

# Nonideality-Aware Training for Accurate and Robust Low-Power Memristive Neural Networks

Dovydas Joksas,\* Erwei Wang, Nikolaos Barmapsalos, Wing H. Ng, Anthony J. Kenyon, George A. Constantinides, and Adnan Mehonic\*

Recent years have seen a rapid rise of artificial neural networks being employed in a number of cognitive tasks. The ever-increasing computing requirements of these structures have contributed to a desire for novel technologies and paradigms, including memristor-based hardware accelerators. Solutions based on memristive crossbars and analog data processing promise to improve the overall energy efficiency. However, memristor nonidealities can lead to the degradation of neural network accuracy, while the attempts to mitigate these negative effects often introduce design trade-offs, such as those between power and reliability. In this work, authors design nonideality-aware training of memristor-based neural networks capable of dealing with the most common device nonidealities. The feasibility of using high-resistance devices that exhibit high  $I$ - $V$  nonlinearity is demonstrated—by analyzing experimental data and employing nonideality-aware training, it is estimated that the energy efficiency of memristive vector-matrix multipliers is improved by almost three orders of magnitude ( $0.715 \text{ TOPs}^{-1}\text{W}^{-1}$  to  $381 \text{ TOPs}^{-1}\text{W}^{-1}$ ) while maintaining similar accuracy. It is shown that associating the parameters of neural networks with individual memristors allows to bias these devices toward less conductive states through regularization of the corresponding optimization problem, while modifying the validation procedure leads to more reliable estimates of performance. The authors demonstrate the universality and robustness of this approach when dealing with a wide range of nonidealities.

## 1. Introduction

Artificial neural networks (ANNs) are now routinely used in machine learning tasks ranging from text generation<sup>[1]</sup> to autonomous driving.<sup>[2]</sup> However, rapidly increasing number of parameters in modern ANNs is making them time- and power-consuming during both training<sup>[3]</sup> and inference<sup>[4]</sup> stages. This makes it challenging to apply machine learning approaches in environments where resources are tightly constrained.<sup>[5,6]</sup>

One of the proposed solutions to improve the efficiency of ANNs has been to adopt different computer architectures. The need to transfer data between memory and computing units in the von Neumann architecture is the main bottleneck in modern computers<sup>[7]</sup>; this is especially evident in machine learning where large amounts of data are utilized. In this specific case, an alternative can be memristor-based ANNs, or memristive neural networks (MNNs). With this approach, memristive crossbar arrays are used to physically compute vector-matrix products, which are one of the most fundamental operations in ANNs.<sup>[8,9]</sup> This is done without the need to constantly move

large amounts of data: matrix entries are encoded into memristor conductances, vector entries—into applied voltages, and the result of the operation is extracted from the output currents produced according to Ohm's law and Kirchhoff's current law.<sup>[8]</sup>

Memristors—when used and programmed as analog devices—can encode values at much higher density, but at a cost of lower precision. A number of nonidealities may prevent from accurately programming device conductances or may cause deviations from intended electrical behavior. Such nonidealities include stuck devices, device-to-device (D2D) and cycle-to-cycle variability, drift in resistance states, line resistance,  $I$ - $V$  nonlinearity, and others.<sup>[10]</sup>

