Diffusive memristor artificial neurons for fully memristive neural network

Zhongrui Wang[†,1], Saumil Joshi[†,1], Sergey Savel'ev[2], Wenhao Song[1], Rivu Midya[1], Yunning Li[1],

Mingyi Rao[1], Peng Yan[1], Shiva Asapu[1], Ye Zhuo[1], Hao Jiang[1], Peng Lin[1], Can Li[1], Jung Ho Yoon[1],

Navnidhi K. Upadhyay[1], Jiaming Zhang[3], Miao Hu[3], John Paul Strachan[3], Mark Barnell[4], Qing

Wu[4], Huaqiang Wu[5], R. Stanley Williams[*,3], Qiangfei Xia[*,1], J. Joshua Yang[*,1]

[1]Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003, USA

[2]Department of Physics, Loughborough University, Loughborough LE11 3TU, UK

[3]Hewlett-Packard Laboratories, Palo Alto, CA 94304, USA

[4]Air Force Research Lab, Information Directorate, Rome, New York 13441, USA

[5]Institue of Microelectronics, Tsinghua University, Beijing, China

[†] *these authors contributed equally*

[*]*e-mails: stan.williams@hpe.com; qxia@ecs.umass.edu; jjyang@umass.edu*

Hardware implementation of neuromorphic computing calls for the development of novel building blocks for energy efficient implementation of neural network algorithms. A volatile memristor based on the Ag diffusive dynamics in dielectric films has been developed to emulate the stochastic leaky integrate-and-fire mechanism of biological neurons, with integration time modulated by the interaction between intrinsic Ag dynamics and capacitor charging. Unsupervised programming of memristive synapses has been demonstrated with memristive neurons, which is used to train a fully connected network that complements convolution and rectified linear unit (ReLU) layers for pattern classification, built on integrated all-memristor neural networks for the first time. The simplicity, scalability, and stackability of such a network has the potential for high density, low cost, and low power neural computing systems.

The computing capability of artificial neural networks (ANNs) was publicly demonstrated by the recent performance of AlphaGo[1,2], which, however, consumed orders of magnitude more power compared to the human brain. Thus, traditional complementary metal–oxide–semiconductor devices and circuits are extremely inefficient in implementing brain-inspired computing paradigms. Devices that behave more directly like synapses and neurons should enable a significantly more efficient implementation of a neural network. Noteworthy progress has been made in building hardware ANNs that incorporate memristors and related devices[3-22] to simulate/emulate synapses by utilizing their tunable conductance as synaptic weights. In all these reported ANNs, the signal processing functions performed by simulated neurons were implemented either by CMOS circuits with about 10 transistors or more or in software running on processors.[12,23] More recently, artificial neurons based on memristors have been reported,[24-31] but there has not yet been a demonstration of a discrete scalable electronic device that performs the leaky integrate-and-fire signal processing and unsupervised learning with memristive synapses, let alone a functional integrated hardware demonstration at the network level comprising only emerging devices. Here we report a new artificial neuron with stochastic dynamics composed of a diffusive memristor based on Ag motion in a host dielectric[32-34]. The delay and decay temporal responses are determined by the interaction between the memristor internal state variables and the total RC time constant of the circuit elements. An integrated circuit was constructed by using drift memristor synapses and diffusive memristor neurons, which has been used to implement convolution, rectified linear unit (ReLU), and fully connected layers of a functional network to demonstrate pattern classification capability enabled by unsupervised synaptic weight updates in a fully memristive neural network for the first time.

2

Important signal processing tasks of a neuron are to integrate inputs received through synapses and (1) to output a signal if a threshold has been reached within a defined time interval or (2) to allow the integrated input signal to decay (i.e. forget) if the interval is exceeded[35-37]. The leaky integrate-and-fire model[38] is often used to describe this behavior in biological neurons and is emulated by volatile memristors, which transition to a high conductance state when their stimulation threshold is exceeded. The 'leaky' membrane potential of the neuron corresponds to the volatile conductance of the memristor, which is a critical dynamical property for forgetting. This enables the neuron to automatically reinstate its resting membrane potential not only after it successfully fires an output pulse, but also if it fails to do so because of insufficient stimulation, thus resetting the original threshold. The decay time determines the memory span of the neuron, which enables short term memory in ANNs[39-41]. In addition to its temporal significance, the signal decay is also crucial in spatial integrations, as it weighs signals from different locations (even simultaneous events) in the network via their transit time along the dendrites[35].

We physically emulated the leaky integrate-and-fire neuron model with a diffusive memristor, fabricated by sandwiching a dielectric material (e.g. SiO_xN_y or SiO_x) carefully doped with Ag nanoclusters between two electrodes. This discrete device, schematically illustrated in Figure 1a, was characterized by applying voltage pulses across the artificial neuron in series with resistors to represent synapses and recording the resulting output current versus time. Figures 1b-e compare experimentally measured data with corresponding physics-based simulation results. (See Methods.) The temporal behavior of the artificial neuron was observed during and after the input of a single super-threshold voltage pulse followed by a train of smaller pulses. (See Supplementary Figure S1.) There was a distinct delay time (τ_d) between the arrival of the voltage pulse and the

rise of the output current, which was caused by the interaction of the RC time constant of the circuit with the internal Ag dynamics of the memristor. With a relatively large circuit capacitance, the RC time constant, that is the time for establishing the switching voltage of the diffusive memristor, dominates the delay time. (See Supplementary Figure S2.) With a smaller capacitance, the RC time becomes shorter and the internal Ag dynamics of the memristor dominates the delay time and thus the integrate-and-fire behavior, as shown in Figure 1. The internal Ag dynamics of diffusive memristors originates from complicated multi-physics effect including field induced Ag mass transport from the electrodes (e.g. Ag diffusion and redox reaction, see Supplementary Note 1.) and the formation of a electrical conducting path.[42-46] We have constructed a physics-based model that agrees well with the microscopic observation of Ag filament growth and rupture during threshold switching as well as the measured temporal response to voltage signals (e.g. Figures 1 and S1).[34,47,48] Although the incorporated mechanisms of the model do not include all of the possible physics, at this stage it provides a good approximation of the rate limiting dynamics of the diffusive memristor, and is thus sufficient for understanding the interplay between the internal Ag dynamics of the memristor and the circuit capacitance.

After the fall of the voltage pulse, the memristor conductance relaxed with a characteristic time (τ_r) determined within our model by the Ag diffusive dynamics to dissolve the nanoparticle bridge and return the neuron to its resting state (See Supplementary Figure S3 for delay and relaxation properties.) The relaxation dynamics also leads to the leakiness of the internal Ag dynamics, which gradually dissolves Ag conducting channel(s) driven by the minimization of interfacial energy between Ag and dielectrics, or Thomson-Gibbs effect.[34,47] When a sequence of sub-threshold pulses was applied to the device, as shown in Figures 1b and 1d, the device fired after some number

of pulses and relaxed back to the resting state after the end of the pulse train. Shown in Figures 1c and 1e are the corresponding experimentally measured and simulated histograms of the firing statistics, respectively, which show that the threshold is not sharp but has an associated probability distribution function, providing the stochastic behavior commonly observed in actual neurons. Since the internal memristor dynamics depend on the behavior of nanoparticles, the leaky integrate and fire mechanism observed here should scale to very small device sizes.

Thus, the function of the diffusive memristor in the artificial neuron is very different from nonvolatile drift memristors or phase change memory devices used as short- or long-term resistive memory elements or synapses.[27,49,50] The diffusive memristor integrates the presynaptic signals within a time window and transitions to a low resistance state only if a threshold has been reached.

