nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04485-1 OPEN

Signal and noise extraction from analog memory elements for neuromorphic computing

N. Gong[1,2], T. Idé[1], S. Kim[1], I. Boybat[3,4], A. Sebastian[3], V. Narayanan[1] & T. Ando[1]

Dense crossbar arrays of non-volatile memory (NVM) can potentially enable massively parallel and highly energy-efficient neuromorphic computing systems. The key requirements for the NVM elements are continuous (analog-like) conductance tuning capability and switching symmetry with acceptable noise levels. However, most NVM devices show non-linear and asymmetric switching behaviors. Such non-linear behaviors render separation of signal and noise extremely difficult with conventional characterization techniques. In this study, we establish a practical methodology based on Gaussian process regression to address this issue. The methodology is agnostic to switching mechanisms and applicable to various NVM devices. We show tradeoff between switching symmetry and signal-to-noise ratio for HfO_2-based resistive random access memory. Then, we characterize 1000 phase-change memory devices based on $Ge_2Sb_2Te_5$ and separate total variability into device-to-device variability and inherent randomness from individual devices. These results highlight the usefulness of our methodology to realize ideal NVM devices for neuromorphic computing.

[1] IBM T. J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA. [2] Department of Electrical Engineering, Yale University, 10 Hillhouse Avenue, New Haven, CT 06511, USA. [3] IBM Research-Zurich, Säumerstrasse 4, 8803 Rüschlikon, Switzerland. [4] Ecole Polytechnique Federale de Lausanne (EPFL), Route Cantonale, 1015 Lausanne, Switzerland. Correspondence and requests for materials should be addressed to T.A. (email: andot@us.ibm.com)

Over several decades, the von Neumann architecture has enabled exponential improvements in system performance. However, as device scaling has slowed and demand to handle big data has soared, the time and energy spent transporting data across the physically separated memory and processing units have started to limit the performance and power efficiency. As potential alternatives, neuro-inspired non-von Neumann computing paradigms have become promising candidates to perform real-world tasks[1, 2]. One avenue of research is referred to as in-memory computing or computational memory, which exploits the physical properties of non-volatile memory (NVM) devices for both storing and processing information[3–6]. Recently, a large-scale experimental demonstration of this concept using an array of one million phase-change memory (PCM) devices has been reported[7]. Another paradigm is hardware acceleration of deep neural network (DNN)[8–12] training via the use of dense crossbar arrays of NVM to perform locally analog computation at the location of the data. As shown in Fig. 1, it is possible to use NVM devices with variable conductance states, such as resistive random access memory (ReRAM)[13] and PCM[14] to represent the synaptic weights and to perform vector-matrix multiplication using the basic electrical principles, i.e., Ohm's and Kirchhoff's laws, thus enabling local and parallel computation on a large scale. By making the conductance change of the NVM element bidirectional, backpropagation algorithm can be implemented. Such a crossbar array of NVMs is expected to achieve significant acceleration factors of DNN training and remarkable reduction in power and area[15, 16]. Another active area of research is spiking neural networks (SNNs) motivated by the need to build more biologically realistic neural network models. Several neuromorphic computing platforms are being developed which are optimized for emulating spike-based computation. These SNNs are typically trained using certain local update rules, such as the spike-timing-dependent plasticity. NVM devices have recently found applications as both synaptic and neuronal elements of such SNNs[17–20].

The key technical challenge for these applications is to realize ideal NVM elements with continuous (analog-like) conductance tuning capability in response to electrical pulses with acceptable noise levels. For acceleration of DNN training, symmetric conductance change with positive and negative pulse amplitudes is another key requirement[15, 16]. The device conductance should go up with a voltage pulse of one polarity and should go down by the same magnitude with a voltage pulse of the opposite polarity. In general, NVM elements do not show this symmetric switching behavior. Therefore, a differential approach is often used in which two conductance values are compared in a unit cell[14]. In this configuration, linearity in switching is required to ensure a symmetric differential signal. In reality, most NVM elements exhibit highly non-linear evolution of conductance as a function of the number of consecutively applied pulses. This results in significant errors in weight updates[13]. In addition, such non-linear conductance change makes separation of signal and noise extremely difficult. Most NVM elements show stochasticity related to the physical origins of switching. When incremental weight updates are performed for analog NVM devices, the magnitude of conductance change approaches the level of inherent randomness[21], manifesting as significant noise components. Therefore, establishing a universally applicable methodology to evaluate signal-to-noise ratio (SNR) of non-linear and analog NVM devices is of paramount importance for neuromorphic computing applications.

In this study, we first establish a practical methodology based on a machine learning algorithm to precisely separate signal and noise components from an analog NVM device with non-linear conductance changes. The methodology is agnostic to the device physics, enabling us to apply it to different types of NVM elements. First, the methodology is applied to HfO_2-based ReRAM to understand the relationship between switching symmetry and SNR. Next, the methodology is applied to PCM devices based on doped-$Ge_2Sb_2Te_5$ (GST). We characterize 1000 devices and separate device-to-device variability and inherent randomness from individual devices.

Results

Analog switching behaviors of ReRAM and PCM. As shown in Fig. 2a, our ReRAM device exhibited analog-like (incremental) change in the device conductance (G) in response to voltage pulses. Consecutive positive voltage (set) pulses (pulse number 1–1000) on the top electrode caused an overall ascending trend of G with some pulse-to-pulse fluctuations. On the other hand, consecutive negative voltage (reset) pulses (pulse number 1001–2000) caused a descending trend of G with similar fluctuations. The change of G in oxide ReRAM device is attributed to change in the configuration of the current conducting filament which consists of oxygen vacancies in a metal oxide film[22, 23] as schematically illustrated in Fig. 2b. The movement of the oxygen vacancies in response to electrical signals has a probabilistic nature and it emerges as inherent randomness in weight updates, which are superimposed on the expected signal[13].