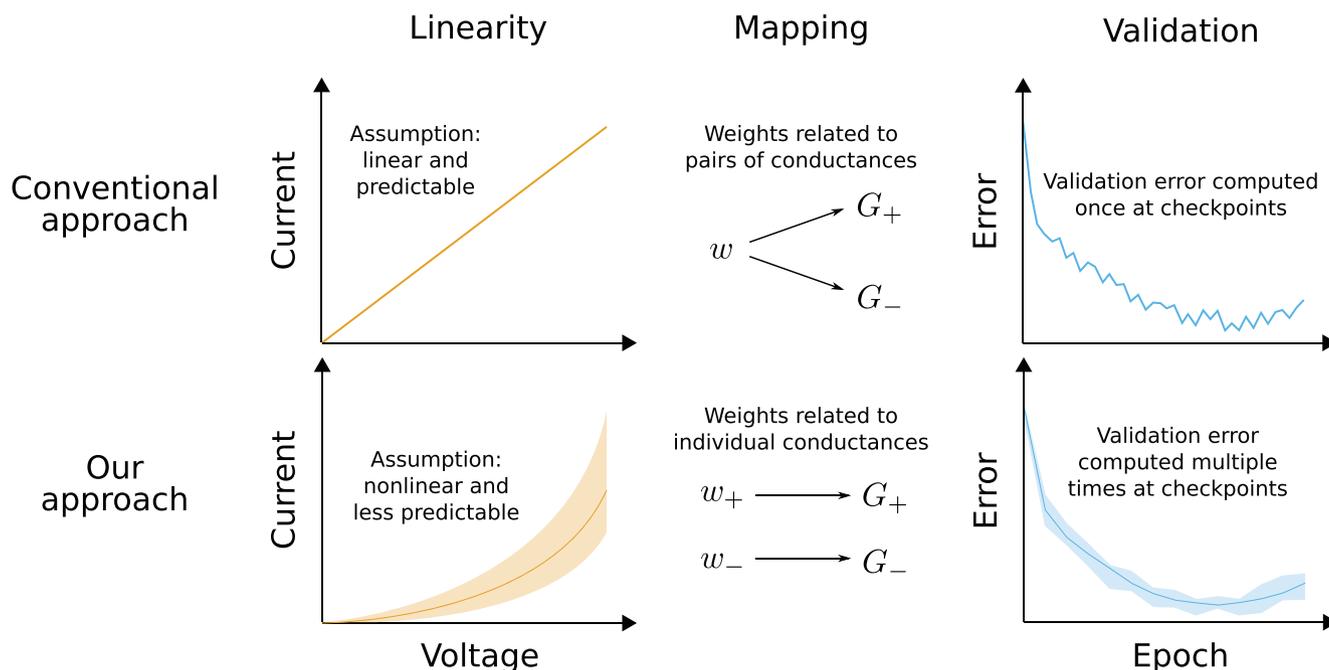
Potential solutions do exist but many of them introduce a number of trade-offs. For example, to ensure more linear  $I$ - $V$  characteristics, one may use low-resistance devices<sup>[11]</sup>; however, this results in higher power consumption. Alternatively, the effects of  $I$ - $V$  nonlinearities may be minimized by utilizing pulse-width modulation,<sup>[12]</sup> but that comes at a cost of increased clock cycles for each encoded input.<sup>[13]</sup>

D. Joksas, N. Barmapsalos, W. H. Ng, A. J. Kenyon, A. Mehonic  
 Department of Electronic and Electrical Engineering  
 University College London  
 London WC1E 7JE, UK  
 E-mail: dovydas.joksas.15@ucl.ac.uk; adnan.mehonic.09@ucl.ac.uk  
 E. Wang, G. A. Constantinides  
 Department of Electrical and Electronic Engineering  
 Imperial College London  
 London SW7 2AZ, UK  
 E. Wang  
 AMD  
 Cambridge CB4 1YG, UK

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/adv.202105784>

© 2022 The Authors. *Advanced Science* published by Wiley-VCH GmbH  
 This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/adv.202105784



**Figure 1.** Overview of the novel ex situ training technique designed for memristive neural networks. Redefining the output operation of synaptic layers allows to take non-ohmic device behavior into account, which makes working with power-efficient high-resistance (and high-nonlinearity) devices feasible. Relating weight parameters in artificial neural networks to individual conductances in crossbar arrays enables to further decrease power consumption through regularization, as well as to adapt to nonidealities that depend on the conductance values. Using an aggregate validation metric provides a more reliable way of assessing memristive neural network performance.

Other techniques of dealing with memristor nonidealities include the following:

- 1) in situ (re)training of weights (or just a subset of them<sup>[14]</sup>) to recover from the effects of nonidealities,<sup>[15–19]</sup> including in convolutional neural networks (CNNs),<sup>[14,20]</sup> recurrent structures,<sup>[20]</sup> and in neural networks used for reinforcement learning<sup>[21]</sup>
- 2) modifying device structure, including inserting a buffer layer,<sup>[22]</sup> inserting an electro-thermal modulation layer,<sup>[23]</sup> and adopting bilayer structure<sup>[24]</sup>
- 3) using additional circuitry<sup>[25,26]</sup> to ensure more stable behavior.

