free ZrO 2/ZrO 2− x/ZrO 2 Resistive Memory with Controllable Interfacial Multistate Switching Behaviour. *Nanoscale Res. Lett.* **2017**, *12* (1), 384.

(22) Chang, Y.-F.; Fowler, B.; Zhou, F.; Chen, Y.-C.; Lee, J. C. Study of Self-Compliance Behaviors and Internal Filament Characteristics in Intrinsic SiOx-Based Resistive Switching Memory. *Appl. Phys. Lett.* **2016**, *108* (3), 33504.

(23) Burr, G. W.; Narayanan, P.; Shelby, R. M.; Sidler, S.; Boybat, I.; di Nolfo, C.; Leblebici, Y. Large-Scale Neural Networks Implemented with Non-Volatile Memory as the Synaptic Weight Element: Comparative Performance Analysis (Accuracy, Speed, and Power). In *Electron Devices Meeting (IEDM), 2015 IEEE International*; IEEE, 2015; p 4.

(24) Burr, G. W.; Shelby, R. M.; Sidler, S.; Di Nolfo, C.; Jang, J.; Boybat, I.; Shenoy, R. S.; Narayanan, P.; Virwani, K.; Giacometti, E. U. Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element. *IEEE Trans. Electron Devices* **2015**, *62* (11), 3498–3507.

(25) Querlioz, D.; Bichler, O.; Dollfus, P.; Gamrat, C. Immunity to Device Variations in a Spiking Neural Network with Memristive Nanodevices. *IEEE Trans. Nanotechnol.*

2013, *12* (3), 288–295.

(26) Frascaroli, J.; Brivio, S.; Covi, E.; Spiga, S. Evidence of Soft Bound Behaviour in Analogue Memristive Devices for Neuromorphic Computing. *Sci. Rep.* **2018**, *8* (1), 7178.

(27) Fusi, S.; Abbott, L. F. Limits on the Memory Storage Capacity of Bounded Synapses. *Nat. Neurosci.* **2007**, *10* (4), 485.

(28) Moraitis, T.; Sebastian, A.; Eleftheriou, E. The Role of Short-Term Plasticity in Neuromorphic Learning: Learning from the Timing of Rate-Varying Events with Fatiguing Spike-Timing-Dependent Plasticity. *IEEE Nanotechnol. Mag.* **2018**, *12* (3), 45–53.

(29) Daoudal, G.; Debanne, D. Long-Term Plasticity of Intrinsic Excitability: Learning Rules and Mechanisms. *Learn. Mem.* **2003**, *10* (6), 456–465.

(30) Markram, H.; Gerstner, W.; Sjöström, P. J. Spike-Timing-Dependent Plasticity: A Comprehensive Overview. *Front. Synaptic Neurosci.* **2012**, *4*, 2.

(31) Edelmann, E.; Cepeda-Prado, E.; Leßmann, V. Coexistence of Multiple Types of Synaptic Plasticity in Individual Hippocampal CA1 Pyramidal Neurons. *Front. Synaptic Neurosci.* **2017**, *9*, 7.

(32) John, R. A.; Yantara, N.; Ng, Y. F.; Narasimman, G.; Mosconi, E.; Meggiolaro, D.; Kulkarni, M. R.; Gopalakrishnan, P. K.; Nguyen, C. A.; De Angelis, F. Ionotronic Halide Perovskite Drift-Diffusive Synapses for Low-Power Neuromorphic Computation. *Adv. Mater.* **2018**, 1805454.

(33) Kim, W.; Chattopadhyay, A.; Siemon, A.; Linn, E.; Waser, R.; Rana, V. Multistate Memristive Tantalum Oxide Devices for Ternary Arithmetic. *Sci. Rep.* **2016**, *6*, 36652.

(34) Chen, B.; Lu, Y.; Gao, B.; Fu, Y. H.; Zhang, F. F.; Huang, P.; Chen, Y. S.; Liu, L. F.; Liu, X. Y.; Kang, J. F. Physical Mechanisms of Endurance Degradation in TMO-RRAM. In *2011 International Electron Devices Meeting*; IEEE, 2011; pp 12–13.

(35) Swaidan, Z.; Kanj, R.; El Hajj, J.; Saad, E.; Kurdahi, F. RRAM Endurance and Retention: Challenges, Opportunities and Implications on Reliable Design. In *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*; IEEE, 2019; pp 402–405.