As for PCM, we investigated the device G changes in response to 20 consecutive set pulses. Figure 2c is a plot of G as a function of pulse number, showing incremental changes with a non-linear

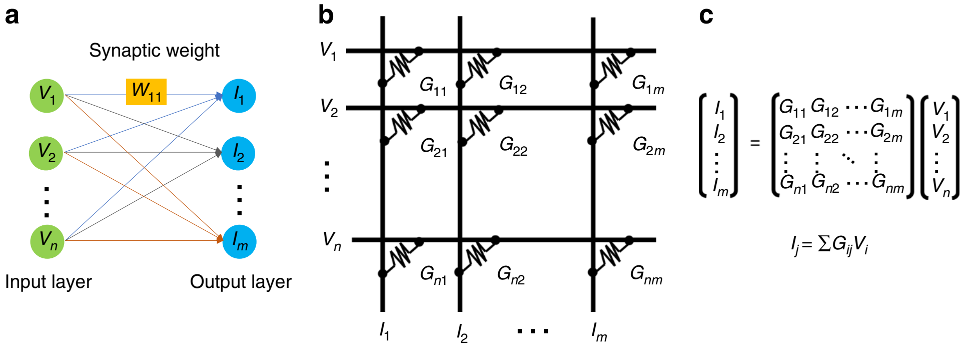
a
Synaptic weight
V_1
W_{11}
I_1
V_2
I_2
V_n
I_m
Input layer
Output layer
b
V_1
V_2
V_n
G_{11}
G_{12}
G_{1m}
G_{21}
G_{22}
G_{2m}
G_{n1}
G_{n2}
G_{nm}
I_1
I_2
I_m
c
$\begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_m \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1m} \\ G_{21} & G_{22} & \cdots & G_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ G_{n1} & G_{n2} & \cdots & G_{nm} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{bmatrix}$
$I_j = \Sigma G_{ij} V_i$

Fig. 1 Neuromorphic computing system based on NVM. **a** Schematic illustration of one-layer neural network with synaptic weights (W) connecting an input layer to an output layer. **b** A synaptic weight is represented by a conductance value of an NVM element at each cross-point in a crossbar array structure. **c** Vector-matrix multiplication is performed by sensing the current (I) for each column, which is the product of the synaptic weight (G) and the input signal (V)

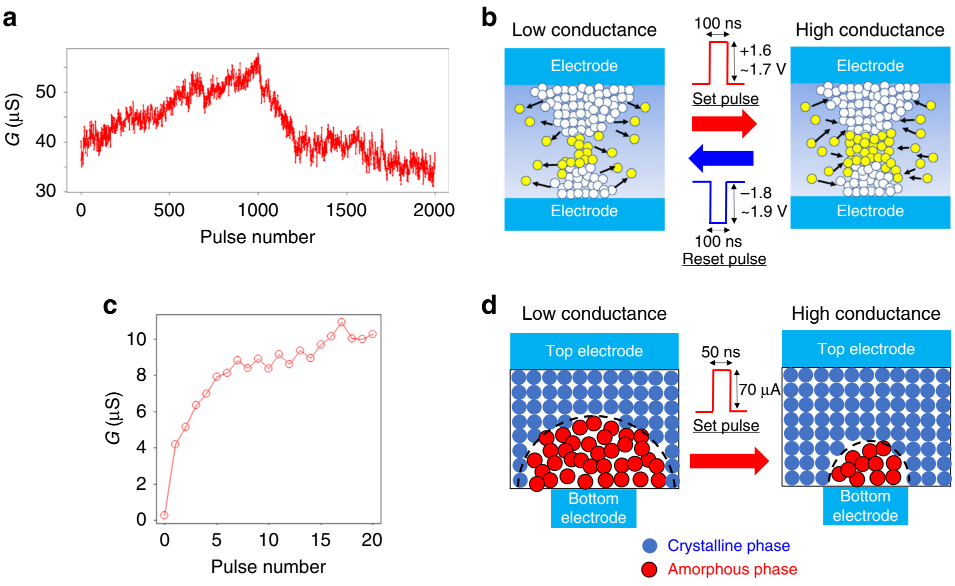

Fig. 2 Analog switching behaviors of ReRAM and PCM. **a** Device *G* as a function of pulse number for our HfO_2-based ReRAM device. 1000 consecutive set pulses, followed by 1000 consecutive reset pulses were applied on the top electrode. **b** The change of *G* is attributed to change in the configuration of the current conducting filament which consists of oxygen vacancies as schematically illustrated. **c** Device conductance *G* as a function of pulse number for our GST-based PCM device. A total of 20 consecutive set pulses were applied. **d** The *G* increase in response to set pulses corresponds to the transition from the amorphous phase to the crystalline phase of GST as schematically illustrated

trace, which is convoluted with pulse-to-pulse fluctuations. The PCM device includes a small part of phase-change material that is sandwiched by top and bottom electrodes. Transition from the low conductance state (amorphous phase) to the high conductance state (crystalline phase) is caused by set pulses that create sufficient joule heating for crystallization of the GST material while the temperature is kept below the melting point as schematically illustrated in Fig. 2d. Due to the stochastic nature in crystallization of the phase-change materials[2, 20, 21, 24, 25], there is significant randomness associated with the weight updates. On the other hand, reset to the low conductance state requires melting of the GST material and this process is known to be abrupt. For the purpose of characterization of analog switching behaviors, we focused on incremental set operations for PCM in this study.