Depending on system configurations, either the circuit RC timing or the memristor dynamics might dominate the artificial neuron temporal behavior. For a clear demonstration of a dominant RC effect, we used a relatively large external capacitor (>1nF) in parallel with the diffusive memristor. (See Supplementary Figure S4.) The leaky integrate-and-fire response of the artificial neuron can be tuned by adjusting the circuit and the physical design around the device, as shown in Figure 2a. The threshold behavior of the diffusive memristor can be compared to that of an ion-channel located near the soma of a neuron, whereas the membrane capacitance and axial resistance are represented by a capacitor C_m parallel to the memristor and a resistor R_a in series with this combination[24,51]. In a neuron, all inputs from the surrounding neurons are fed through synapses and integrated near the soma, the membrane capacitance charges up, activating the ion channels if the charge reaches the threshold, and the neuron fires. When input pulses are applied to the element shown in Figure 2a, the circuit capacitance charges with a time constant (R_aC_m), increasing the

5

voltage across the diffusive memristor. If the threshold is reached, a Ag conduction channel is formed between the electrodes, which switches the memristor and discharges (fires) the capacitor. We present data that shows the capacitor charging and the subsequent firing of a current pulse by the memristor in Figure 2b. A smaller capacitance makes the integration process and spiking faster, while a larger axial input resistance slows down the charge build-up, delaying or preventing the firing, as summarized in both Figures 2b and 2c. The current spike across the diffusive memristor coincides with the discharging of the capacitor, indicating the active release of the charge stored in the capacitor. Just as the physical characteristics and environment of a biological neuron affects it's properties[52], the structure of the hybrid device and its surrounding circuit design controls it's response to input stimuli. This allows us to tailor the properties of the artificial neuron to achieve desirable response characteristics for specific applications. (See Supplementary Table S1 for the factors affecting firing properties.)

Next, we experimentally demonstrate the interactions between the artificial neurons and synapses, which serves as the basis for the learning of all biological neural systems. (See Supplementary Figure S5 and S6 for the input waveform design with RC timing effect.) A drift memristor synapse with a small weight (low conductance) is in series with the artificial neuron, consisting of a diffusive memristor in parallel with a capacitor to simulate a large circuit capacitance in this case. (See Supplementary Figure S7a.) The synapse has a low efficiency, i.e. the voltage drop across it is large, which results in a slow build-up of charge across the circuit capacitance during the rising edge of the applied pulse. The artificial neuron integrates the input but does not fire because it cannot reach the required threshold within the duration of this pulse. On the other hand, a synapse with a larger weight (or a larger conductance of the drift memristor synapse) results in a faster

buildup of charge across the capacitance and a successful firing event, as shown in Supplementary Figure S7b. For the case with a negligible circuit capacitance, a synapse with a small weight produces a small voltage division across the artificial neuron which consists of a diffusive memristor in parallel with a resistor. (This parallel resistor may or may not be needed depending on the relative resistance of synapses and neurons.) (See Supplementary Figure S7c) However, a large weight of the synapse leads to the observed firing of the artificial neuron as the voltage drop across the diffusive-memristor/resistor(parallel) becomes larger and exceeds the threshold. (See Supplementary Figure S7d)

In order to experimentally illustrate unsupervised synaptic weight update caused by neuron firing, we used a 2×2 drift memristor synapse array connected to diffusive memristor artificial neurons at each output as shown in Figure 3. All the synapses were initialized to small weights, with some variation due to the stochastic nature of their switching, as shown in the Figure 3a. We applied a triangular voltage pulse (1st column in Figure 3) or a train of rectangular spikes (3rd column in Figure 3) to the first row of synapses to emulate large and small circuit capacitance, respectively. The second row is kept at nearly zero bias. The '10' digital input vector pattern is used in this demonstration, but analog inputs could be used in principle. As shown in the 1st and 3rd column of Figure 3a and b, the neuron N_2 connected to the right column fires because the synapse S_{12} had a slightly larger initial weight. The firing of the neuron pulls down the voltage of the bottom electrodes of S_{12} and S_{22}, resulting in a large voltage spike (green lines of the middle panels in Figure 3b) across S_{12}, further enhancing its weight. Next, we verified the network response for an input vector '11'. When either triangular voltage pulses (2nd column in Figure 3a) or trains of rectangular spikes (4th column in Figure 3a), corresponding to large and small circuit capacitance,

respectively, are applied to both rows of the 2×2 network, both neurons fire, enhancing the weights of synapses S_{12} and S_{21}. (See 2nd and 4th column in Figure 3b)

We then went a step further to demonstrate inference on a prototype chip of fully integrated memristive neural network. Figure 4a shows the overview of the integrated chip consisting of 1-Transistor-1-Memristor (1T1R) synaptic array and diffusive memristor neurons. The synapses were built by integrating drift memristors with foundry-made transistor arrays using back-end-of-the-line (BEOL) processes. (See Methods.) Each $Pd/HfO_2/Ta$ memristors is connected to a series n-type enhancement-mode transistor. Figure 4b shows the detailed structure of a single 1T1R cell and associated connections. When all the transistors are turned on, the 1T1R array works as a fully connected memristor crossbar. Structural analysis using high-resolution transmission electron microscopy was performed on the integrated memristors, which reveals an amorphous HfO_2 layer sandwiched between Pd and Ta electrodes in Figure 4c. Figure 4d illustrates the junction of a single diffusive memristor. A transmission electron micrograph of its cross-section shows the amorphous nature of the background SiO_x dielectric lattices and the nano-crystalline Ag layer in Figure 4e.

Pre-synaptic signals could be classified by such a fully memristive neural network. Here for demonstration purpose, the synapses were pre-programmed to have different weights, which could be the result of any kind of learning process. Four letter patterns "U", "M", "A", and "S" with artificially added noise were used as example inputs. The red and blue squares in Figure 4f represent the input differential voltages fed to the rows of the synaptic array. For example, a red square means a +0.8V/-0.8V input pair and a light blue square means a -0.6V/+0.6V input pair. The input pattern is divided into 4 sub-images of a 2×2 size, with a stride of two. Each sub-image is unrolled into 1 column input vector (8 voltages) and fed into the network (8 rows) at each time. For each possible sub-image

8

there is a corresponding convolutional filter implemented by 8 memristor synapses in a column, with a total of 8 filters (8 columns) in the 8×8 array. The measured weights were depicted in Figure 4g after programming. The negative values of the convolution matrices are mapped to the conductance of memristor cells by grouping memristors from adjacent rows to form a differential pair. The result of the convolution of the 8 filters to each sub-image are concurrently revealed by the firing of their corresponding diffusive memristor artificial neurons which serve the role of the ReLUs. This network can produce unique response for each input pattern, as illustrated in the Figure 4h and i, in the form of integration time and the maximum fire current. Supplementary Figure S8b depicts the temporal current responses of the neurons upon the noisy 'UMAS' inputs. We have also verified the repeatability of the network by feeding the 8 noise-free patterns in cycles to the network and record the average firing delay and current of neurons (See Supplementary Figure S9b). Compare Figure 4h with Figure S9b, the integration time of a noisy input is generally longer due to smaller inputs and thus smaller convolution results. Correspondingly, inputs with positive additive noise will usually fire faster. This proof of principle demonstration of the fully integrated memristive neural network comprising memristor-based artificial synapses and artificial neurons can be expanded to implement learning systems of larger complexity in an energy efficient manner, such as multilayer neuron networks.[53,54] (See Supplementary Note 2 for power consumption analysis.)

