nature materials

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41563-017-0001-5

In the format provided by the authors and unedited.

SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations

Shinhyun Choi[1,2], Scott H. Tan[1,2], Zefan Li[1,2], Yunjo Kim[1,2], Chanyeol Choi[1,2], Pai-Yu Chen[3], Hanwool Yeon[1,2], Shimeng Yu[3] and Jeehwan Kim[1,2,4]*

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona, USA. [4]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Shinhyun Choi and Scott H. Tan contributed equally to this work. *e-mail: jeehwan@mit.edu

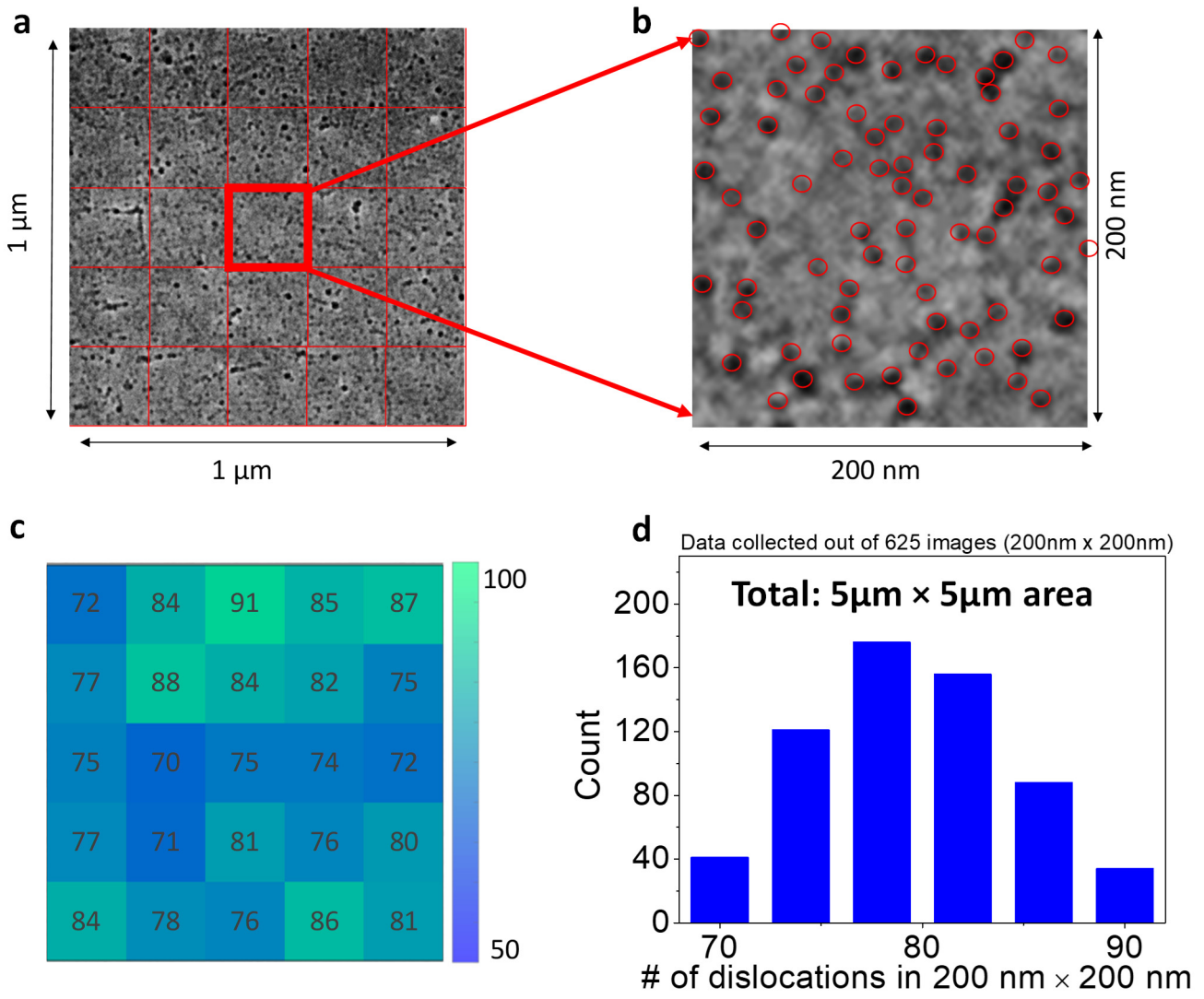


Figure S1. Dislocation density analysis of heteroepitaxial SiGe on Si. The dislocation density is in the range of $10^{11}/cm^2$. This suggests that the epiRAM can be scaled down to tens of nanometers. a. a SEM image showing decorated dislocations in 1 µm × 1 µm area. **b.** Magnified SEM with dislocation pinholes highlighted by red circles in the SEM image. In the 200 nm × 200 nm area, 75 dislocations can be observed **c.** Color map highlighting the distribution of threading dislocations across the entire 1 µm × 1 µm area. **d.** Histogram showing the dislocation counts in 5 µm × 5 µm area.

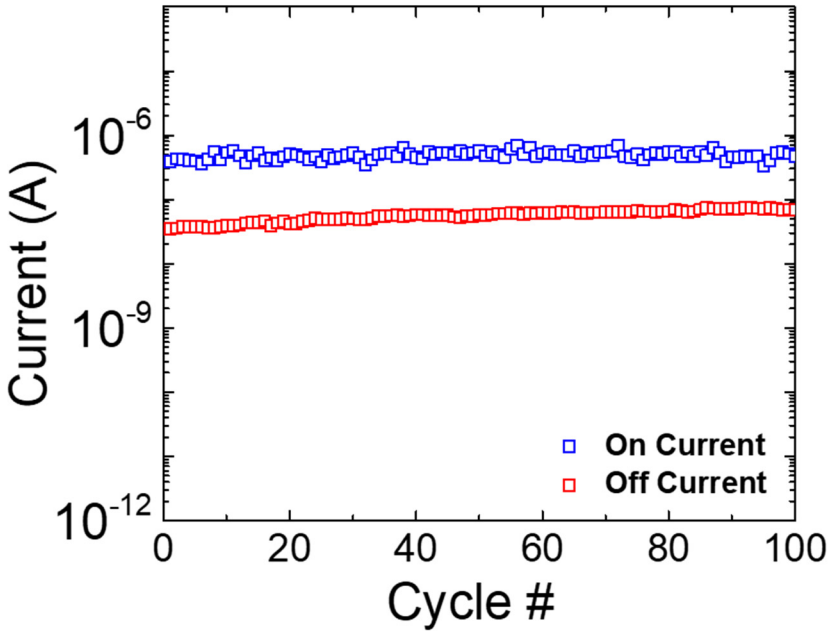


Figure S2. Uniform On/Off current (measured at 0.8 V) over 100 DC *I-V* sweeps of SiGe epiRAM without Schimmel etch.

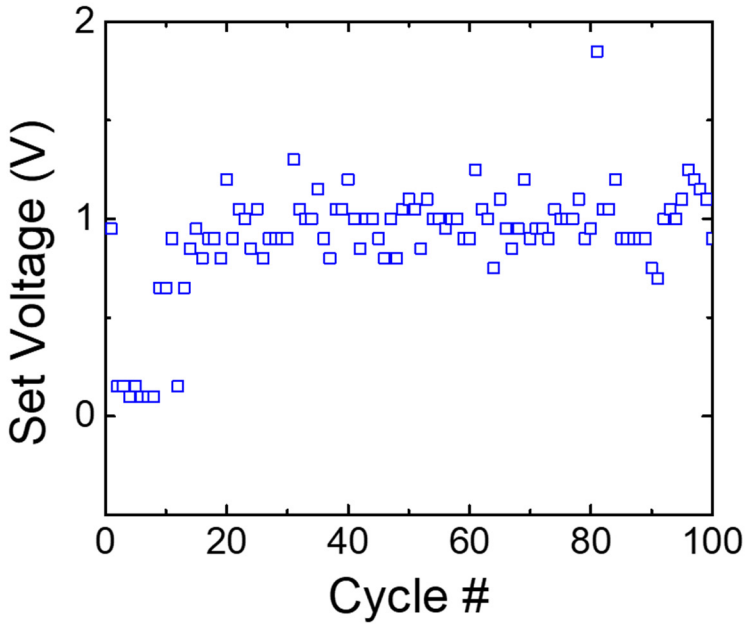


Figure S3. The temporal set voltage variation when replacing SiGe epilayer with amorphous-Si. The set voltage variation ($\sigma/\mu = 0.28$) is significantly higher than SiGe epiRAM since stochastic fluctuations can occur in three-dimensions.