Characterization of NVM elements. To evaluate the performance of analog NVM elements for neuromorphic computing applications, one has to extract noise-free signals from experimental data. A conventional approach is to assume a parametric model for expected conductance changes, derived from relatively simple assumptions on underlying physics. For ReRAM devices, an exponential formula has been proposed to capture the nonlinear trend[13]. However, the pre-assumed exponential relationship often causes significant errors when fitting weight update as a function of number of applied pulses. In addition, different NVM elements generally need different fitting formulas, making it difficult to compare key performance parameters, such as switching symmetry and SNR, on a common ground. To address this issue, we leverage a machine learning algorithm called Gaussian process regression (GPR)[26]. GPR is a non-parametric Bayesian regression method, which does not assume any specific functional form such as linear and exponential. The main motivation for implementing GPR in the analysis of analog NVM elements is to let experimental data give predictions of noise-free signals by themselves. The major assumption we used is the smoothness of the curve. For analog NVM devices, we exploit continuous changes in switching media (e.g., filament configuration for ReRAM, volume of crystalline region for PCM) rather than non-continuous phenomena to achieve incremental conductance changes. This makes analog switching data highly compatible with the assumption of smoothness. The key ingredient of GPR is the kernel matrix (Eq. (6) in Methods), which controls the smoothness of the estimated functional curve. We established a practical approach to optimize the kernel matrix by combining the Bayesian marginalized likelihood maximization with the frequentists' cross-validation approach. This enabled us to precisely separate signal and noise for our large dataset while avoiding numerical instability. The proposed inference procedure also assumes that a prior probability distribution over underlying functions follows a multivariate Gaussian distribution, which consists of a linear combination of finite random variables. This assumption is consistent with the switching mechanism of analog memory devices where the device conductance is governed by parallel configurations of randomly distributed conducting filaments comprising oxygen vacancies or crystalline phase-change materials. The measured device conductance values indeed follow a Gaussian distribution around noise-free signals and this was verified by observing the distribution of noise in our experimental data for ReRAM (Supplementary Note 1). The details of our GPR-based methodology are described in Methods section.

We performed cross-validation[27] using our ReRAM data and confirmed that the GPR-based methodology extracted the inherent features irrespective of the sampling size (Supplementary Note 2). We confirmed the robustness of our methodology against the variation of duration of input pulses from 5 to 100 ns, covering the range of interest for neuromorphic computing (Supplementary Note 3). We also confirmed the robustness of our methodology against the variation of test temperature (Supplementary Note 4). For the rest of the analysis, we used a pulse duration of 100 ns and tested the devices at room temperature. Next, we extracted key performance metrics using the GPR fitting. We applied the methodology to our ReRAM data with 1000 consecutive set pulses, followed by 1000 consecutive reset pulses, for the purpose of characterizing switching symmetry. As shown in Fig. 3a, the GPR fitting gave predicted noise-free curves

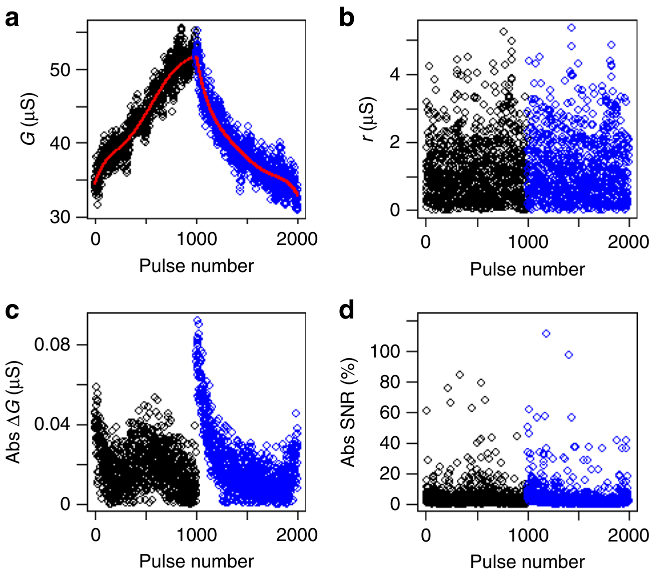

Fig. 3 Extraction of key metrics for neuromorphic computing. The GPR-based methodology is applied to our ReRAM data with 1000 consecutive set pulses, followed by 1000 consecutive reset pulses. **a** Device G as a function of pulse number for set (black) and reset (blue) pulse sequences. The predicted noise-free signals are shown in red lines. **b** Extracted r values as a function of pulse number for set (black) and reset (blue) pulse sequences. **c** Extracted ΔG values as a function of pulse number for set (black) and reset (blue) pulse sequences. **d** Extracted SNR values as a function of pulse number for set (black) and reset (blue) pulse sequences

(red lines) for both set (black) and reset (blue) pulse sequences. Once the noise-free curves are estimated, the G change per pulse, denoted by ΔG, is easily computed, based on which we define SNR as

$$\mathrm{SNR} \underline{\underline{\mathrm{def}}} \frac{\Delta G}{r}, \tag{1}$$

where r represents the absolute difference between predicted and observed G values (i.e., residuals). The impact of SNR on the accuracy of neural network was previously discussed[21]. Since relatively long sequences were used for both ReRAM and PCM devices to minimize fluctuations in read signals, we attribute r to inherent randomness associated with the physical origin of weight update. In artificial neural network implementations, fast reading is particularly preferred to decrease the overall cycle time and consequently accelerate the computational operations. This should increase the contribution of read noise. In this case, we need to optimize the read operation to balance the overall performance and the noise level, which is beyond the scope of this work. The extracted r value is shown as a function of pulse number in Fig. 3b. The absolute ΔG values for set and reset pulses are denoted by ΔG_+ and ΔG_-, respectively. The ΔG_+ (black) and ΔG_- (blue) are plotted as a function of pulse number in Fig. 3c. Figure 3d shows absolute SNR, calculated locally at each pulse from ΔG and r. For characterization of switching symmetry, we introduce symmetry factor (SF), which is defined as

$$\mathrm{SF} \underline{\underline{\mathrm{def}}} \frac{\Delta G_+ - \Delta G_-}{\Delta G_+ + \Delta G_-}. \tag{2}$$

With this definition, the degree of symmetry is quantified as a value between −1 and 1, with 0 corresponding to the perfect symmetry. Asymmetry in both directions (larger ΔG_+ versus ΔG_-) are equally weighted around 0 and can be compared with absolute values. In order to compute SF and SNR at a given G level, we need to express ΔG_+, ΔG_-, and r as functions of G. Therefore, we divided the total G range into 100 sub-ranges and computed a mean value of ΔG and a root mean square value of r within each G sub-range. In this way, one can obtain SF and SNR for each G sub-range. The local extraction (i.e., at a certain pulse number or G level) of SF and SNR is a powerful feature of our methodology. The symmetry requirement for acceleration of DNN training specified in ref. [15] (<5% difference between ΔG_+ and ΔG_-) corresponds to |SF| <0.025.