STDP is a prevalent protocol of synaptic weight update in spiking neural networks.[4,55] Here we derive a simple STDP scheme based on the observations in Figure 3 to train a fully connected layer in an unsupervised approach, which naturally complements the convolution and ReLU layers in Figure 4 and further enables a functional convolutional network. Since the drift memristor synapses encode the conditional probability[18], the neurons will tend to respond to the means of

9

inputs associated with fire events, essentially carrying out clustering of the inputs. This is experimentally demonstrated in Figure 5. Software pooling and signal conversion is used to fit the output of ReLU layer to the input of fully connected layer. (See Figure 5a and Methods.) Lateral inhibition is deployed, which is typical in fully connected feedforward networks to enhance the discrimination of the inputs and make the self-adapting network energy efficient[18,56-59]. (See Methods.) After a few cycles of uncertainty where the conductance of synapses concentrates around the initial values (~100μS), the synapses are clearly programmed by the simple STDP rules. As shown in Figure 5d, undergoing either potentiation or depression, patterns of synapses associated with the N1, N2, and N3 neurons quickly gain similarities by self-organizing processes to one of the prototypical patterns in Figure 5a, respectively. (i.e. '11110000', '11000011', '00001100') It is also noted that synapses may show different response to the learning rules. For instance, the 3rd synapse of N1 and the 7th synapse of N2 are much less potentiated, which may be due to the device-to-device variation of threshold conditions of drift memristors. The quick divergence of conductance of drift memristors indicates a fast learning rate which is dependent on the firing time or pulse width of diffusive memristor neurons. Such convergence is also reflected by the magnitude (or threshold) of input patterns in Figure 5b. The magnitude of a specific pattern reduces in the first few cycles and then becomes stable. This is because diverged conductance of drift memristors tend to saturate so that further increment (decrement) in conductance will become less effective when they are close to the upper (lower) bound of the conductance range.

Conclusion

We have demonstrated a stochastic leaky integrate-and-fire artificial neuron based on a discrete scalable diffusive memristor, featuring Ag dynamics similar to that of actual neuron ion channels,

representing to date the simplest and yet faithful realization of electronic neural functionality, in contrast to traditional approaches requiring tens to hundreds of CMOS devices. Physics-based simulations reproduce our experimental observations and thus enhance our understanding of the interplay between memristor dynamics and circuit RC effects. Utilizing the integrate-and-fire function, the artificial neurons have performed unsupervised synaptic weight updating and pattern classification for the first time on integrated convolution neural networks comprising only memristors.

Reference

1   Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484-489, (2016).

2   Chouard, T. The Go Files: AI computer clinches victory against Go champion. *Nature*, (2016).

3   Strukov, D. B., Snider, G. S., Stewart, D. R. & Williams, R. S. The missing memristor found. *Nature* **453**, 80-83, (2008).

4   Jo, S. H. *et al.* Nanoscale Memristor Device as Synapse in Neuromorphic Systems. *Nano Lett.* **10**, 1297-1301, (2010).

5   Yu, S., Wu, Y., Jeyasingh, R., Kuzum, D. & Wong, H. S. P. An Electronic Synapse Device Based on Metal Oxide Resistive Switching Memory for Neuromorphic Computation. *IEEE Trans. Elect. Dev.* **58**, 2729-2737, (2011).

6   Ohno, T. *et al.* Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. *Nat. Mater.* **10**, 591-595, (2011).

7   Suri, M. *et al.* in *Electron Devices Meeting (IEDM), 2011 IEEE International.* 4.4.1-4.4.4 (IEEE).

8   Indiveri, G. *et al.* Neuromorphic silicon neuron circuits. *Front. Neurosci.* **5**, 73, (2011).

9   Pershin, Y. V. & Di Ventra, M. Neuromorphic, Digital, and Quantum Computation With Memory Circuit Elements. *Proc. IEEE* **100**, 2071-2080, (2012).

10   Lim, H., Kim, I., Kim, J. S., Hwang, C. S. & Jeong, D. S. Short-term memory of TiO_2-based electrochemical capacitors: empirical analysis with adoption of a sliding threshold. *Nanotechnology* **24**, 384005, (2013).

11   Sheridan, P., Ma, W. & Lu, W. in *Circuits and Systems (ISCAS), 2014 IEEE International Symposium on.* 1078-1081 (IEEE).

12   Merolla, P. A. *et al.* A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668-673, (2014).

13   Eryilmaz, S. B. *et al.* Brain-like associative learning using a nanoscale non-volatile phase change synaptic device array. *Front. Neurosci.* **8**, 205, (2014).

14   La Barbera, S., Vuillaume, D. & Alibart, F. Filamentary Switching: Synaptic Plasticity through Device Volatility. *ACS Nano* **9**, 941-949, (2015).

15   Prezioso, M. *et al.* Training and operation of an integrated neuromorphic network based on metal-oxide memristors. *Nature* **521**, 61-64, (2015).

16   Burr, G. W. *et al.* Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element. *IEEE Trans. Elect. Dev.* **62**, 3498-3507, (2015).

17   Hu, S. G. *et al.* Associative memory realized by a reconfigurable memristive Hopfield neural network. *Nat. Commun.* **6**, 7522, (2015).

18   Serb, A. *et al.* Unsupervised learning in probabilistic neural networks with multi-state metal-oxide memristive synapses. *Nat. Commun.* **7**, 12611, (2016).

19   Park, J. *et al.* TiO$_x$-based RRAM synapse with 64-levels of conductance and symmetric conductance change by adopting a hybrid pulse scheme for neuromorphic computing. *IEEE Elect. Dev. Lett.* **37**, 1559-1562, (2016).

20   Ambrogio, S. *et al.* Unsupervised Learning by Spike Timing Dependent Plasticity in Phase Change Memory (PCM) Synapses. *Front. Neurosci.* **10**, 56, (2016).

21   van de Burgt, Y. *et al.* A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing. *Nat. Mater.* **16**, 414-418, (2017).

22   Shulaker, M. M. *et al.* Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. *Nature* **547**, 74-78, (2017).

23   Sourikopoulos, I. *et al.* A 4-fJ/Spike Artificial Neuron in 65 nm CMOS Technology. *Front. Neurosci.* **11**, 123, (2017).

24   Pickett, M. D., Medeiros-Ribeiro, G. & Williams, R. S. A scalable neuristor built with Mott memristors. *Nat. Mater.* **12**, 114-117, (2013).

25   Al-Shedivat, M., Naous, R., Cauwenberghs, G. & Salama, K. N. Memristors empower spiking neurons with stochasticity. *IEEE Trans. Emerg. Sel. Topics Circuits Syst.* **5**, 242-253, (2015).

26   Lim, H. *et al.* Reliability of neuronal information conveyed by unreliable neuristor-based leaky integrate-and-fire neurons: a model study. *Sci. Rep.* **5**, 9776, (2015).

27   Tuma, T., Pantazi, A., Le Gallo, M., Sebastian, A. & Eleftheriou, E. Stochastic phase-change neurons. *Nat. Nanotechnol.* **11**, 693-699, (2016).

28   Lim, H. *et al.* Relaxation oscillator-realized artificial electronic neurons, their responses, and noise. *Nanoscale* **8**, 9629-9640, (2016).

29   Mehonic, A. & Kenyon, A. J. Emulating the Electrical Activity of the Neuron Using a Silicon Oxide RRAM Cell. *Front. Neurosci.* **10**, 57, (2016).

30   Gupta, I. *et al.* Real-time encoding and compression of neuronal spikes by metal-oxide memristors. *Nat. Commun.* **7**, 12805, (2016).