However, many of these approaches are technology-specific and thus difficult to apply to different types of devices.

When optimizing the performance of memristive systems (as opposed to individual devices), software approaches may be preferable because they are usually technology-agnostic. For general applications, mapping or redundancy schemes can be used to mitigate the effects of faulty devices<sup>[27]</sup> or line resistance.<sup>[28]</sup> In the specific context of MNNs, multiple smaller nonideal networks may replace a large one and increase the accuracy in this way.<sup>[29]</sup> Alternatively, modifying ex situ training has been proposed: altering the cost function<sup>[30]</sup> or injecting noise into the synaptic weights<sup>[31]</sup> can make MNNs more robust to the effects of nonidealities.

Memristor-oriented ex situ training is indeed a very promising method of making MNNs feasible. However, it has been applied by considering only a limited number of nonidealities, while the robustness of this technique is not well understood. In this work,

we propose a number of improvements to memristive ex situ training, which are summarized in **Figure 1**.

First, our novel training technique addresses the problem of nonlinearity. Deviations from the linear relationship between inputs and outputs in crossbar arrays are a major obstacle; however, none of the aforementioned methods directly address this issue. Although existing works often take *conductance* deviations into account during ex situ training, crossbar arrays are still usually modeled as structures computing a product of a vector of voltages and a matrix of conductances. To reflect non-ohmic behavior of memristive devices (illustrated in Figure 1) during training, we propose incorporating nonlinearities into the node functions of MNNs. Our aim is to embrace memristor nonlinearity so that the network can learn to be robust toward this nonideality—or even take advantage of it—during training based on stochastic gradient descent. Existing works attempting to take device variations into account during ex situ training often result in potentially lower training accuracy.<sup>[32]</sup> Depending on the nature of the nonideality that MNNs are trained on, our method could potentially perform even *better* than conventional ANNs. For example, if nonlinear response of the device is sufficiently consistent, nonlinearities may increase the network capacity.<sup>[33]</sup>

We also show how to improve nonideality-aware training by exposing conductances to the training process in a more direct way. By constraining MNN weights to be nonnegative, they can be related to conductances in a linear way. This allows to

- 1) minimize the effects of nonidealities in cases where their severity is dependent on the conductance values of the memristors

- 2) employ regularization as a high-level tool for controlling power consumption of MNNs

Finally, we propose improving validation techniques applied during the training of MNNs. Most memristive nonidealities are stochastic in nature, therefore computing validation metrics only once at specified checkpoints may provide unreliable estimates of neural network performance, as illustrated in Figure 1. We propose computing validation metrics multiple times and using an aggregate value (like the median) to determine which version of the MNN to save for inference.

In this work, we explore how to make nonideality-aware training more effective. We employ experimental data from a  $\text{SiO}_x$  memristor device and a  $128 \times 64$  Ta/HfO<sub>2</sub> memristor crossbar array, and consider multiple nonidealities:  $I$ - $V$  nonlinearity, stuck devices, and D2D variability. We give special focus to  $I$ - $V$  nonlinearity because existing works often prefer low-resistance devices as they exhibit more linear  $I$ - $V$  behavior. We show that our proposed training function enables the use of the more power-efficient high-resistance memristors by minimizing the accuracy loss due to high nonlinearity and variability, which typically manifest themselves in high-resistance devices.<sup>[34]</sup> We demonstrate how the proposed mapping between nonnegative weights and memristor conductances enables further power savings through  $\mathcal{L}_1$  regularization. We additionally show how MNN validation during training may be improved by taking the stochastic nature of many of the nonidealities into account.

Importantly, we demonstrate the feasibility of our novel training technique. Networks produced using nonideality-aware ex situ training are better adapted to nonidealities and can even handle significant uncertainty in device behavior or designers' understanding of what that behavior is. Weights learned *once* on a digital system can be transferred onto *multiple* crossbar-based physical implementations, even if the nonidealities manifest themselves differently in each system—our methodology does not require retraining once the weights are mapped onto crossbars. To assess the robustness of nonideality-aware training, we even expose MNNs to different setups than they were trained for—we find that the networks employing this technique are much more robust than the networks trained using the previous, more conventional approaches.