Switching symmetry and SNR of ReRAM devices. We applied the GPR-based methodology on our ReRAM devices with different metal oxide thicknesses (device A: 5 nm, device B: 4 nm). The devices were tested under different set and reset voltages and the SNR and SF values were extracted locally at each G level, as shown in Fig. 4a. For SNR, we took mean values for set and reset traces. Representative G versus pulse number traces are shown in insets. Figure 4b shows a cross-sectional two-dimensional plot of SNR versus SF taken at G ~20 μs from Fig. 4a. At this G level, low |SF| values were achieved at relatively low SNR values, and vice versa. Data points are absent in the upper-left corner of Fig. 4b, indicating that there is a fundamental tradeoff between SNR and SF values. In order to investigate the relationship between SNR and SF values for multiple device/pulse conditions spanning different G levels, they were grouped according to SNR values and cumulative distribution function of |SF| were compared, as shown in Fig. 4c. The reproducibility of the trend was confirmed up to 10 different devices of device type B (Supplementary Note 5). One can clearly observe that the device/pulse conditions that lead to higher SNR values tend to result in poor switching symmetry. The tradeoff can be directly observed in the G versus pulse number plots (the insets of Fig. 4a). We speculate that higher switching symmetry is achieved by making the movement of oxygen vacancies more incremental and thereby changing the width of current conducting filament rather than completely rupturing and reforming it. ΔG is smaller for the former case and it should eventually approach the level of inherent randomness, resulting in lower SNR values. Such a tradeoff makes it difficult to improve both switching symmetry and SNR at the same time and it remains as a key challenge for ReRAM devices for neuromorphic computing applications. However, if these key metrics are accurately quantified like we demonstrated with our GPR-based methodology, one can optimize the device and pulse conditions to find the optimum point within the tradeoff. As reviewed in a previous section, switching symmetry is a critical requirement to implement backpropagation algorithm for DNNs. In reality, learning accuracy is compromised due to non-ideal (asymmetric) switching characteristics of synaptic elements. Therefore, we optimized the device condition (device A) and the pulse condition (set: 1.6 V, reset: −1.8 V) using the GPR-based methodology to minimize SF. The beauty of our methodology is the capability to extract SF, agnostic to switching mechanisms and irrespective of data size. This enabled us to compare our ReRAM data with various resistive switching devices in literature[28–35]. There have been reports on improved switching symmetry using pulses with varying amplitude[28, 30, 31]. These cases were benchmarked together and marked separately in Fig. 4d. One can see a general trend of improved symmetry using pulses with varying amplitudes. This approach, however, requires sensing of current states of individual devices and adjustment of voltage amplitudes, which is not compatible with local and parallel computation. It should be noted that our optimized ReRAM data showed good switching symmetry compared with all