31   Stoliar, P. *et al.* A Leaky-Integrate-and-Fire Neuron Analog Realized with a Mott Insulator. *Adv. Funct. Mater.*, 1604740, (2017).

32   Yang, Y. *et al.* Observation of conducting filament growth in nanoscale resistive memories. *Nat. Commun.* **3**, 732, (2012).

33   Liu, Q. *et al.* Real-time observation on dynamic growth/dissolution of conductive filaments in oxide-electrolyte-based ReRAM. *Adv. Mater.* **24**, 1844-1849, (2012).

34   Wang, Z. *et al.* Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nat. Mater.* **16**, 101-108, (2016).

35   Magee, J. C. Dendritic integration of excitatory synaptic input. *Nat. Rev. Neurosci.* **1**, 181-190, (2000).

36   Stuart, G., Spruston, N. & Häusser, M. *Dendrites*. Third edition. edn, (Oxford University Press, 2016).

37   Tsien, J. Z. The memory. *Sci. Am.* **297**, 52-59, (2007).

38   Gerstner, W. & Kistler, W. M. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. (Cambridge University Press, 2002).

39   Johnson, N. & Hogg, D. Learning the distribution of object trajectories for event recognition. *Image and Vision Computing* **14**, 609-615, (1996).

40   Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735-1780, (1997).

41   Gers, F. A., Schmidhuber, J. & Cummins, F. in *9th International Conference on Artificial Neural Networks: ICANN '99*. 850-855 (Institution of Engineering and Technology).

42   Tsuruoka, T. *et al.* Effects of Moisture on the Switching Characteristics of Oxide-Based, Gapless-Type Atomic Switches. *Adv. Funct. Mater.* **22**, 70-77, (2012).

43   Valov, I. *et al.* Atomically controlled electrochemical nucleation at superionic solid electrolyte surfaces. *Nat. Mater.* **11**, 530-535, (2012).

44   Valov, I. *et al.* Nanobatteries in redox-based resistive switches require extension of memristor theory. *Nat. Commun.* **4**, 1771, (2013).

45   Messerschmitt, F., Kubicek, M. & Rupp, J. L. M. How Does Moisture Affect the Physical Property of Memristance for Anionic-Electronic Resistive Switching Memories? *Adv. Funct. Mater.* **25**, 5117-5125, (2015).

46   Valov, I. & Lu, W. D. Nanoscale electrochemistry using dielectric thin films as solid electrolytes. *Nanoscale* **8**, 13828-13837, (2016).

47   Midya, R. *et al.* Anatomy of Ag/Hafnia-Based Selectors with 10^{10} Nonlinearity. *Adv. Mater.* **29**, 1604457-n/a, (2017).

48   Jiang, H. *et al.* A novel true random number generator based on a stochastic diffusive memristor. *Nat. Commun.* **8**, (2017).

49   Wong, H.-S. P. *et al.* Phase change memory. *Proc. IEEE* **98**, 2201-2227, (2010).

50   Jeyasingh, R., Liang, J., Caldwell, M. A., Kuzum, D. & Wong, H.-S. P. in *Custom Integrated Circuits Conference (CICC), 2012 IEEE.* 1-7 (IEEE).

51   Chua, L., Sbitnev, V. & Kim, H. Hodgkin-Huxley Axon is Made of Memristors. *Int. J. Bifurcat. Chaos* **22**, 1230011, (2012).

52   Mainen, Z. F. & Sejnowski, T. J. Influence of dendritic structure on firing pattern in model neocortical neurons. *Nature* **382**, 363-366, (1996).

53   Hinton, G. E. & Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **313**, 504-507, (2006).

54   Roweis, S. T. & Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290**, 2323-2326, (2000).

55   Kim, S. *et al.* Experimental Demonstration of a Second-Order Memristor and Its Ability to Biorealistically Implement Synaptic Plasticity. *Nano Lett.* **15**, 2203-2211, (2015).

56   Yu, S. *et al.* A low energy oxide-based electronic synaptic device for neuromorphic visual systems with tolerance to device variation. *Adv. Mater.* **25**, 1774-1779, (2013).

57 Tuma, T., Le Gallo, M., Sebastian, A. & Eleftheriou, E. Detecting Correlations Using Phase-Change Neurons and Synapses. *IEEE Elect. Dev. Lett.* **37**, 1238-1241, (2016).

58 Pantazi, A., Wozniak, S., Tuma, T. & Eleftheriou, E. All-memristive neuromorphic computing with level-tuned neurons. *Nanotechnology* **27**, 355205, (2016).

59 Sebastian, A. *et al.* Temporal correlation detection using computational phase-change memory. *Nat. Commun.* **8**, 1115, (2017).

60 Li, C. *et al.* Analogue signal and image processing with large memristor crossbars. *Nature Electronics*, (2017).

MATERIALS AND METHODS

Fabrication of discrete diffusive memristor and drift memristor

The diffusive memristor devices were fabricated on p-type (100) Si wafer with 100 nm thermal oxide. The bottom electrodes were patterned by photolithography followed by evaporation and liftoff of ~20/2nm Pt/Ti. The ~16 nm thick doped dielectric layer was deposited at room temperature by reactively co-sputtering Si and Ag in Ar, N_2, and O_2. The ~30 nm Pt top electrodes were subsequently patterned by photolithography followed by evaporation and liftoff processes. Electrical contact pads of the bottom electrodes were first patterned by photolithography and then subjected to reactive ion etching with mixed CHF_3 and O_2 gases.

The drift memristors share same substrate and bottom electrodes with diffusive memristors. The HfO_2 switching layer was deposited by atomic layer deposition at 250 °C, which was subsequently patterned for reactive ion etching. Finally, top electrodes of 50/10nm Ta/Pd were sputtered and lifted-off.

Fabrication of the fully integrated memristive neural network

The synapses used in the demonstration are 1T1R array with $Pd/HfO_2/Ta$ memristors. The front-end and part of the back-end process for the transistors array was fabricated in a commercial fab. To make a good connection between the fab metal layers and the memristors, argon plasma treatment was done to remove the native metal oxide layers followed by the deposition of 5nm Ag and 200nm Pd by sputtering and lift-off process, and annealing at 300 °C for 0.5 hour. A 5nm Ta adhesive layer and 60nm Pd bottom electrode were then deposited by sputtering and patterned by lift-off. The HfO_2 switching layer was deposited by atomic layer deposition at 250 °C. The patterning of the switching layer was done by photolithography and reactive ion etching. Top

electrodes of 50nm Ta were sputtered and lifted-off. The bottom electrodes of diffusive memristors were patterned by photolithography followed by evaporation and liftoff of ~2/20/2nm Ag/Pt/Ti. To enhance the contact between the diffusive memristor electrodes and the column wires of the drift memristors, 100nm Pd patches were patterned, sputtered, and lifted-off. The ~10nm thick doped dielectric layer was patterned and deposited at room temperature by co-sputtering SiO_2 and Ag in Ar, followed by lift-off. The ~2/30 nm Ag/Pt top electrodes were subsequently patterned by photolithography followed by evaporation and liftoff processes.

Electrical measurements

We used a Keysight B1530 to perform the electrical measurements for the results shown in Figure 1. Using one channel of the Keysight B1530, we applied voltage pulses across the diffusive memristor in series with a resistor and measured the current using the other channel.