## 2. Design

### 2.1. Nonlinearity-Aware Training

The main focus of this work is the design of a novel ex situ MNN training scheme that could handle nonidealities characterized by the nonlinear relationship between inputs and outputs. Nonidealities like stuck devices, programming inaccuracies, or random telegraph noise can typically be represented with a change in memristor conductance *alone*. They are thus more straightforward to take into consideration during ex situ training because one can inject noise into the conductance array to represent their effect. We refer to these nonidealities as **linearity preserving**. On the other hand, nonidealities like  $I$ - $V$  nonlinearity or line resistance cannot be represented by simply disturbing the conductances of crossbar devices. We refer to such nonidealities as **linearity nonpreserving**. In this work, we redefine

the output operations of the synaptic layers to take the effect of linearity-nonpreserving nonidealities into account. To the best of our knowledge, this is the first time these nonidealities are addressed during ex situ training.

Existing works usually model memristor crossbars as structures computing linear dot products, while only activation functions are assumed to introduce nonlinearities. Specifically, outputs  $y_j \in \mathbf{y} \in \mathbb{R}^{1 \times N}$  are calculated using inputs  $x_i \in \mathbf{x} \in \mathbb{R}^{1 \times M}$ , weights  $w_{ij} \in \mathbf{w} \in \mathbb{R}^{M \times N}$ , and a nonlinear activation function  $f$ , as shown in Equation (1). During inference,  $\mathbf{x}$ ,  $\mathbf{w}$ , and  $\mathbf{y}$  are then mapped onto and from voltages, conductances, and currents, respectively. However, this creates a discrepancy between the linear dot-product training nodes and memristor nonidealities that deviate from ohmic device behavior.

$$y_j = f\left(\sum_{i=1}^M x_i w_{ij}\right) \quad (1)$$

Therefore, we suggest modifying the output operation of the synaptic layers to reflect the nature of linearity-nonpreserving nonidealities. Specifically, in cases where the nonlinearity is limited to individual devices (i.e. where devices experience  $I$ - $V$  nonlinearity), we propose replacing the approach in Equation (1) with the approach in Equation (2). That is, the activation function is unchanged but the product is replaced with a nonlinear function  $g$  that captures memristors' non-ohmic  $I$ - $V$  behavior.

$$y_j = f\left(\sum_{i=1}^M g(x_i, w_{ij})\right) \quad (2)$$

The exact form of  $g$  will depend on

- 1) the mapping scheme, examples of which are explored in Section 2.2
- 2) non-ohmic behavior model, which might typically be determined by the devices used; in Section 2.4, we present one possibility motivated by experimental device data

### 2.2. Modified Weight Implementation

As mentioned previously, synaptic weights of MNNs are typically implemented using conductances of memristive devices. While weights can usually take any real value, conductances are nonnegative—this presents design challenges. In this work, we propose training double the number of weights,<sup>[35]</sup> but constrain them to be nonnegative and associate them with individual devices. This approach creates a more natural mapping between neural network parameters and conductances, as well as enables weight regularization to act as a method for decreasing the power consumption.

#### 2.2.1. Conventional Approaches

In typical MNN implementations, both inputs  $x \in \mathbf{x}$  and outputs  $y \in \mathbf{y}$  are mapped onto voltages<sup>[36,37]</sup>  $V$  and from currents  $I$ , respectively, in a proportional way using scaling factors  $k_V$  and  $k_I$ ,

as shown in Equation (3).

$$V = k_V x \quad (3a)$$

$$\gamma = \frac{I}{k_I} \quad \text{where } k_I = k_V k_G \quad (3b)$$

where  $k_G$  is the conductance scaling factor typically made equal to  $\frac{G_{\text{on}} - G_{\text{off}}}{\max|\mathbf{w}|}$  and  $k_V$  is determined before training.