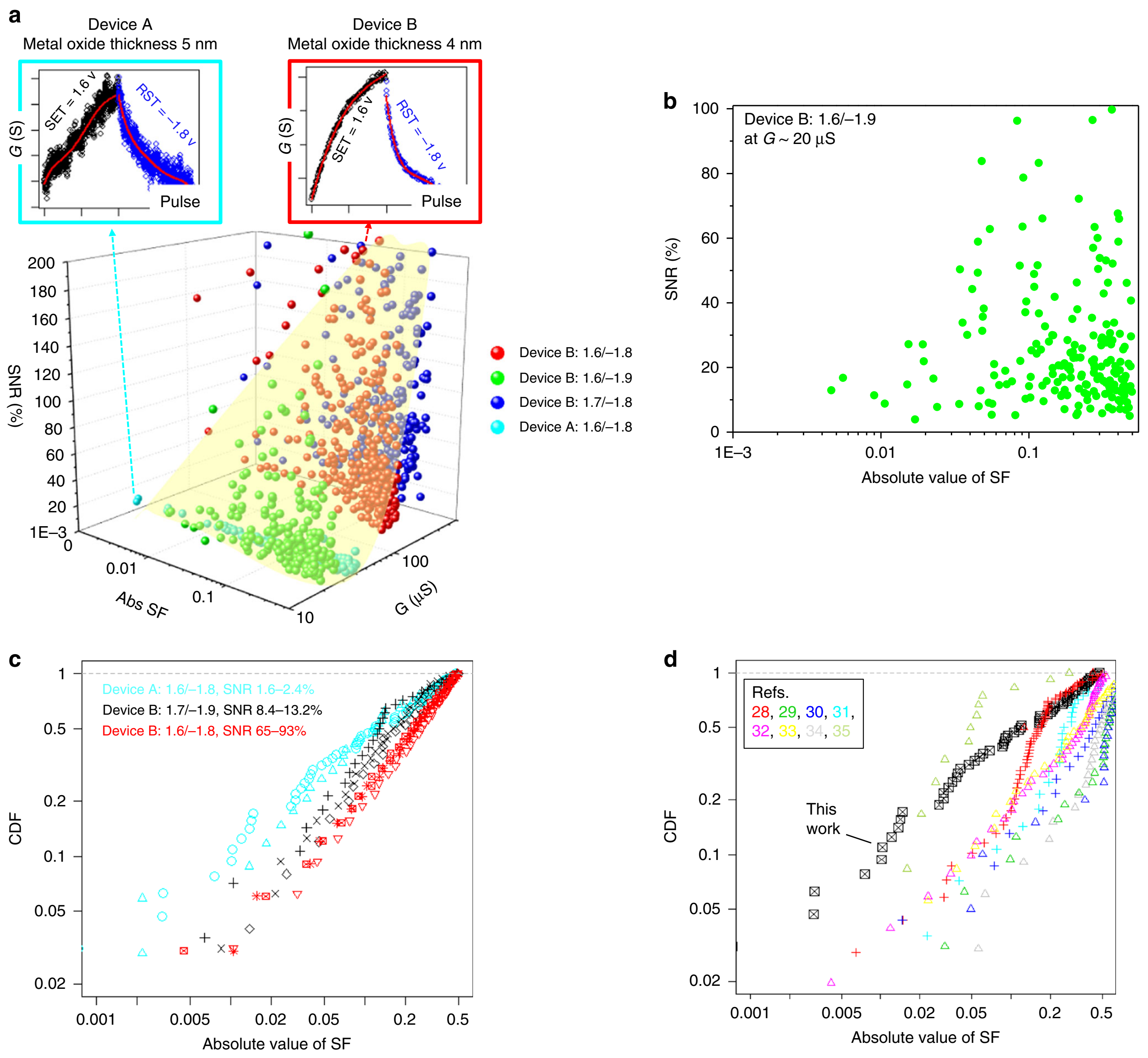

Fig. 4 Tradeoff between switching symmetry and SNR for ReRAM. The GPR-based methodology was applied to our ReRAM data with different metal oxide thicknesses (device A: 5 nm, device B: 4 nm) and different set (1.6-1.7 V) and reset (−1.8 to −1.9 V) voltages to investigate the relationships between the key metrics. **a** SNR and SF as functions of *G*. The original data for two representative device and pulse conditions are shown in the insets to contrast the tradeoff. **b** Cross-sectional 2D plot of SNR versus SF taken at *G* ~20 μs from **a**. **c** Plot of cumulative distribution function (CDF) of absolute values of SF for data in different SNR ranges. **d** Benchmark plot of SF. The optimized ReRAM data (device A, set: 1.6 V, reset: −1.8 V) were compared with refs. [28-35]. The benchmark data points are color-coded (see the legend) and shown as symbols for same amplitude pulses and as crosses for varying amplitude pulses

benchmark data with identical voltage pulses. This is a significant step forward to realize online training capability in a parallel manner. Future work needs to focus on simultaneously achieving sufficiently high SNR values with materials optimizations.

Breakdown of variability components in 90-nm PCM devices. A conventional approach to extract inherent randomness associated with weight updates is to test multiple devices and to obtain statistical distributions[21]. The variability obtained in this manner, however, includes device-to-device variability in addition to inherent randomness from individual devices. These variability components need to be quantified separately in order to accurately assess potentials of certain NVM elements for neuromorphic computing applications. We tested 1000 PCM devices and extracted signal and noise from individual devices using our GPR-based methodology. This enabled us to further break down the total variability to the inherent randomness of individual devices and the device-to-device variability. These two variability components are illustrated in Fig. 5a with two representative PCM devices (devices 1 and 2) that were fabricated with the identical process. The GPR fitting was performed to predict noise-free signals as shown in red and blue solid lines, respectively, in Fig. 5a. The predicted signals for devices 1 and 2 deviate from each other due to device-to-device variability. In addition, the experimental data points (shown in circles) fluctuate around the individual fitted lines, which is attributable to inherent randomness of weight updates since the read noise was minimized by

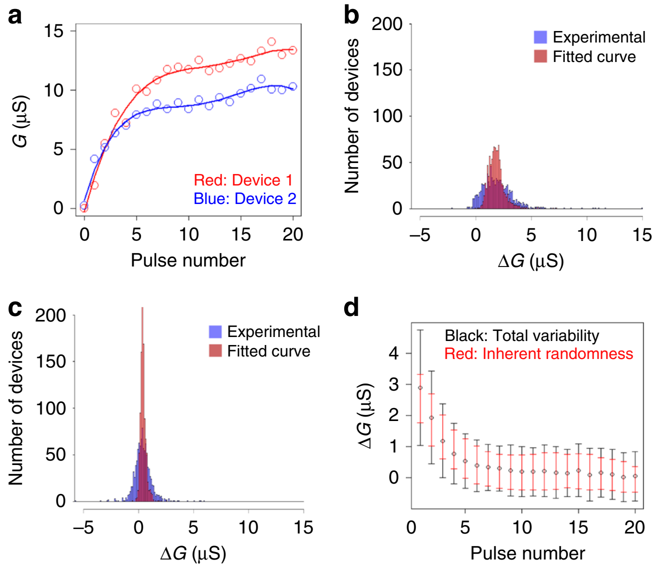

Fig. 5 Separation of device-to-device variability and inherent randomness for PCM. **a** Device G as a function of pulse number for two representative PCM devices (devices 1 and 2), fabricated with the identical process. The predicted noise-free signals from the GPR fitting are shown in red and blue solid lines. The difference of two fitted lines corresponds to device-to-device variability, whereas the fluctuations of the experimental data around the fitted lines are attributed to inherent randomness from individual devices. Histograms of ΔG values extracted from experimental data (blue) and fitted curves (red) (**b**) after the second pulse and (**c**) after the sixth pulse. **d** The mean and standard deviation of ΔG obtained from the experimental data (shown in black circles and error bars) were compared with the root mean square of inherent randomness (r) obtained from the GPR-based methodology (shown in red error bars) as a function of pulse number. The inherent randomness accounts for 67% of the total variability