Electrical measurements of Figure 2 and 3 were performed using the Keysight 33622A arbitrary waveform generator, the Keysight MSOX3104 mixed signal oscilloscope, and the Keysight B1530 WGFMU. Voltage pulses were applied by the Keysight 33622A. The analog oscilloscope channels were used to measure the voltages at the output of the function generator, and across the diffusive memristor. The current across the diffusive memristor was monitored using the Keysight B1530. We used electrolytic capacitors and general-purpose resistors. For the ON duration study, we used a 50kΩ resistor and a 5nF capacitor with a 100μs pulse interval; the voltage study used 100μs ON and 50μs duration with a 10nF capacitor and a 47gΩ resistor; for varying interval, a 50kΩ resistor and a 5nF capacitor with a 100μs ON duration.

An in-house customized measurement system is developed to operate the fully memristive neural network.[60] As shown in Supplementary Figure S10a, the system work in two different modes,

switched by the customized multiplexer (MUX) array. In the first mode, the row electrodes of the synaptic array (Pd electrodes of the Pd/HfO$_x$/Ta memristors) are connected to waveform generators which output triangular waveforms. The current through the diffusive memristor neurons are sampled by the transimpedance amplifiers (TIAs) and microcontroller unit 2 (MCU 2). In the second mode, the rows (columns) of the drift memristor array are connected to the customized row (column) printed circuits boards, leaving diffusive memristors float.

The training scheme is detailed in Supplementary Figure S10b. The experiment runs 30 8-pixel patterns presented to the network. Each pattern is derived from the neuron outputs in Figure 4i. (See Supplementary Figure S11.) Basically, input voltages are proportional to the maximum currents of neurons in scanning one of four letters ("U", "M", "A", and "S", see Figure 4f) by software pooling. The 8-pixel outputs are generated via the 4 channels of function generators by averaging each pair. (The ideal output patterns are '11110000', '11110000', '00001100', '11000011', which allows representation in space with reduced dimensions.) The current to voltage conversion is done by software with added artificial noise.

The lateral inhibition is realized with the training scheme and hardware assistance. The input pattern is scaled so that its maxima is 0.5V at the beginning of each training cycle. The voltage of the input pattern gradually increases until a neuron fires. In principle, a sufficient slow ramping rate could limit the number of concurrently fired neurons. In addition, we also program the MCU 2 to float the columns of loser neurons once a fire event is identified in each cycle to assure that only the winner neuron could successfully trigger plasticity at its synapses. The depression of drift memristor synapses is done after each fire event by applying RESET pulses via the customized row boards to all drift memristors receiving low inputs of the winner neuron.

19

Modeling the integrate-and-fire behavior of the diffusive memristor/capacitor

Diffusive-memristor/capacitor hybrid dynamical simulations. For modeling the dynamics of a diffusive memristor, we consider an interplay of electric, heat and Ag-nanoparticle degrees of freedom.[34,47,48] Ag-nanoparticle diffusion is described by the Langevin equation:

$$\eta \frac{dx_i}{dt} = -\frac{\partial U(x_i)}{\partial x_i} + \alpha \frac{V(t)}{L} + \sqrt{2\eta k_B T} \zeta_i. \quad (1)$$

Here we introduce location, x_i, of ith Ag-nanoparticle, which drifts with time t in the potential landscape $U(x_i)$ under the action of the friction force $\eta \frac{dx_i}{dt}$ with particle viscosity η, the electric force $\alpha \frac{V(t)}{L}$ with induced charge α and distance L between electrodes, and the random force described by the unbiased δ-correlated white noise ζ_i, $\langle \zeta_i(t) \rangle = 0, \langle \zeta_i(0)\zeta_j(t) \rangle = \delta_{i,j}\delta(t)$ (Here $\delta(t)$ is the Dirac delta function, and $\delta_{i,j}$ is Kronecker delta.).. The particular shape of the potential does not qualitatively change the result and should take into account the interaction attracting Ag-nanoparticles to the large clusters as well as pinning of Ag-nanoparticles to the inhomogeneities of the insulating matrix resulting in a large number of smaller potential wells. The relative strength of the potential with respect to the thermal fluctuation energy $k_B T$ (with the Boltzmann constant k_B and the local Ag-nanoparticle temperature T, which can significantly differ from the device ambient temperature) determines the diffusion kinetics. Due to Joule dissipation, the temperature T changes in time according to the Newton cooling law:

$$\frac{dT}{dt} = \frac{V^2}{\mathbb{C}_T R} - \kappa(T - T_0), \quad (2)$$

where κ is the heat transfer coefficient describing heat flux from the device and $\mathbb{C}_T$ is the system heat capacitance. The input power is determined by the memristor resistance $R(x_1, x_2, .., x_N)$ and voltage $V(t)$ across the device. The resistance is controlled by the sequential tunneling processes of electrons from one Ag-nanoparticle to another and can be written as $= R_t \sum_0^N e^{(x_{i+1}-x_i)/\lambda}$, where $x_0 = -L$ and $x_{N+1} = L$ are the positions of the device terminals, R_t is the resistance amplitude and λ is the tunneling length. As a unit of resistance in our simulations we used its minimum value $R_{min} = (N+1)R_t e^{2L/((N+1)\lambda)}$ (occurring when all Ag-nanoparticles are equally separated), while the voltage is normalized to the switching threshold value determined self-consistently as the value when the probability of switching is close to one (see Supplementary Figure S1b).

As for any distributed system with a high resistance, the diffusive memristor has an intrinsic capacitance C_M. Considering the circuit shown in the inset of Figure 1e, we derive the equation for the voltage across the memristor driven by the applied voltage $V_{ex}(t)$:

$$\tau_0 \frac{dV}{dt} = V_{ex}(t) - \left(1 + \frac{R_{ex}}{R(x)}\right) V \quad (3)$$

Where the "RC" time is defined as $\tau_0 = C_M R_{ex}$ with the resistance R_{ex} of the external or signal input wires connected in series with the memristor (for simulations we used $\kappa\tau_0 = 0.2$ and $\frac{R_{ex}}{R_{min}} = 1$).