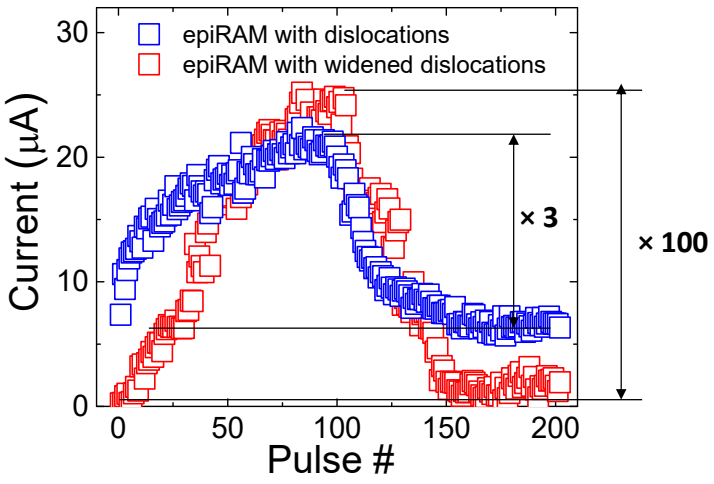


Figure S4. Analog switching potentiation and depression profiles for etched and unetched epiRAM. The current level at 2 V read pulses (1 ms) changes throughout the train of 100 set pulses (5 V, 5 μs) and 100 reset pulses (-3V, 5 μs). The magnitude of the Burger's vector in SiGe is ~0.3 nm and the diameter of Ag ions are ~0.23 nm, which result in tight spatial accommodation. As a result, the analog on/off ratio is only 3. On the other hand, defect-selective etching provides additional free sites for Ag occupation along the pipeline for ion migration, which boost analog on/off ratio to above 100 using the same pulse train conditions. This confirms that securing enough space at dislocation allows linearity in conductance response. The diameter of dislocation is estimated to be few nanometers which offers enough space to accommodate Ag ions (few angstrom radius).

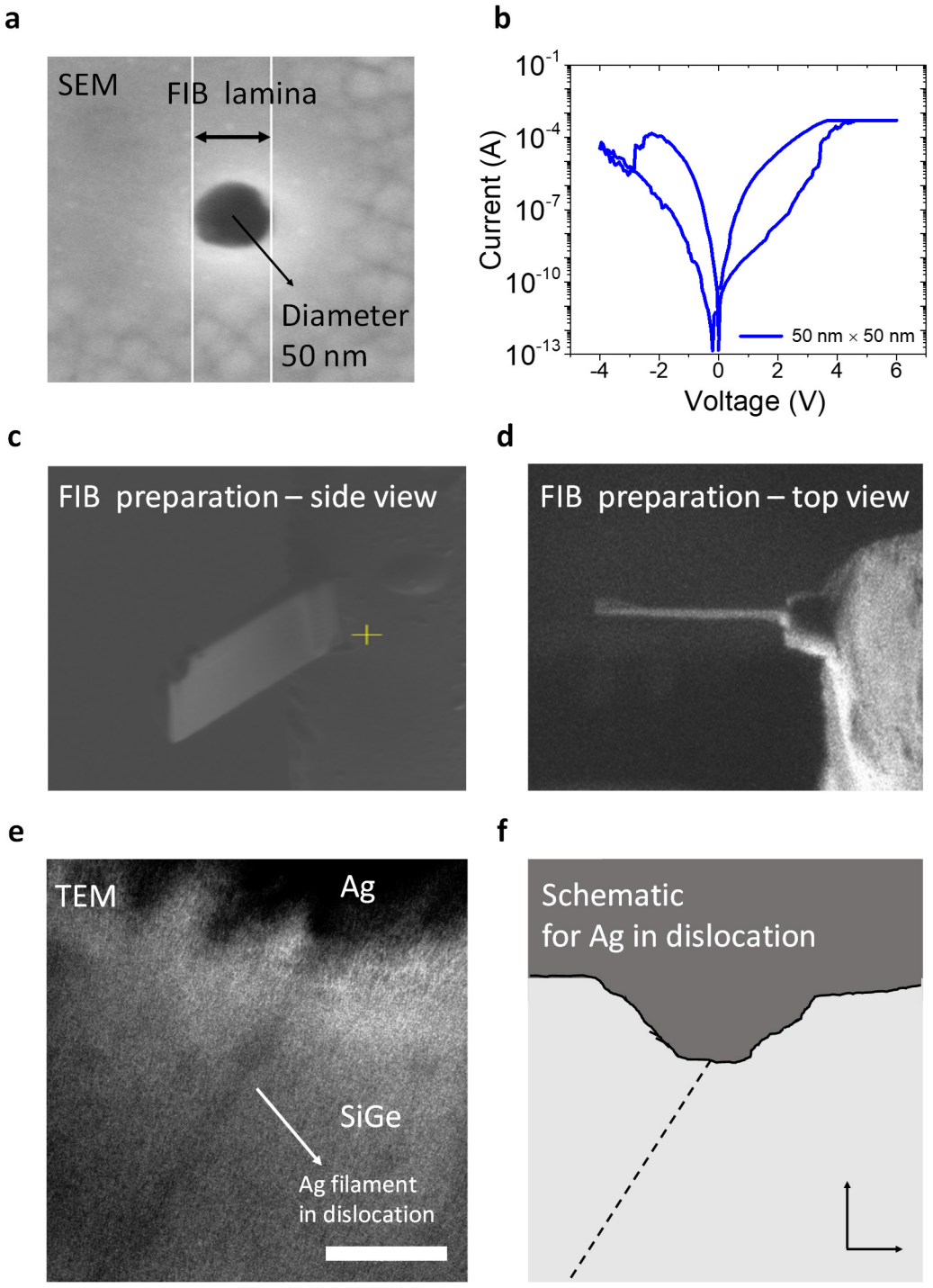


Figure S5. Ag filament confinement in a widened dislocation. a. Plan-view SEM image of active area (50 nm diameter). The other parts are covered by 100nm SiO_2 layer. **b.** I-V curve measured from the active area shown in Fig. S5a after Ag deposition. After the electrical measurement, the device is set to low resistance state before FIB preparation. **c,d.** The images during FIB preparation for TEM observation in side-view and top view, respectively. **e.** Cross-sectional TEM image of Ag filament confined in the dislocation pipe. Scale bar: 12 nm. **f.** Schematic for Ag filament confinement in a dislocation.

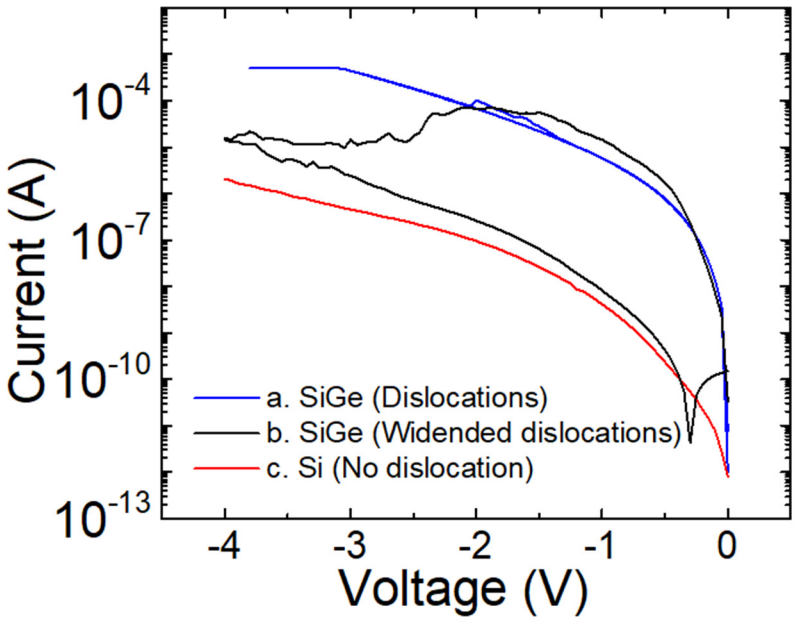


Figure S6. Negative bias DC *I-V* characteristics for a. SiGe epiRAM without defect-selective etch, b. SiGe epiRAM with 5s etch and c. Ag/ intrinsic Si/ p-type Si device. EpiRAM without defect-selective etch does not fully reset. EpiRAM with 5s defect-selective etch resets to low current state similar to intrinsic Si device, indicating Ag can successfully be retracted. Furthermore, Ag/ intrinsic Si/ p-type Si device does not switch indicating that dislocations are the paths for ionic transport in epiRAM.