the test sequence as described in Methods section. We compared the histograms of ΔG values extracted from experimental data and fitted curves after the pulse numbers 2 (Fig. 5b) and 6 (Fig. 5c). The statistical distribution of the fitted curves (red) is the contribution from device-to-device variability, whereas the statistical distribution of the experimental data (blue) includes inherent randomness superimposed on top of that. The latter distribution was much wider, clearly showing significant contribution of inherent randomness. The peak ΔG value decreased and the device-to-device variability (red) tightened from the second to the sixth pulse. On the other hand, the inherent randomness remained relatively constant. This resulted in the tail of total distribution (blue) extending into the negative ΔG regime, which is undesirable (Fig. 5c). The mean and standard deviation of ΔG obtained from the experimental data (shown in black circles and error bars) were compared with the root mean square of inherent randomness (r) obtained from the GPR-based methodology (shown in red error bars) as a function of pulse number in Fig. 5d. The total standard deviation became comparable with ΔG for incremental weight updates. Since the learning accuracy is known to degrade when the ratio of standard deviation to ΔG becomes >1[21], reduction of variability is indispensable. Our analysis revealed that a large portion of total variability is attributed to inherent randomness of individual devices (~67%) for a mature technology based on the 90 nm CMOS baseline. The median SNR value calculated from inherent randomness is ~35% for PCM devices, which is comparable to our ReRAM device switching at a similar G level (cf. Fig. 4b). This indicates that variability due to inherent randomness is a common challenge for ReRAM and PCM for neuromorphic computing applications. Innovations in device and material are needed to suppress this component. Our methodology based on GPR enables precise extraction of inherent randomness from individual devices and provides useful guidelines for further improvement.

Discussion

We established a practical methodology based on GPR to precisely separate signal and noise components from analog NVM elements with non-linear conductance changes. This solves key technical challenges for characterization of artificial synapses of neuromorphic computing system, namely extraction of switching symmetry and SNR. The methodology is agnostic to switching mechanisms and therefore applicable to various types of NVMs. We applied the methodology to HfO_2-based ReRAM devices and found the tradeoff between switching symmetry and SNR. Using SF as a guideline, substantial improvement in switching symmetry was achieved compared to reported ReRAM devices in literature. By systematic analysis of 1000 GST-based PCM devices, we clearly demonstrated that a large portion of variability in weight update is attributable to inherent randomness from individual devices and this is the key component to be suppressed in order to achieve high classification accuracy.

Finally, the proposed methodology helps neuromorphic system engineers in two ways depending on phases of technology development. In an exploratory phase, our methodology enables extraction of switching symmetry and SNR from individual devices and expedites search for ideal materials. The conventional methodology requires fabrication of many devices with tight device-to-device variability for extraction of SNR, which is difficult to attain in the early stage when exotic material options need to be screened. In a relatively mature technology phase, our methodology helps find the optimum input signals (e.g., duration and amplitude of pulses) that provide the best switching symmetry (linearity) and SNR within the tradeoff for the entire neuromorphic system.

Methods

ReRAM device fabrication and test. We fabricated 2-terminal oxide ReRAM with device dimensions of $50 \times 50\ \mu m^2$. First, a SiO_2 underlayer was grown on a 200 mm Si wafer. Then, a 100 nm-thick TiN film was deposited by reactive sputtering as a bottom electrode, followed by deposition of a HfO_2 layer by atomic layer deposition as a switching layer where a current conducting filament is formed. We varied the thickness of the switching layer (device A: 5 nm, device B: 4 nm) to investigate its impact on switching symmetry. Next, a 20 nm-thick TiN was deposited by reactive sputtering as a top electrode. The device area was defined by photolithography and reactive ion etching of the TiN electrode. To test switching symmetry and SNR of our ReRAM devices, we applied a sequence of weight update (write) pulses with the same voltage amplitude for each polarity. We used high-resolution source measure unit (SMU) to read the device conductance state between the write pulses. We applied a small read voltage of 0.1 V to prevent disturbance in the resistance state. While keeping the read voltage applied across the device, we took multiple read steps with a 16.67 ms integration time until the measured values read at the instrument stabilized (typically within 3–10 repetitive read measurements in the device resistance range of interest). Then, we chose the last measurement as the representative value. We did not detect random telegraph noise with this read sequence. The write pulses had duration of 100 ns (unless otherwise mentioned) and various voltage amplitudes (set pulse: 1.6–1.7 V; reset pulse: −1.8 to −1.9 V) were compared to investigate the impacts on switching symmetry and SNR. In order to separate noises from weight update and those from weight read, we also carried out read-only test, where only read steps were repeated up to 1000 times without weight updates in between. Our linear regression analysis showed that the residual standard error of read-only trace is 2.51×10^{-7} S, which is almost one order lower than that of read-after-write trace ($1.38–1.57 \times 10^{-6}$ S). Therefore, we attribute a majority of noise components of our ReRAM devices to inherent randomness in weight updates.

PCM device fabrication and test. The PCM devices were integrated into a chip fabricated in the 90 nm CMOS technology[36]. The phase-change material is doped $Ge_2Sb_2Te_5$. The bottom electrode has a radius of ~20 nm and was defined using a sub-lithographic key-hole transfer process[37]. The phase-change material is ~100 nm-thick and extends to the top electrode. All experiments in this work were done on an array comprising 1 million devices, which is organized as a matrix of 512 word lines (WLs) and 2048 bit lines (BLs). The selection of one PCM device is done by serially addressing a WL and a BL. A single selected device can be programmed by forcing a current through the BL with a voltage-controlled current source. For reading a PCM cell, the selected BL is biased to a constant voltage of 0.3 V. The resulting read current is integrated by a capacitor, and the resulting voltage is then digitized by an on-chip 8-bit cyclic ADC. The ADCs are calibrated by means of on-chip reference poly-silicon resistors. As for characterization of incremental device G change, each device was first initialized to a state that has almost zero conductance. After the initialization, a set pulse of 70 µA was applied followed by conductance read steps. The read step was repeated 50 times to obtain mean G values in order to minimize read noise and to focus on characterization of write noise. This sequence was repeated 20 times to obtain G values as a function of pulse numbers.