Acknowledgements

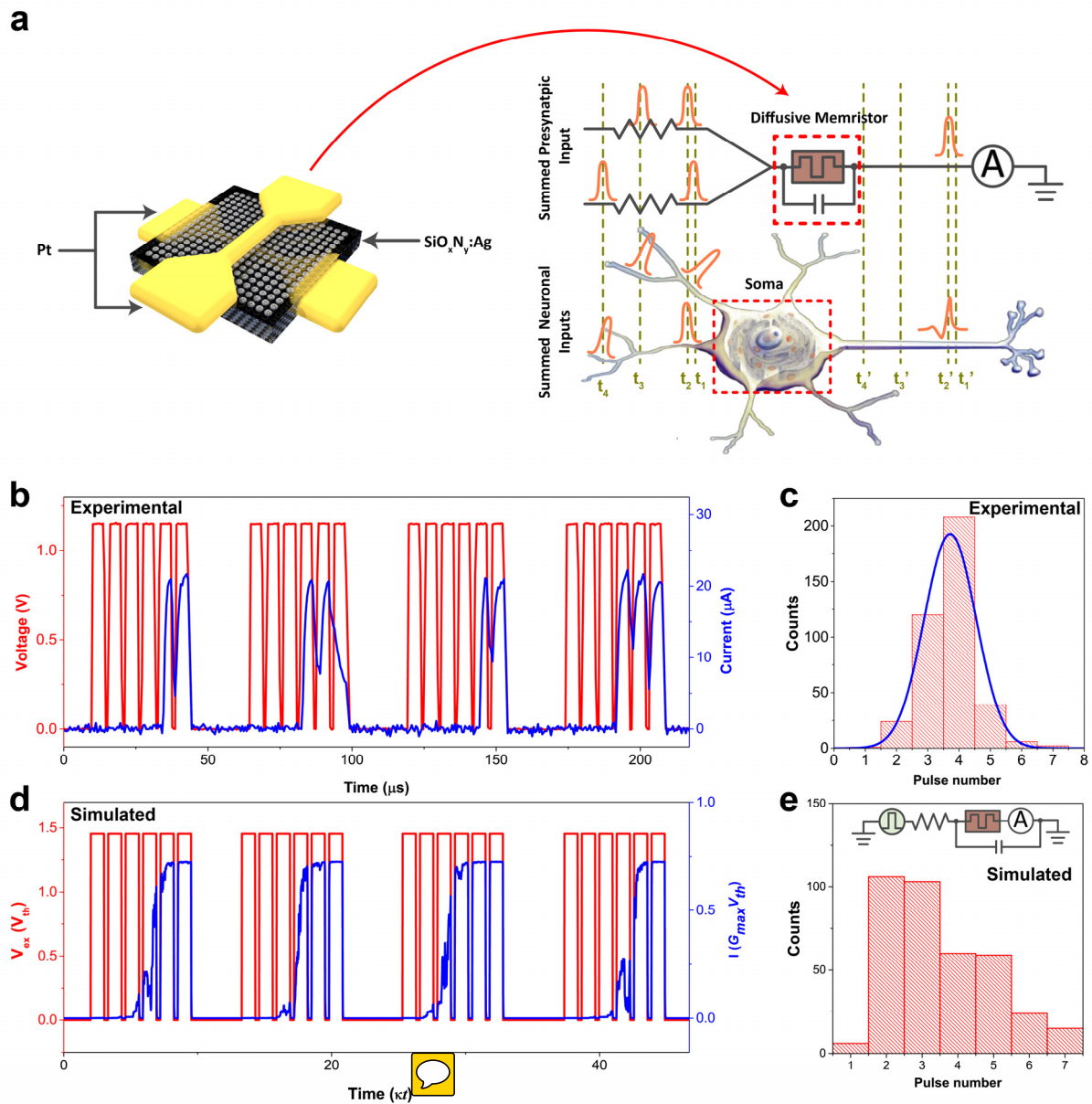**Figure 1** Diffusive memristor artificial neuron. **a**, Schematic illustration of a crosspoint diffusive memristor, which consists of a SiO_xN_y:Ag layer between two Pt electrodes. The ==artificial== neuron receives software summed weighted presynaptic inputs via a pulsed voltage source and an equivalent synaptic resistor (e.g. 20μS in this case). (See Supplementary Note 3 for the principle of software spatial summation.) Both the ==artificial== and biological neurons integrate input stimuli

23

(orange) beginning at t_1 and fire when the threshold condition is reached (i.e. at t_2'). The integrated signal decays over time such that input stimuli spaced too far apart will fail to reach threshold (i.e. the delay between t_3 and t_4). **b**, Experimental response of the device to multiple subthreshold voltage pulses followed by a rest period of 200 μs (only 20 μs is shown for convenience). The device required multiple pulses to reach the threshold and 'fire'. **c**, Histogram of the number of subthreshold voltage pulses required to successfully fire the artificial neuron (red) compared to a Gaussian distribution (blue). **d**, Simulated response of the device to multiple subthreshold voltage pulses as in **b** and showing similar behaviour to experiment, with the resting time between pulse trains chosen to allow the Ag in the device to diffuse back to the OFF state. (Only 10% of the rest period is shown for convenience) **e**, Simulated switching statistics with respect to pulse numbers (within each train), consistent with the experimental results in **c**. The inset illustrates the circuit diagram used in the simulation.
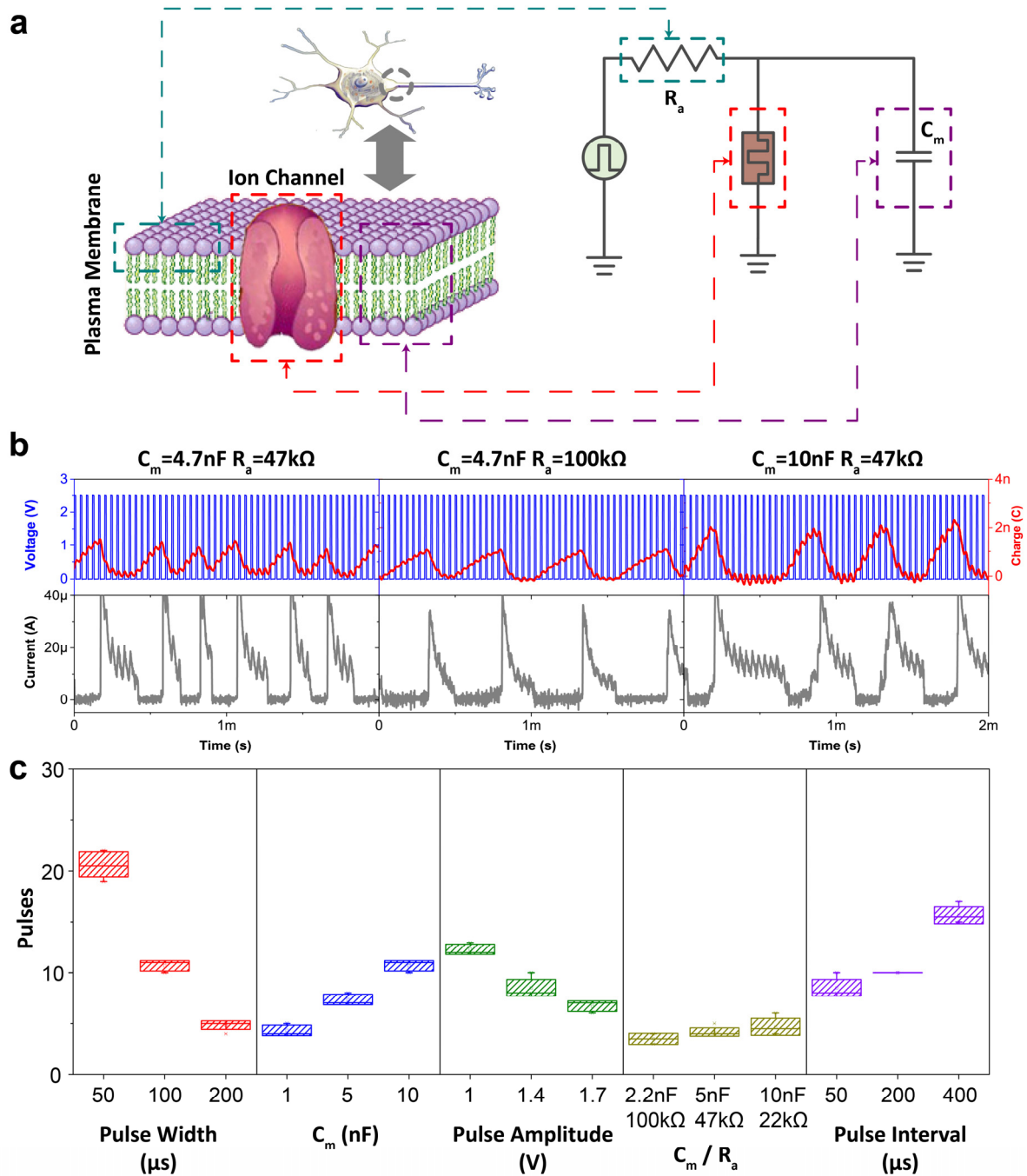
Figure 2 Controlled firing of a diffusive memristor artificial neuron. **a**, Illustration of an ion channel embedded in the cell membrane near the soma of a biological neuron. The inputs from the