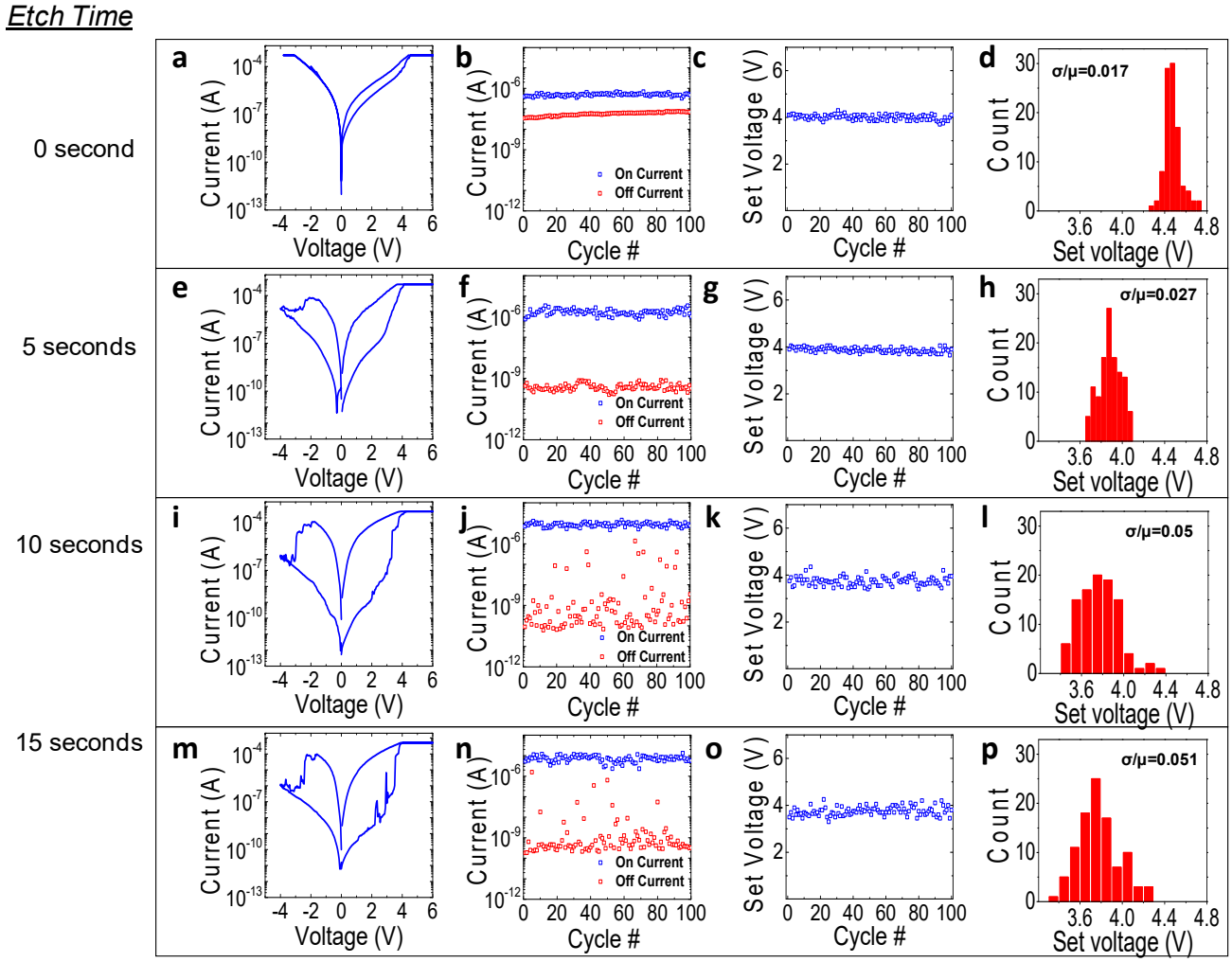


Figure S7. EpiRAM device performance for different defect-selective etch times; no etch (1st row), 5 sec (2nd row), 10 sec (3rd row), and 15 sec (4th row). Shown are DC *I-V* characteristics, on/off current states at 0.8 V over DC *I-V* cycles, and set voltage distributions, respectively, for a-d. unetched epiRAM, e-h. 5s-etched epiRAM, i-l. 10s-etched epiRAM, and m-p. 15s-etched epiRAM. Defect-selective etching increases on/off current ratio. However, etching for over 5s results in unstable on/off current states and more set voltage variation due to excessive dislocation widening.

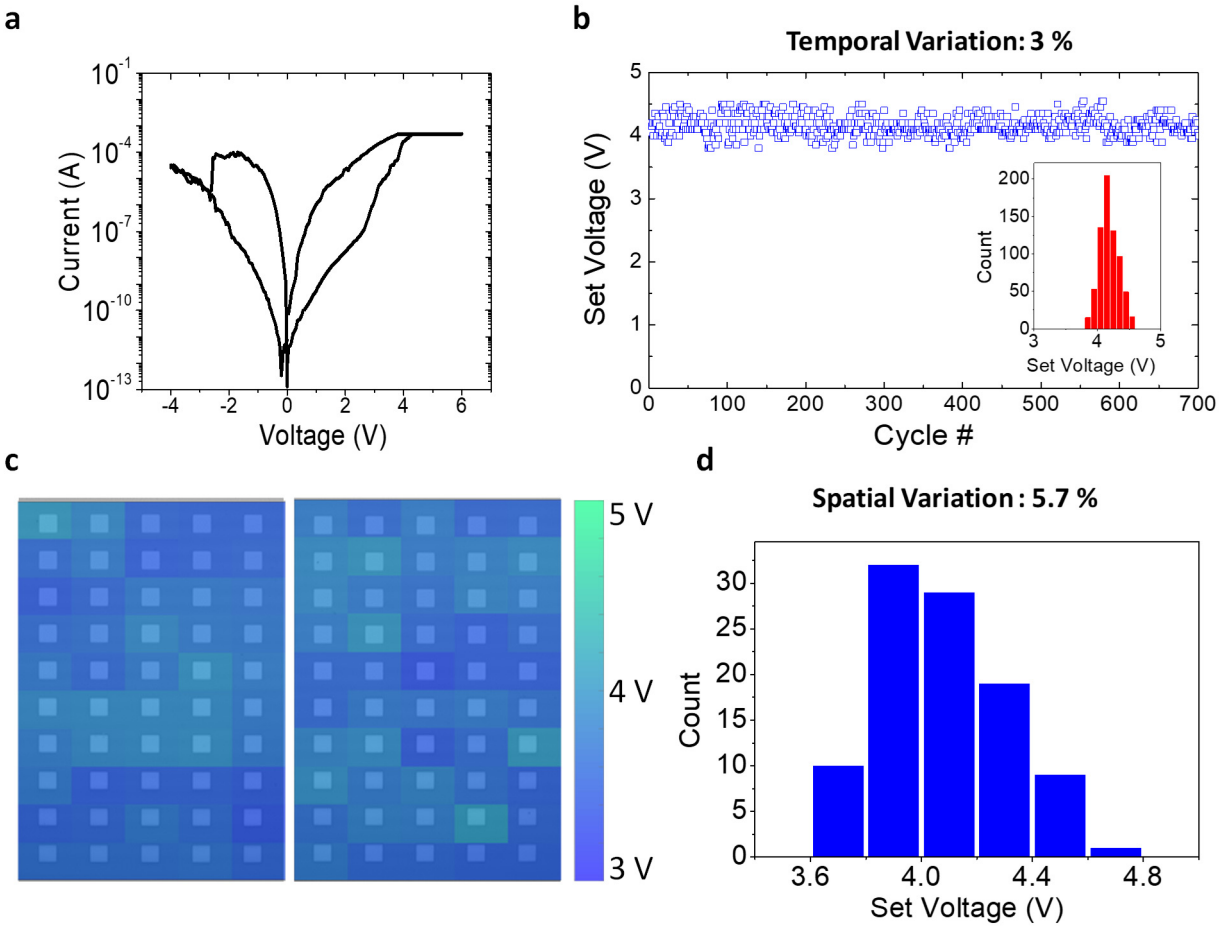


Figure S8. Spatial and temporal variation of 25 nm x 25 nm epiRAM devices. a. I-V characteristics of 25 nm × 25 nm device. **b.** Temporal set voltage variation after 700 cycles (inset: histogram of set voltage). **c.** Color map of set voltages at 100 devices overlapped on optical micrographs of the epiRAM devices. **d.** histogram for spatial set voltage distribution shown in Fig. S8c.