GPR-based methodology. The goal of GPR is to learn a probability distribution of the output signal, y, conditioned on the input signal, x, from data $\left\{\left(x^{(n)}, y^{(n)}\right) | n = 1, \ldots, N\right\}$, where N is the number of samples and the superscript (n) denotes the n-th sample in the data. The distribution is given by

$$p(y|x) = \mathcal{N}\left(y|m(x), s^2(x)\right), \tag{3}$$

$$m(x) = \mathbf{k}^{\mathrm{T}}(\mathbf{K}+\mathbf{I})^{-1}\mathbf{y}_N, \tag{4}$$

$$s^2(x) = \sigma^2\left[2 - \mathbf{k}^{\mathrm{T}}(\mathbf{K}+\mathbf{I})^{-1}\mathbf{k}\right], \tag{5}$$

where $\mathcal{N}\left(y|m(x), s^2(x)\right)$ denotes the Gaussian distribution of y with the mean $m(x)$ and the variance $s^2(x)$. Also, σ^2 denotes the variance that corresponds to measurement noise, $\mathbf{I}$ denotes the identity matrix, and $\mathbf{y}_N = \left(y^{(1)}, \ldots, y^{(N)}\right)^{\mathrm{T}}$, where the superscript T denotes the matrix transpose.

The key ingredient of GPR is the kernel matrix $\mathbf{K}$, which controls the smoothness of the estimated functional curve. We use a non-dimensional kernel $\mathbf{K}$ whose (i, j) element is given by

$$K\left(x^{(i)}, x^{(j)}\right) \overset{\text{def}}{=} \exp\left(-\frac{|x^{(i)} - x^{(j)}|^2}{2\sigma_K^2}\right). \tag{6}$$

The n-th entry of the N-dimensional vector $\mathbf{k}(x)$ is also given by $K\left(x, x^{(n)}\right)$. The parameters σ_K^2, σ^2 are learned from the data, as explained later. The idea is to use the predictive mean, $m(x)$, at the input value (pulse number) x, as a noise-free version of the output signal (G).

Determining GPR parameters. The parameter σ^2 is determined by maximizing the log marginalized likelihood[26], which is given by

$$E(\sigma) \overset{\text{def}}{=} -\frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\mathbf{y}_N^{\top}(\mathbf{K}+\mathbf{I})^{-1}\mathbf{y}_N - \frac{1}{2}\ln\det(\mathbf{K}+\mathbf{I}) + c, \tag{7}$$

in our parameterization, where c denotes an unimportant constant, and det is the matrix determinant. Assuming σ_K is given for now and taking the derivative with respect to σ^{-2}, we have

$$\sigma^2 = \left(\frac{1}{N}\right)\mathbf{y}_N^{\mathrm{T}}(\mathbf{K}+\mathbf{I})^{-1}\mathbf{y}_N. \tag{8}$$

To compute this, we need a value of σ_K. In theory, we could find it by maximizing E simultaneously with σ. This approach, however, involves a complex non-linear optimization procedure and often results in numerical instability in our application.

Here we propose a practical approach that combines the Bayesian marginalized likelihood maximization with the frequentists' cross-validation approach. Specifically, to determine σ_K, we maximize the predictive leave-one-out (LOO) likelihood, as defined by

$$L(\sigma_K) \overset{\text{def}}{=} \sum_{i=1}^{N}\ln\mathcal{N}\left(y^{(i)}|m_{-i}\left(x^{(i)}\right), s_{-i}^2\left(x^{(i)}\right)\right), \tag{9}$$

where m_{-i} and s_{-i}^2 are the predictive mean and variance of GPR (Eqs. (4) and (5)) obtained from the dataset excluding the i-th sample. To find the maximizer of $L(\sigma_K)$, we can leverage the fact that the observed variance does not depend heavily on the input across the entire domain. By replacing s_{-i}^2 with a constant, the LOO likelihood criterion is reduced to the task of finding a minimizer of the mean square of the residual (i.e., r), which is easily done independently of σ^2. In this study, we use the following procedure and criterion to find an appropriate σ_K value from the experimental data. We vary σ_K to cover a wide range and identify an optimum range where the change of σ_K negligibly affects extracted r values. This is practically equivalent to maximizing the predictive LOO likelihood. Our criterion is r change of <1% for σ_K change of 10% and this is met with a σ_K value of around $3 \times N$ for our dataset (Supplementary Note 6).

Data availability. The data that support the findings of this study are available from the corresponding author upon request.