dendrites are integrated on the capacitance of the membrane and the ion channel opens if the threshold condition is reached. Also shown is the analogous electrical integrate-and-fire circuit of the artificial neuron in which the diffusive memristor functions acts as the ion channel and the capacitor acts as the membrane. **b**, Response of the integrate-and-fire circuit to multiple consecutive pulses and the influence of varying membrane capacitance C_m and axial resistance R_a shows how the number of pulses required to charge the capacitor up to the memristor threshold increases with enlarging C_m or R_a. The current pulse across the diffusive memristor coincides with the discharge of the capacitor, clearly demonstrating that the device is actively firing a pulse of stored charge. **c**, Controlled firing response of the integrate-and-fire circuit under different input and circuit conditions. A similar effect as in **b** can be observed by changing the input parameters such as the pulse width (shorter pulses result in larger number of pulses before firing), pulse interval (shorter intervals result in smaller pulse number), and circuit parameters such as capacitance (larger capacitance delays the firing). Changing the input resistance while keeping the RC constant results in a small or no change in the firing.
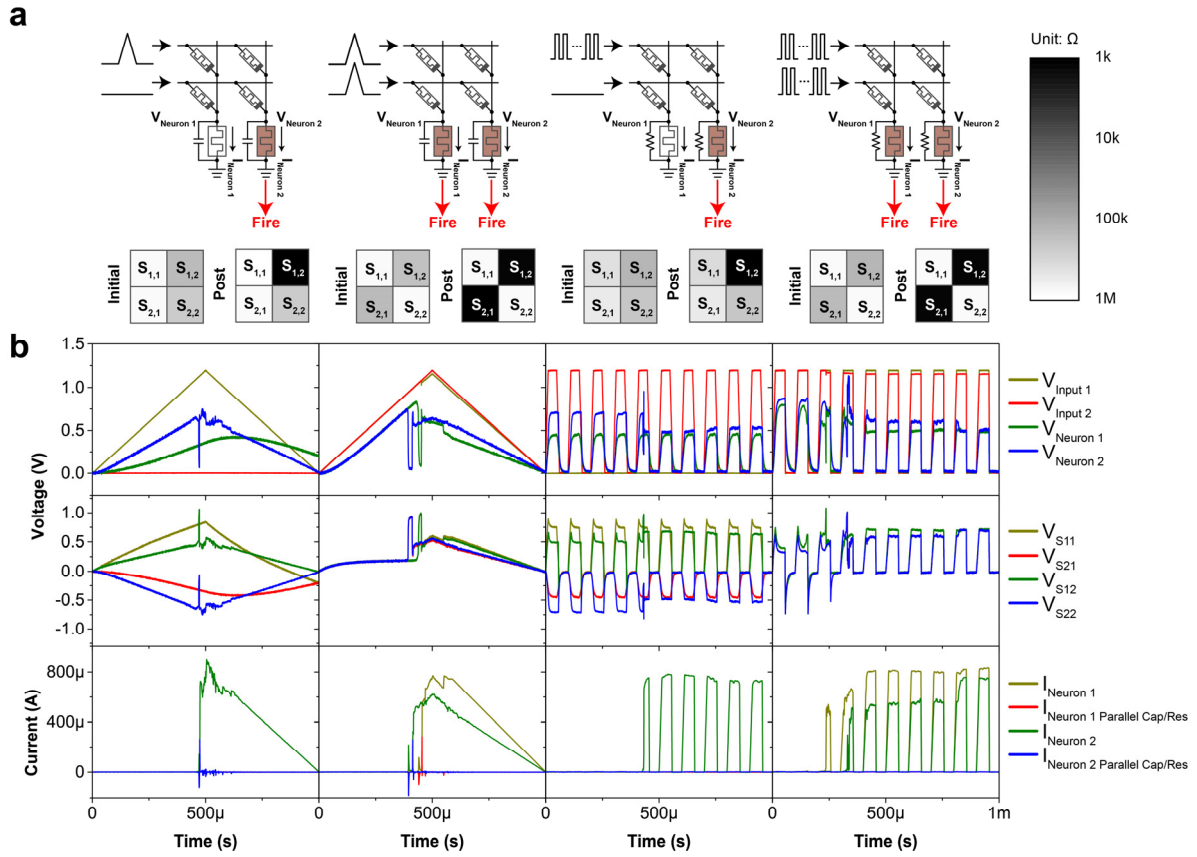
Figure 3 Experimental demonstration of unsupervised synaptic weight update using a 2×2 drift memristor array interfaced with two diffusive memristor artificial neurons at the output of each column, illustrating circuits with large and small capacitance. **a**, Schematics of the circuits, the pre-synaptic inputs, the post neuron outputs, and conductance map of the synapse array before and after training, respectively. All synapses were initialized to the high resistance state with some stochastic variation before training. **b**, The measured pre-synaptic signals, the potentials across neurons and synapses, and the neural currents. Upon receiving a '10' input vector, the right neuron fires with both RC (1st column) and internal Ag dynamics (3rd column) mechanisms, which programs the synapse S_{12}. The input vector '11' results in the firing of both neurons and programs

27

both S_{12} and S_{21} at the same time, with both RC (2nd column) and internal Ag dynamics (4th column) mechanisms.
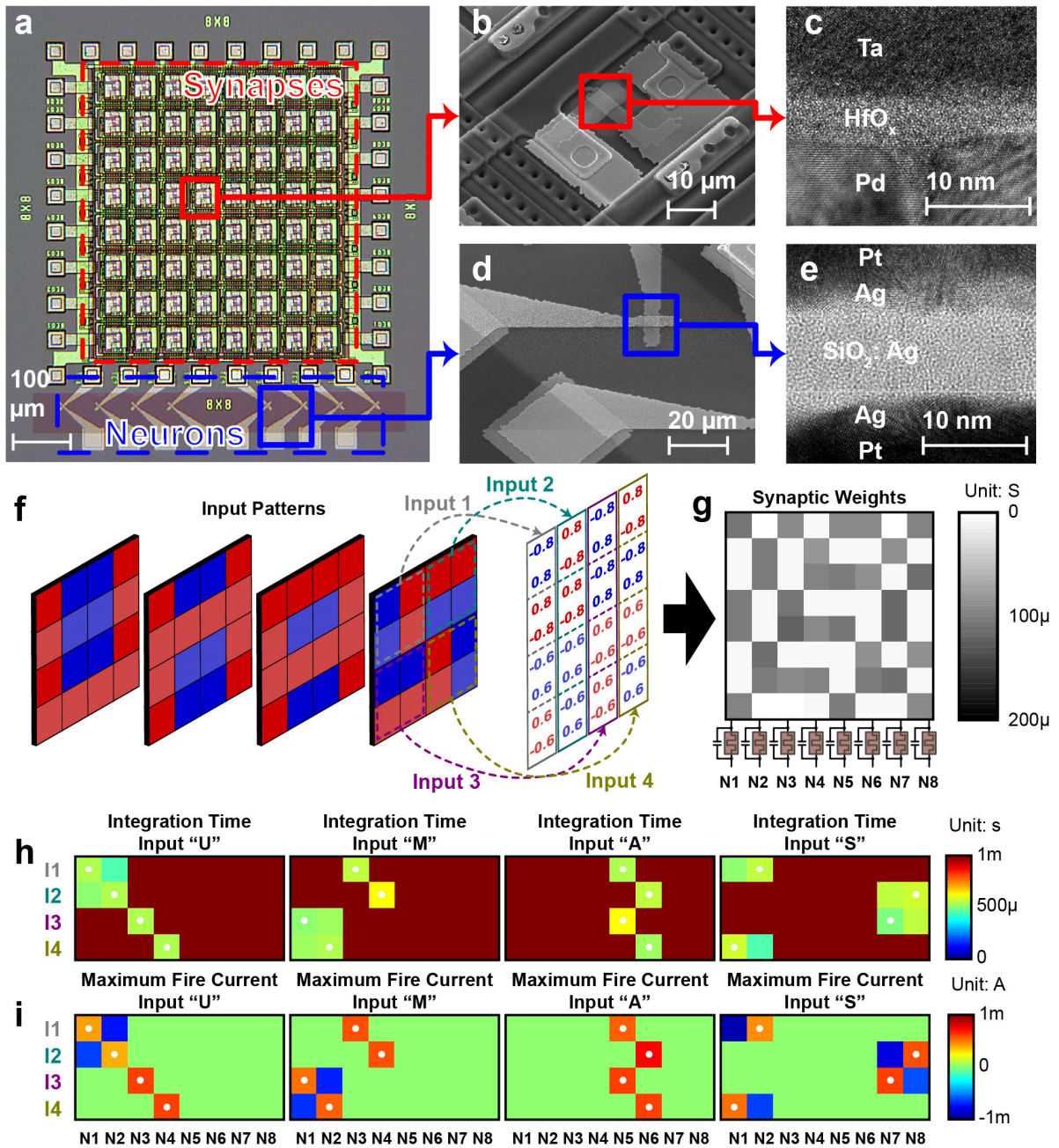
Figure 4 Fully integrated memristive neural network for pattern classification. **a**, Optical micrograph of the integrated memristive neural network, consisting of an 8×8 1T1R memristive synapse crossbar interfacing with 8 diffusive memristor artificial neurons (Each neuron used in