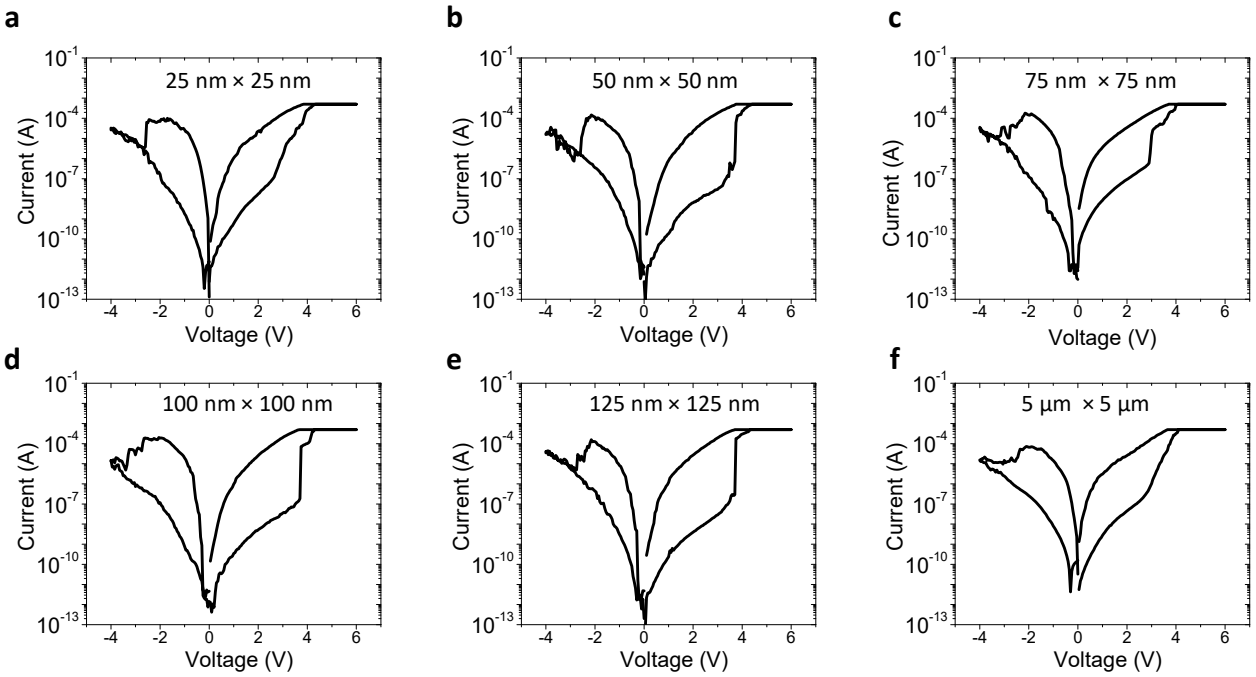


Figure S9. Semi-logarithmic DC *I-V* characteristics of a. 25 nm × 25 nm, b. 50 nm × 50 nm, c. 75 × 75 nm, d. 100nm × 100 nm, e. 125 nm × 125 nm, and f. 5 µm × 5 µm devices. Similar I-V characteristics suggest that switching only occurs in localized areas with a limited number of dislocations responsible for the majority of ionic movement among multiple dislocations.

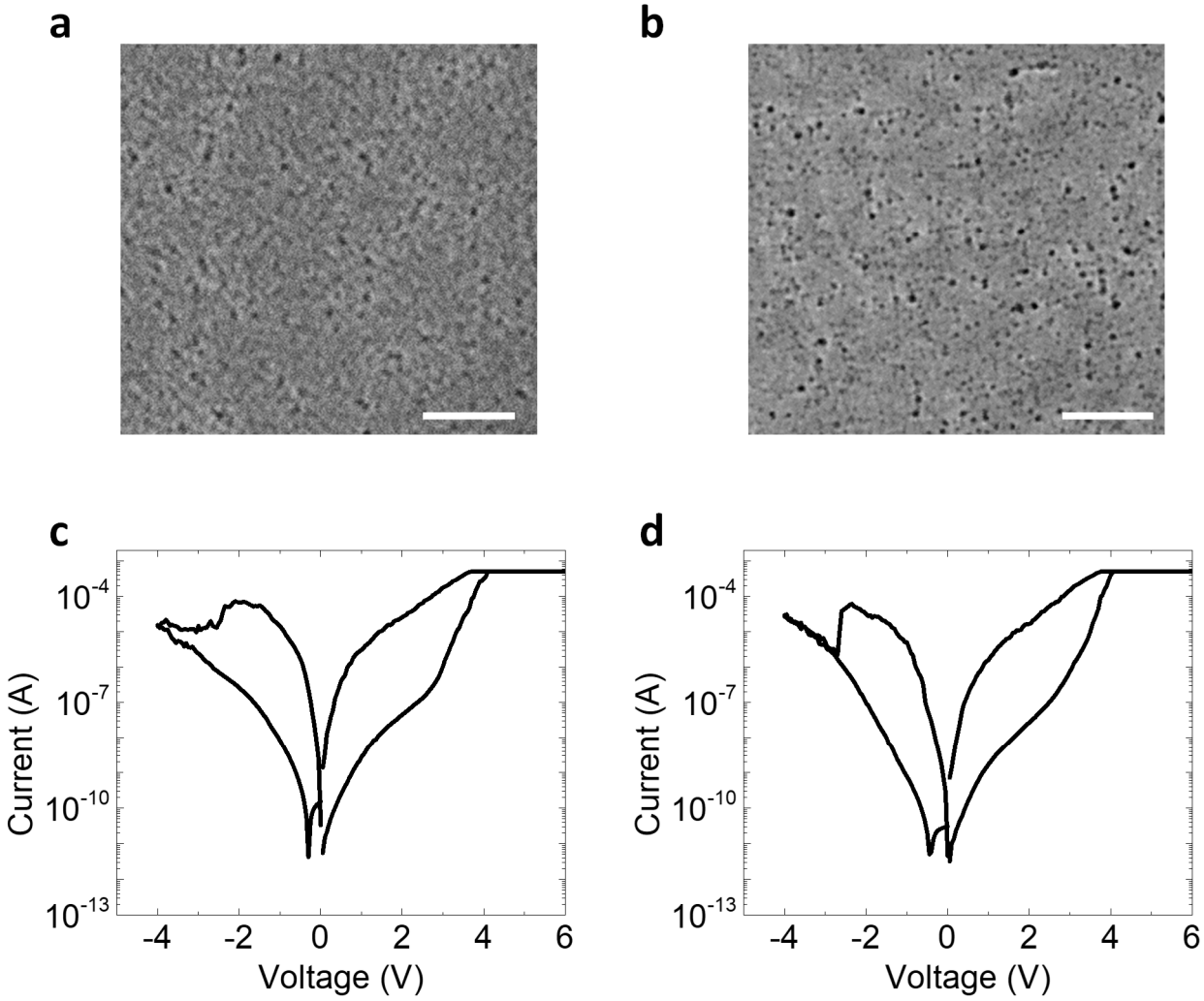


Figure S10. I-V DC sweeps with different Ge composition in crystalline SiGe. a. a SEM image of a 60 nm 5 sec-etched crystalline SiGe layer with 10% Ge composition. Scale bar: 200nm. **b.** a SEM image of a 60 nm 5 sec-etched crystalline SiGe layer with 30% Ge composition. Scale bar: 200nm. **c.** *I-V* DC sweep of a 5 sec-etched epiRAM with 10% Ge composition. **d.** *I-V* DC sweep of a 5 sec-etched epiRAM with 30% Ge composition. This confirms that the dislocation density changed by Ge composition does not affect device characteristics.

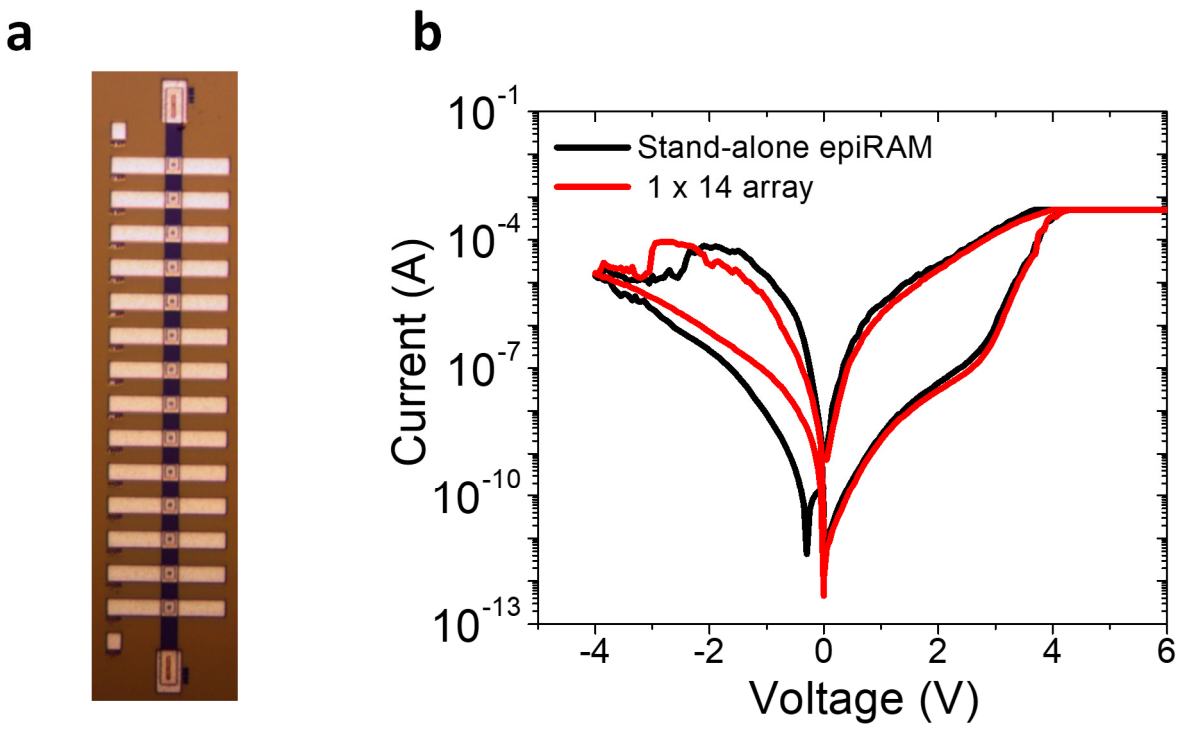


Figure S11. a. Optical microscope image of fabricated 1 × 14 epiRAM arrays. **b.** I-V characteristics from stand-alone and 1 × 14 crossbar structure.