To encode both positive and negative weights, *pairs* of conductances are employed. Conductances  $G_+$  and  $G_-$  are introduced into “positive” and “negative” bit lines of crossbar arrays, where the output currents of the latter are subtracted from the output currents of the former; this is referred as **differential pair architecture**. Each weight is typically made proportional to the difference of  $G_+$  and  $G_-$  (with  $k_G$  acting as the constant of proportionality), which enables to encode any real number within a finite interval. However, infinite conductance combinations will produce the same difference,<sup>[38]</sup> thus the network designer may have to make an arbitrary choice of how to perform this mapping. For example, to encode weights  $w \in \mathbf{w}$ , the two conductances may be picked symmetrically around the average value,<sup>[38]</sup> as shown in Equation 4.

$$G_{\pm} = G_{\text{avg}} \pm \frac{k_G w}{2} \quad (4)$$

$$\text{where } G_{\text{avg}} \equiv \frac{G_{\text{off}} + G_{\text{on}}}{2}.$$

Although there might be advantages to using the mapping scheme like the one in Equation (4),<sup>[39]</sup> the choice of mapping could be *explicitly* tied to certain objectives. For example, Ref. [25] points out that differential pair architecture with aware mapping can be advantageous for mitigating the effects of stuck devices. Alternatively, a mapping scheme that optimizes some metric (like power consumption) may be employed. Indeed, such a scheme is used throughout this work for mapping the weights of conventionally trained ANNs onto conductances—we minimize power consumption by ensuring that at least one device in the pair is set to  $G_{\text{off}}$ , as demonstrated in Equation (5).

$$\begin{aligned} G_+ &= G_{\text{off}} + \max\{0, k_G w\} \\ G_- &= G_{\text{off}} - \min\{0, k_G w\} \end{aligned} \quad (5)$$

However, choosing the optimal scheme manually is a low-level approach that requires understanding the physical characteristics of MNNs. Thus, even if training is done *ex situ*, network designer has to make choices about not only the conventional training hyperparameters (like learning rate), but also how the system will be implemented physically because that will affect MNN performance. We see this as an additional obstacle to making memristive implementations of ANNs feasible in the real world.

### 2.2.2. Our Approach—Double Weights

In this work, instead of tweaking the mapping function, we decided to change the characteristics of the weights that are be-

ing trained. Specifically, we train two sets of *nonnegative* weights,  $w_{ij}^+ \in \mathbf{w}_+ \in \mathbb{R}_{\geq 0}^{M \times N}$  and  $w_{ij}^- \in \mathbf{w}_- \in \mathbb{R}_{\geq 0}^{M \times N}$ , which we refer to as **double weights**; similar approach has been explored in Ref. [40]. We map double weights onto conductances in the aforementioned “positive” and “negative” bit lines, respectively. Although all the weight parameters are nonnegative, the negative contribution of  $i^{\text{th}}$  input on  $j^{\text{th}}$  output can still be encoded because of the differential pair architecture in the *physical* system. Only the nonlinearity-aware *ex situ* training function in Equation (2) has to be adjusted leading to the form in Equation (6).

$$y_j = f \left( \sum_{i=1}^M g(x_i, w_{ij}^+) - g(x_i, w_{ij}^-) \right) \quad (6)$$

The adoption of double weights allows to relate *every* weight in  $\mathbf{w} \equiv [\mathbf{w}_+ \ \mathbf{w}_-]$  to the corresponding conductance in the *same* way, i.e.,  $w_{\pm} \in [0, \max(\mathbf{w})]$  are linearly mapped onto  $G_{\pm} \in [G_{\text{off}}, G_{\text{on}}]$ , as shown in Equation (7).

$$G_{\pm} = k_G w_{\pm} + G_{\text{off}} \quad (7)$$

A clear advantage of our approach is that double weights allow for more direct optimization of MNN behavior. Exposing raw device characteristics—i.e., conductances—to the training algorithm, enables it to select the combination that has both the optimal performance (as defined by some metric like loss) and high robustness. For example, if a certain nonideality manifests itself to a greater degree at low conductances, the training algorithm would be able to push double weight pairs (and by extension, conductances) toward higher values. Because of the differential pair architecture, setting  $G_+$  and  $G_-$  to 1.0 and 2.0 mS, respectively, will—at least in the case of linearity-preserving nonidealities—have the same effect as setting them to 3.0 and 4.0 mS, respectively. Therefore, with double weights, the training algorithm should be able to choose conductance combinations that minimize the negative influence of nonidealities.