this demonstration has an external capacitor.). **b**, Scanning electron micrographs of a single 1T1R cell. Memristive synapses of the same row share bottom electrode lines while those of the same column share top electrode and transistor gate lines. **c**, Cross-sectional transmission electron microscopy image of the integrated Pd/HfO$_x$/Ta drift memristor prepared by focused-ion-beam cutting. **d**, Scanning electron micrograph of a single diffusive memristor junction. **e**, High-resolution transmission electron micrograph of the cross-section of the Pt/Ag/SiO$_x$:Ag/Ag/Pt diffusive memristor showing amorphous background SiO$_x$ with nano-crystalline thin Ag layers. **f**, The input pattern consists of 4 letters 'UMAS' with artificially added noise. Each input pattern consists of 4×4 pixels which are divided into four inputs (Input 1, Input 2, Input 3, and Input 4). Each input covers a sub-array of 2×2 size (4 pixels) of the original pattern using differential pairs as listed. Triangular voltage waveforms are fed to the 8 rows of synapses of the network. **g**, Measured conductance weights of the memristors after programming the 8 convolutional filters (1 filter per column) onto the 8×8 array using a differential pair scheme. Each of the 8 columns is interfacing with a diffusive memristor neuron at the end of the column. **h-i**, Measured integration time and maximum amplitude of fire current of the artificial neurons as responses to the 'UMAS' input patterns. Each individual input pattern is associated with its unique firing pattern of the 8 artificial neurons. The ideal output patterns are marked by the white dots for neurons with positive fire current flowing out of the network.
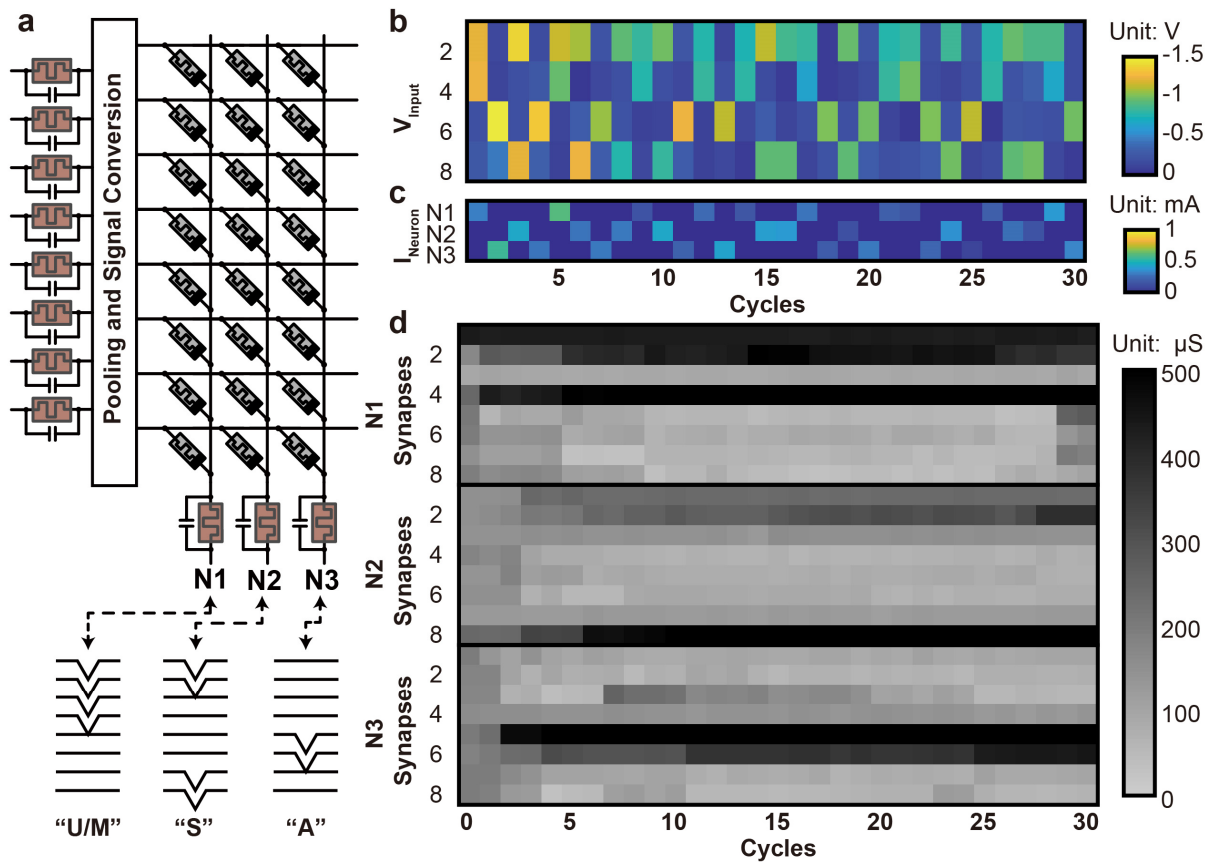
Figure 5 Unsupervised training of a fully connected network based on the integrated all-memristive neural network. **a**, The schematic diagram of the 8×3 network with inputs based on the outputs of the neurons in Figure 4. The prototypical patterns of neurons after training correspond to the input letters "U/M", "S", and "A" in Figure 4, respectively. **b-d**, The input patterns (peak voltages of triangular waveforms), peak neuronal currents, and synaptic weights at each training cycle. The synapses of the N1, N2, and N3 neurons quickly diverge from the initial 100μS and evolve by self-organizing processes to patterns with increasing similarities to one of the prototypical patterns in **a**. The magnitude of input patterns in **b** reduces in the first few cycles and becomes stable due to conductance saturation of the diverged drift memristor synapses.