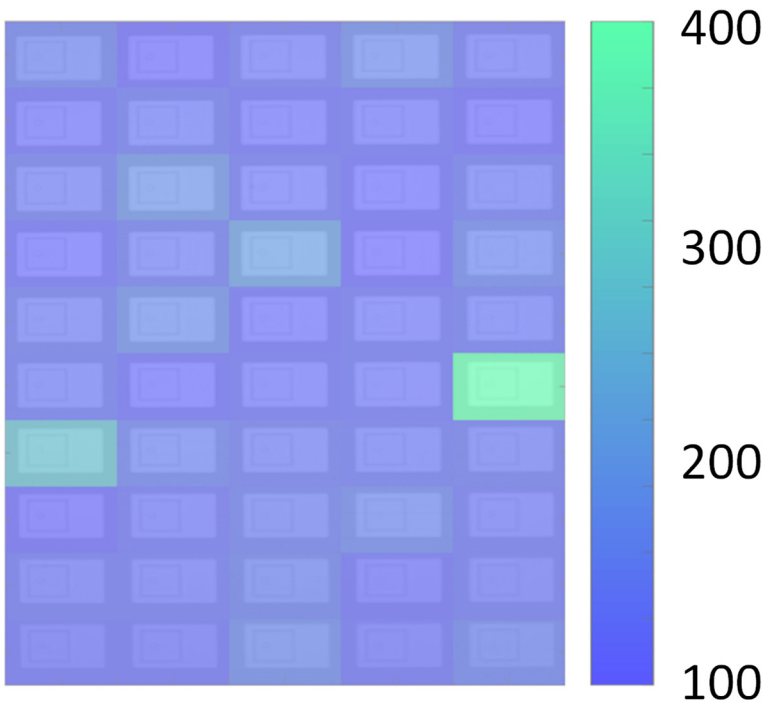


Figure S12. Color Map displaying high analog on/off current ratios of 50 etched epiRAMs overlapped on optical microscope image. The analog on/off current ratio expressed in the color bar is defined as the current measured by read pulse (2V, 1 ms) after 100 potentiation pulses (5 V, 5 µs) divided by the current measured by read pulse after 100 depression pulses (-3 V, 5 µs). High analog on/off current ratio (>100) is achieved by all measured epiRAM, which is important for achieving high learning accuracy.

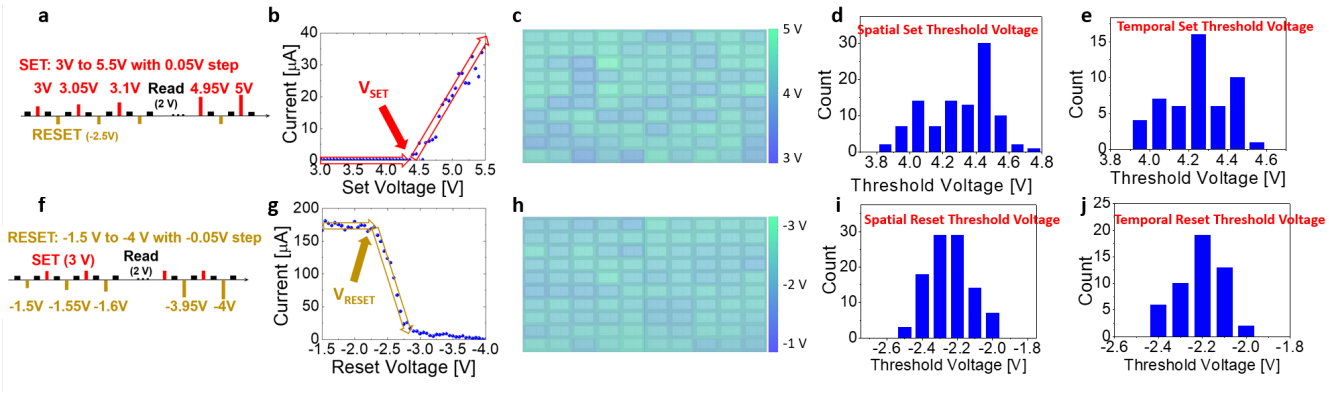


Figure S13. Analog set and reset voltage analysis. **a, b.** Applied pulse trains for measuring set voltage and reset voltage, respectively. **c, d.** EpiRAM conductance first begins to increase (from HRS) at the set threshold voltage and starts to decrease (from LRS) at the reset threshold voltage, respectively. **e, f.** Histograms illustrating the spatial variation (4.8%) for set threshold voltage and spatial variation (5.3%) for reset threshold voltage, respectively. **g, h.** Temporal variation (3.9%) for set threshold voltage and temoral variation (4.8%) for reset threshold voltage, respectively. The pulse widths for set/reset pulses are 5 μs and pulse widths for read pulses are 1 ms.

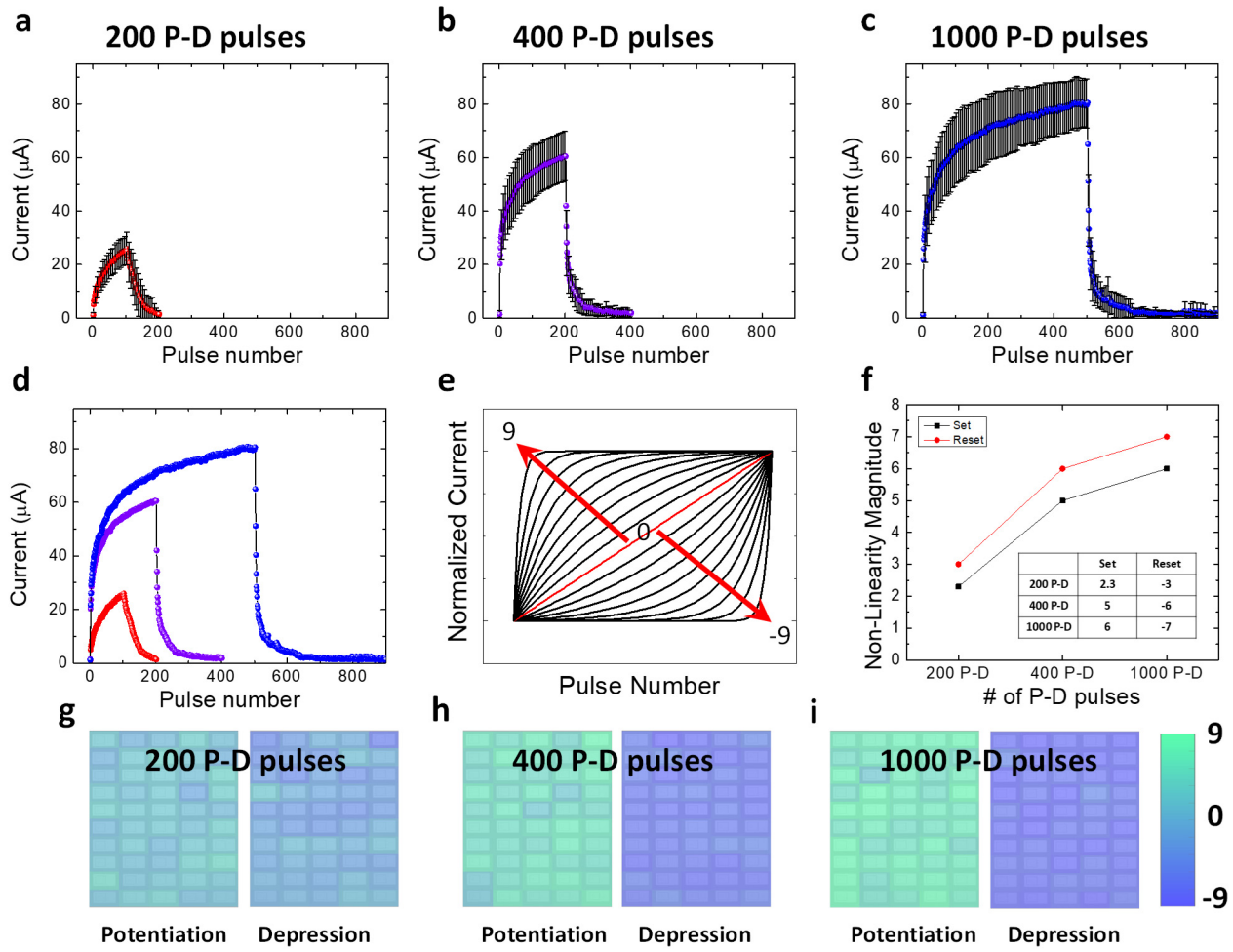
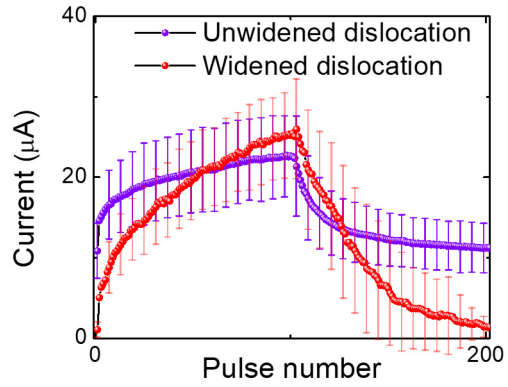