Received: 27 November 2017 Accepted: 1 May 2018
Published online: 29 May 2018

References

1. Merolla, P. A. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **345**, 668–673 (2014).
2. Burr, G. W. et al. Neuromorphic computing using non-volatile memory. *Adv. Phys. X* **2**, 89–124 (2016).
3. Gallo, M. L. et al. Mixed-precision in-memory computing. *Nat. Electron.* **1**, 246–253 (2018).
4. Sheridan, P. M. et al. Sparse coding with memristor networks. *Nat. Nanotechnol.* **12**, 784–789 (2017).
5. Wright, C. D., Hosseini, P. & Diosdado, J. A. V. Beyond von-Neumann computing with nanoscale phase-change memory devices. *Adv. Funct. Mater.* **23**, 2248–2254 (2012).
6. Hosseini, P., Sebastian, A., Papandreou, N., Wright, C. D. & Bhaskaran, H. Accumulation-based computing using phase-change memories with FET access devices. *IEEE Electron Device Lett.* **36**, 975–977 (2015).
7. Sebastian, A. et al. Temporal correlation detection using computational phase-change memory. *Nat. Commun.* **8**, 1115 (2017).
8. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
9. Collobert, R. & Weston, J. A unified architecture for natural language processing. In *Proc. 25th International Conference on Machine Learning - ICML 08* (ACM, Helsinki, Finland, 2008).
10. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* **115**, 211–252 (2015).
11. Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* **29**, 82–97 (2012).
12. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
13. Chen, P.-Y. et al. Mitigating effects of non-ideal synaptic device characteristics for on-chip learning. In *2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* https://doi.org/10.1109/iccad.2015.7372570 (Publisher IEEE, Austin, USA, 2015).
14. Burr, G. W. et al. Experimental demonstration and tolerancing of a large-scale neural network (165000 synapses) using phase-change memory as the synaptic weight element. *IEEE Trans. Electron Devices* **62**, 3498–3507 (2015).
15. Gokmen, T. & Vlasov, Y. Acceleration of deep neural network training with resistive cross-point devices: design considerations. *Front. Neurosci.* **10**, 333 (2016).
16. Agarwal, S. et al. Resistive memory device requirements for a neural algorithm accelerator. In *2016 International Joint Conference on Neural Networks (IJCNN)* https://doi.org/10.1109/ijcnn.2016.7727298 (Publisher IEEE, Vancouver, Canada, 2016).
17. Kuzum, D., Jeyasingh, R. G. D., Lee, B. & Wong, H.-S. P. Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing. *Nano Lett.* **12**, 2179–2186 (2011).
18. Kim, S. et al. NVM neuromorphic core with 64k-cell (256-by-256) phase change memory synaptic array with on-chip neuron circuits for continuous in-situ learning. In *2015 IEEE International Electron Devices Meeting (IEDM)* https://doi.org/10.1109/iedm.2015.7409716 (Publisher IEEE, Washington DC, USA, 2015).
19. Saïghi, S. et al. Plasticity in memristive devices for spiking neural networks. *Front. Neurosci.* **9**, 51 (2015).
20. Tuma, T., Pantazi, A., Gallo, M. L., Sebastian, A. & Eleftheriou, E. Stochastic phase-change neurons. *Nat. Nanotechnol.* **11**, 693–699 (2016).
21. Boybat, I. et al. Stochastic weight updates in phase-change memory-based synapses and their influence on artificial neural networks. In *2017 13th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)* https://doi.org/10.1109/prime.2017.7974095 (Publisher IEEE, Giardini Naxos, Italy, 2017).
22. Miranda, E., Jimenez, D. & Sune, J. The quantum point-contact memristor. *IEEE Electron Device Lett.* **33**, 1474–1476 (2012).

23. Ielmini, D. Modeling the universal set/reset characteristics of bipolar RRAM by field- and temperature-driven filament growth. *IEEE Trans. Electron Devices* **58**, 4309–4317 (2011).
24. Wong, H.-S. P. et al. Phase change memory. *Proc. IEEE* **98**, 2201–2227 (2010).
25. Gallo, M. L., Tuma, T., Zipoli, F., Sebastian, A. & Eleftheriou, E. Inherent stochasticity in phase-change memory devices. In *2016 46th European Solid-State Device Research Conference (ESSDERC)* https://doi.org/10.1109/essderc.2016.7599664 (Publisher IEEE, Lausanne, Switzerland, 2016).
26. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, United States 2008).
27. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R* (Springer, New York, United States 2017).
28. Jang, J.-W., Park, S., Burr, G. W., Hwang, H. & Jeong, Y.-H. Optimization of conductance change in $Pr_{1-x}Ca_xMnO_3$-based synaptic devices for neuromorphic systems. *IEEE Electron Device Lett.* **36**, 457–459 (2015).
29. Jo, S. H. et al. Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.* **10**, 1297–1301 (2010).
30. Wang, I.-T., Chang, C.-C., Chiu, L.-W., Chou, T. & Hou, T.-H. 3D Ta/TaOx/TiO2/Ti synaptic array and linearity tuning of weight update for hardware neural network applications. *Nanotechnology* **27**, 365204 (2016).
31. Chen, W. et al. A CMOS-compatible electronic synapse device based on Cu/SiO2/W programmable metallization cells. *Nanotechnology* **27**, 255202 (2016).
32. Yao, P. et al. Face classification using electronic synapses. *Nat. Commun.* **8**, 15199 (2017).
33. Marinella, M. J. et al. Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator. Preprint at http://arxiv.org/abs/1707.09952 (2017).
34. Wu, W. et al. Improving analog switching in HfOx-based resistive memory with a thermal enhanced layer. *IEEE Electron Device Lett.* **38**, 1019–1022 (2017).
35. Woo, J. et al. Improved synaptic behavior under identical pulses using AlO_x/HfO_2 bilayer RRAM array for neuromorphic systems. *IEEE Electron Device Lett.* **37**, 994–997 (2016).
36. Close, G. F. et al. Device, circuit and system-level analysis of noise in multi-bit phase-change memory. In *2010 International Electron Devices Meeting* https://doi.org/10.1109/iedm.2010.5703445 (Publisher IEEE, San Francisco, USA, 2010).
37. Breitwisch, M. et al. Novel lithography-independent pore phase change memory. In *2007 IEEE Symposium on VLSI Technology* https://doi.org/10.1109/vlsit.2007.4339743 (Publisher IEEE, Kyoto, Japan, 2007).

Acknowledgements

We would like to thank Marwan Khater, Hiroyuki Miyazoe, Adam Pyzyna, and the staff of Microelectronics Research Laboratory at IBM T.J. Watson Research Center for their contributions in device fabrication. We would also like to thank Wilfried Haensch for management support and valuable discussions.

Author contributions

T.A. conceived the idea. N.G. and T.A. performed the experiments and analyzed all data. T.I., N.G., and T.A. developed the GPR-based methodology. S.K. performed the experiments on ReRAM and analyzed the data. I.B. and A.S. performed the experiments on PCM and analyzed the data. V.N. provided managerial support and critical comments. N.G. and T.A. wrote the manuscript with input from all the authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04485-1.

Competing interests: The authors declare no competing interests.