Figure S14. Potentiation and depression linearity analysis. a-c. The current recorded at 2 V read pulses (1 ms) after 200/400/1000 P-D pulses consisting of 100/200/500 write pulses (5 V, 5 μs) and 100/200/500 erase pulses (-3V, 5 μs), respectively. This suggests that applying more pulses increases nonlinearity of conductance change. **d.** The slopes of potentiation and depression for different number of pulses can be represented by non-linearity magnitude. **e.** The definition of non-linearity magnitude **f.** The non-linear magnitude depending of number of P-D pulses. Inset: table for non-linearity magnitudes for 200, 400, and 1000 potentiation and depression (P-D) pulses. **g-i.** Color maps showing spatial variation of linearity for 200, 400, and 1000 P-D pulses.

Reproducible training requires stabilization steps typically composed of repeating 30 sets of 100/200/500 pulses after which the filament is settled into the dislocation. The reason of the variation on slopes depending on pulse number is from conduction channel difference of the initial states. The results are all collected from one set of pulse chain. Therefore, the initial states could be different based on the pulse number. The physical state of the conduction channel after reset pulses are applied is crucial for understanding the origins of this phenomenon. It is expected that more erase pulses give more resistive states. For example, with 100 erase pulses, the conduction channel of the initial state could be not erased fully compared to the case of 200/500 erase pulses. These changes make the initial slope difference among 200/400/1000 P-D pulses and opens a possibility of linear conductance change with 200 P-D pulses. For the training algorithm for neural network requires linear weight update. The linearity of conductance change for linear weight update could reduce the complexity of periphery circuit that compensates the non-linear conductance update. Repeated measurement with statistics

show consistent slopes depending on the pulse numbers. For further verification, we have measured 100 devices with different pulses and their statistics are shown in Fig. S14.

a



b

	Set	Reset
Unwidened Dislocation	4.7	-4.5
Widened Dislocation	2.3	-3

c

d

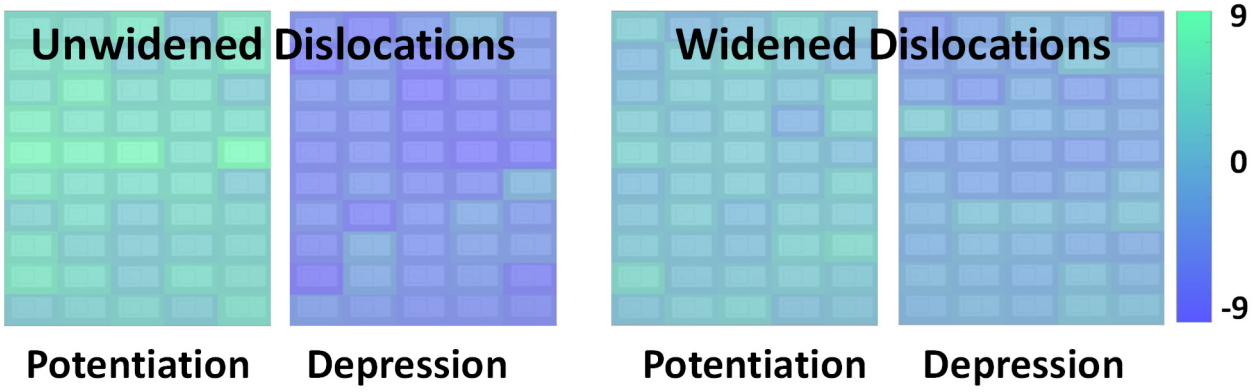


Figure S15. Analog potentiation and depression of epiRAM with unwidened/widened dislocations. **a.** Average and standard deviation of P-D responses for devices with unwidened/widened dislocations **b.** The average values of non-linearity magnitude of set/reset for unwidened/widened dislocations. **c.** Color maps showing spatial variation of linearity for unwidened dislocations. **d.** Color maps showing spatial variation of linearity for widened dislocations.

When the dislocation is not etched, the magnitude of burgers vector of the SiGe threading dislocation and the diameter of Ag are comparable. This means ion transport is spatially limited in virgin dislocations. As a result, current range by the pulse response on non-etched sample is very limited and shows very low on/off ratio and non-linear conductance change. With Schimmel etched sample, high on/off ratio and linear conductance update have been observed. Widening the dislocations by Schimmel etch opens additional free sites for Ag occupation along the pipeline. This allows for efficient tuning of resistance.

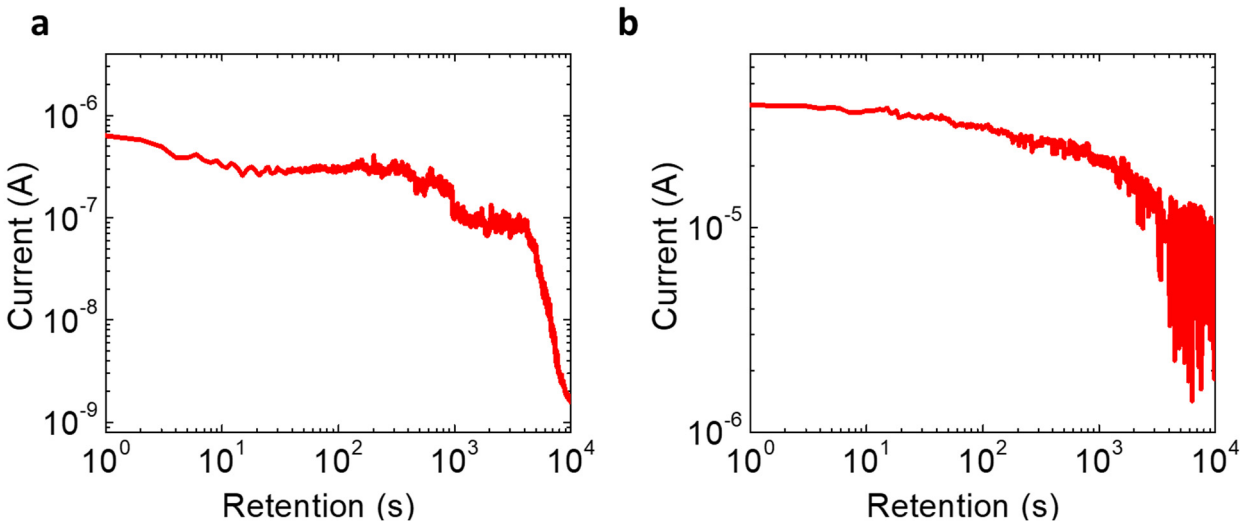


Figure S16. Retention test at room temperature for a. epiRAM with untreated dislocations, and b. epiRAM with over-etched (10s) dislocations (read voltage: 1.5 V). The injected Ag in untreated narrow dislocations is unstable and lose the conduction channel due to mechanical stress. This results in poor retention in epiRAM with untreated dislocations. The epiRAM with over-etched dislocations (10s) shows poorer retention compared to that containing dislocations with optimal spacing (5s etch) as shown in Fig. 3c in the maintext. The over-etched dislocation space would allow the motion of Ag in the filament eventually resulting in breaking the continuity of Ag filaments, leading to poor retention.

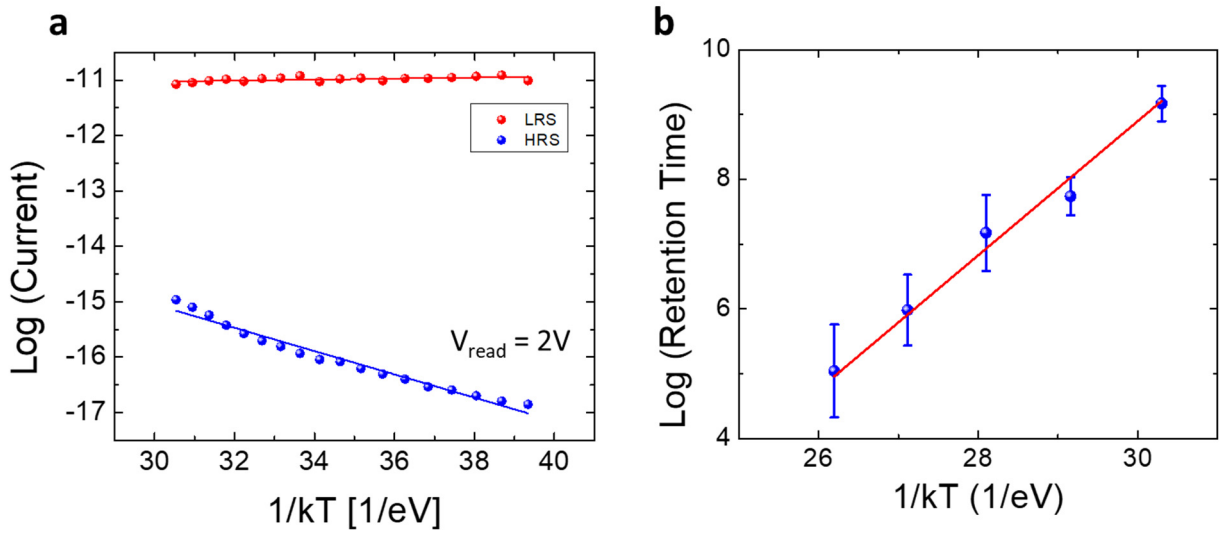


Figure S17. a. Arrhenius plot for low-resistance state (LRS) and high-resistance state (HRS) current and retention at LRS. a. LRS remains nearly constant (activation energy E_{LRS} = -9 meV), which suggests the conduction channel is mainly metallic. On the other hand, HRS is Arrhenius with temperature (E_{HRS} = 200 meV), indicating epiRAM is semiconducting for electron transport in the HRS state. The current is measured at 2V read bias. **b.** Retention test was performed at elevated temperatures. The activation energy of 1.04 eV for the Ag ion diffusion is extracted from Arrhenius plot, which is a similar value reported for Ag diffusion in single-crystalline Si with dislocations[S1]. The extrapolation of the plot to room temperature indicates 1.87 years of retention. The extrapolation of the plot to room temperature indicates 1.87 years of retention. The data are collected twice from 10 devices at each temperature between 398K to 443K.

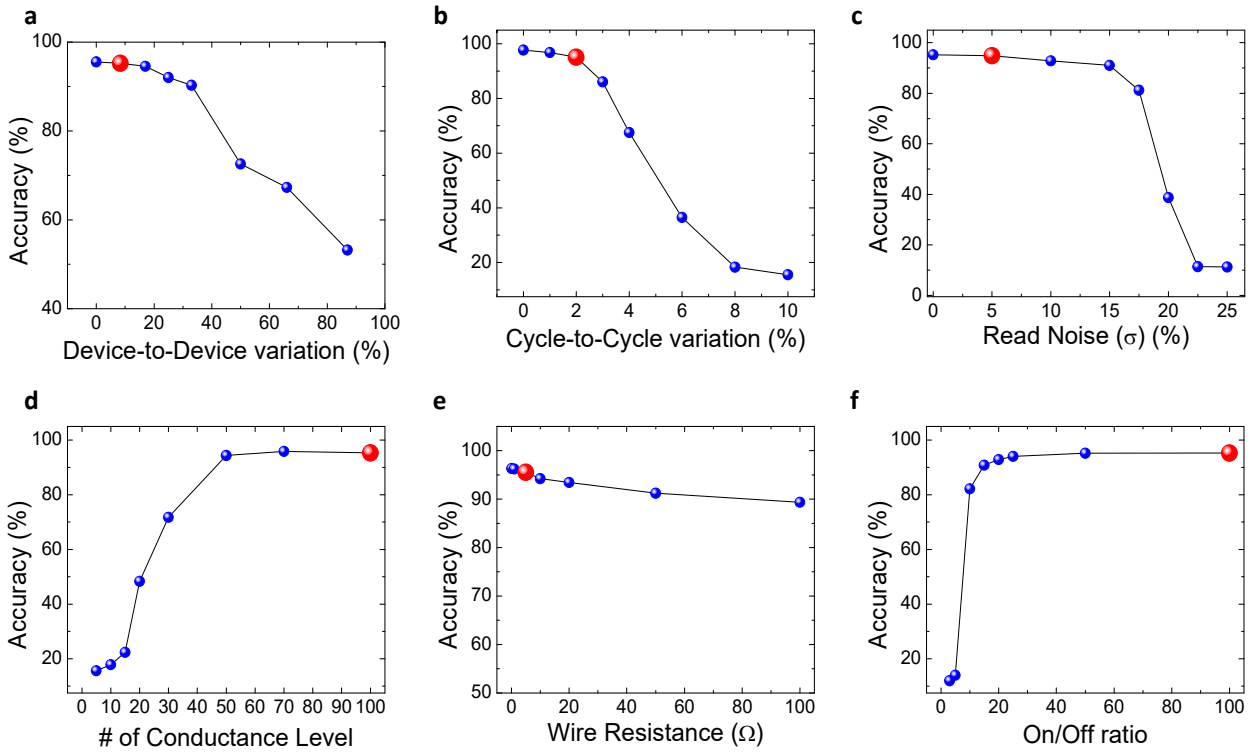


Figure S18. Dependence of recognition accuracy on a. Device-to-device variation **b.** Cycle-to-cycle variation, **c.** Read noise, **d**. Conductance level, **e.** Wire resistance, and **f.** On/off ratio. The red dots indicate the value extracted from epiRAM. The parameters of epiRAM satisfy the standard to achieve high accuracy.

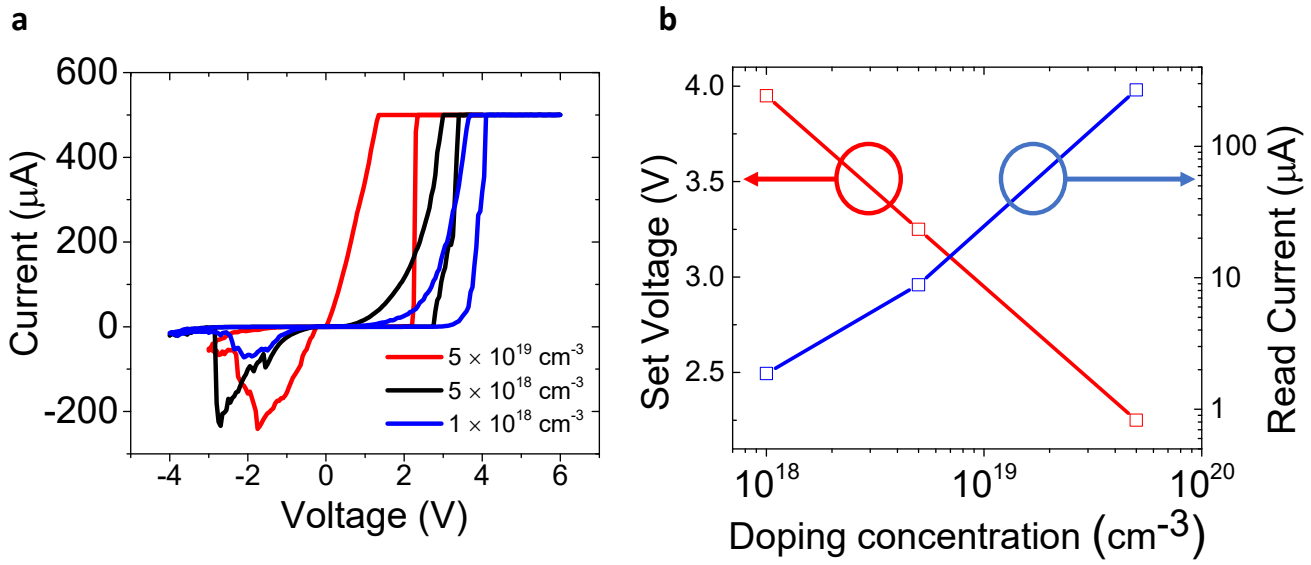


Figure S19. Tunability of set voltage and read current of epiRAM by modulating Schottky barriers at the Ag-filament/Si interface. a. *I-V* plots of epiRAM with different doping concentration of Si layer below SiGe, **b.** set voltage and read current(at 0.8 V) plotted as a function of doping concentration suggesting that higher doping concentration results in lower set voltage and higher read current due to the lower Schottky barrier. Using this technique, epiRAM characteristics can be optimized for specific end-use applications. For example, linear I-V read out is better for accuracy while a non-linear curve is ideal to reduce power consumption and obtain selector-free arrays without sneak path. These parameters can be adjusted by changing the Schottky barrier.

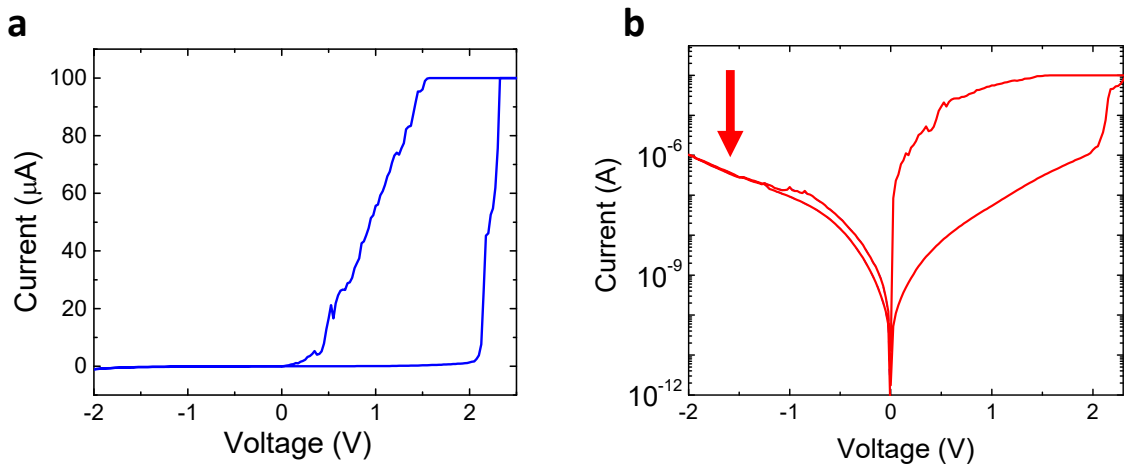


Figure S20. a. Linear-scale and **b.** Logarithmic-scale I-V curve of p-i-p SiGe back-to-back epiRAM. LRS-state current is rectified at negative bias. This property can be utilized to suppress sneak path currents in crossbar arrays.

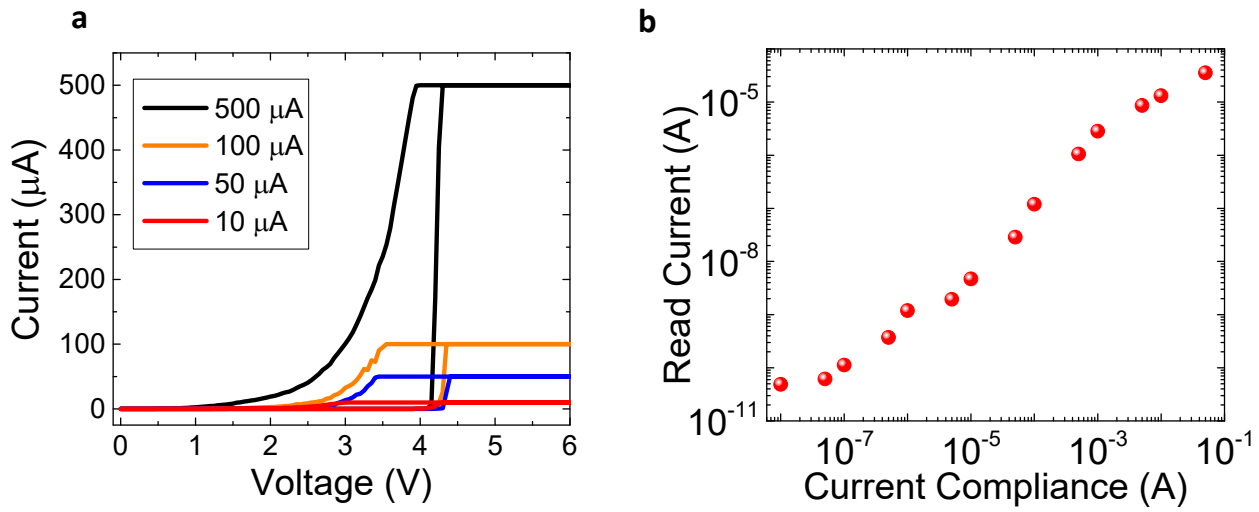


Figure S21. The multiple resistance states controlled by external circuitry for digital applications. **a.** linear-scale DC *I-V* plots showing that different device states achieved by different current compliance while the set voltage remains. **b.** multiple current levels (read at 0.8 V) accessible between 10^{-11} A and 10^{-5} A with current compliance from 10^{-8} A to 5×10^{-2} A.

The effect of Ge composition and the possibility of device scaling

During heteroepitaxial growth of SiGe on Si, lattice mismatch induces strain, which results in the formation of threading dislocations in the SiGe layer upon relaxation. Because the Ge composition directly corresponds to the magnitude of SiGe:Si lattice mismatch, Ge composition can be used to control dislocation density. As shown in Fig. S10a and Fig. S10b, the dislocation density increases with Ge composition. Fig. S10c and Fig. S10d represent DC *I-V* characteristics of epiRAM with 10% and 30% Ge composition, respectively. Set voltage, reset voltage and on/off current levels are similar for both compositions, which suggests that Ge composition (determining the dislocation density) does not largely affect electrical characteristics of epiRAM. It implies that the device size can be reduced to tens of nanometer to achieve ultrahigh density arrays.

CMOS compatibility of our epiRAM

Our epiRAM allows monolithic integration of neuromorphic synaptic arrays with Si or SiGe CMOS neurons at the peripheries of the arrays at the front end of the line. Since SiGe epitaxial layers can be shared to form epiRAM and transistors, epiRAM can be integrated on premade Si or SiGe transistors neurons as it is processed at room temperature. Optionally combined with layer transfer techniques[S2], entire epiRAM arrays can be peeled off and stacked on CMOS circuits at the back end of the line.

Behavioral device model for MLP simulation:

In the behavioral model capturing nonlinear conductance update, conductance change with number of pulses (P) is described with the following equations[S3].

$$G_{LTP} = B(1 - e^{(-\frac{P}{A})}) + G_{\min}$$

$$G_{LTD} = -B(1 - e^{(\frac{P - P_{\max}}{A})}) + G_{\max}$$

$$B = \frac{G_{\max} - G_{\min}}{1 - e^{(-\frac{P_{\max}}{A})}}$$

G_{LTP} and G_{LTD} are the conductance for LTP and LTD, respectively. $G_{\max}$, $G_{\min}$, and $P_{\max}$ are directly extracted from the experimental data, which represents the maximum conductance, minimum conductance, and the maximum pulse number required to switch the device between the minimum and maximum conductance states, respectively. A is determined by the nonlinearity of weight update and can be positive or negative. $A_{LTP} = 0.5032, A_{LTD} = -0.3868$ (Normalized by $P_{\max}$) is used to fit data in the main text. B is simply a function of A within the range of $G_{\max}$, $G_{\min}$, and $P_{\max}$.

MLP Simulation Parameters

We have employed simple circuitry to compensate the non-linear conductance change with pulse number[S4]. During each training epoch, one pattern is selected randomly from the 60,000 training set patterns. After one million training epochs, accuracy is tested with 10,000 separated testing sets.

Accuracy of this algorithm using software with similar network size is 98%[S4,S5]. Using binary input signal for the first two layers instead of gray-scale, the accuracy limitation of the network is 97%[S3]. Our simulation suggests that epiRAM can achieve 95.1% accuracy taking the finite on-off ratio, number of conductance level, cycle-to-cycle variation, wire resistance, and read noise into account. The evolution of accuracy to training epoch is shown in Fig. 4c for the ideal software and epiRAM device. Both converge quickly after half a million epochs. Figure S18 show the impact of device-to-device, cycle-to-cycle variation, read noise, conductance level, wire resistance, and on/off ratio on accuracy.

S1. Fisher, D. J. Diffusion in Silicon. 10 Years of Research. 555 (2000).

S2. Kim, Y. *et al.* Remote epitaxy through graphene enables two-dimensional material-based layer transfer. *Nature* **544,** 340–343 (2017).

S3. Yu, S. *et al.* Binary neural network with 16 Mb RRAM macro chip for classification and online training. in *2016 IEEE International Electron Devices Meeting (IEDM)* 16.2.1-16.2.4 (IEEE, 2016). doi:10.1109/IEDM.2016.7838429

S4. Kataeva, I., Merrikh-Bayat, F., Zamanidoost, E. & Strukov, D. Efficient training algorithms for neural networks based on memristive crossbar circuits. in *Proceedings of the International Joint Conference on Neural Networks* **2015–Septe,** 1–8 (IEEE, 2015).

S5. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86,** 2278–2323 (1998).