Additionally, double weights allow regularization to act as a high-level tool for controlling the importance of power consumption. We propose training with the  $\ell_1$  sparsification regularizer,<sup>[41]</sup> which can not only improve training,<sup>[42]</sup> but also promote lower conductances because they are linearly related to weight parameters, as demonstrated in Equation (7). During backward propagation, the regularizer influences training loss, inducing conductance pairs to descend toward  $G_{\text{off}}$ . Instead of manually tweaking the mapping function, network designer can decide to what extent energy efficiency should be prioritized by simply adjusting, say, regularization factor in  $\ell_1$  regularization. This can be incorporated into typical hyperparameter tuning process that is performed before deploying ANNs in practice.

### 2.3. Modified Validation

We also propose a modified model validation scheme more fit for MNNs. To determine when to stop the training (or which version of the network to use when the training takes a predetermined set of epochs), a validation dataset is commonly employed—a metric (like error or loss) is computed for this dataset at certain epochs

and is used for picking the optimal version of the network.<sup>[43]</sup> However, many of memristor nonidealities are at least partly stochastic in nature, thus, say, validation accuracy at any given epoch may not be entirely representative of the model's quality purely due to random chance. Because of this, we suggest computing the validation metric multiple times and using an aggregated value for higher reliability. Choices can be made about

- aggregate value that is used
- how frequently validation is performed
- how many times validation metric is computed at each check-point

In the simulations of this work, we computed validation error 20 times every 20 epochs and saved the model whenever the *median* validation error decreased.

## 2.4. Nonidealities

In this work, we explore a wide range of memristor nonidealities and utilize experimental data. We use two different memristor technologies—SiO<sub>x</sub>- and Ta/HfO<sub>2</sub>-based resistive random-access memory (RRAM). More details on the two technologies can be found in Section 5 and our previous publications.<sup>[10,29]</sup>

### 2.4.1. *I-V Nonlinearity*

One of the most common ways to characterize deviations from ohmic behavior in memristive devices has been by considering two points on an *I-V* curve.<sup>[44,45]</sup> For example, one may define nonlinearity at voltage  $V_{\text{ref}}$  as the ratio of the conductance at that voltage to the conductance at half that voltage,<sup>[44]</sup> as shown in Equation (8). Nonlinearity of 1 can then be characterized as indicative of ohmic behavior; similarly, any deviations from that value indicate *I-V* nonlinearity. This metric can be useful in describing non-ohmic behavior at different voltages but it is more challenging to utilize it for modeling purposes.

$$\text{nonlinearity} \equiv \frac{G(V_{\text{ref}})}{G(V_{\text{ref}}/2)} \quad \text{where } G(V) \equiv \frac{\text{current at voltage } V}{\text{voltage } V} \quad (8)$$

In this work, we utilized silicon oxide devices to investigate the effects of current–voltage nonlinearity. SiO<sub>x</sub> devices can undergo resistance switching—typical *I-V* switching curve is shown in Figure S1 (Supporting Information), while a more detailed analysis of resistance switching performance can be found in our previous study.<sup>[34]</sup> For the purposes of this work, to achieve a wide range of resistance states and to analyze *I-V* nonlinearity, incremental positive sweeps were used to gradually reset the device from the low-resistance state to the high-resistance state. *I-V* curves of two subsets of all achieved states are shown in Figure 2a,b. Low-resistance discrete states (in Figure 2a) exhibit more linear behavior and experience little variability in nonlinearity. On the other hand, high-resistance states (in Figure 2b) are more nonlinear and the nonlinearity is less predictable.

Multiple conduction mechanisms have been proposed for modeling *I-V* behavior of memristors. The models may incorporate principles like Fowler–Nordheim tunneling, thermionic emission, and Poole–Frenkel emission.<sup>[46]</sup> Data in Figure 2a,b were best fit using Poole–Frenkel model in Equation (9).

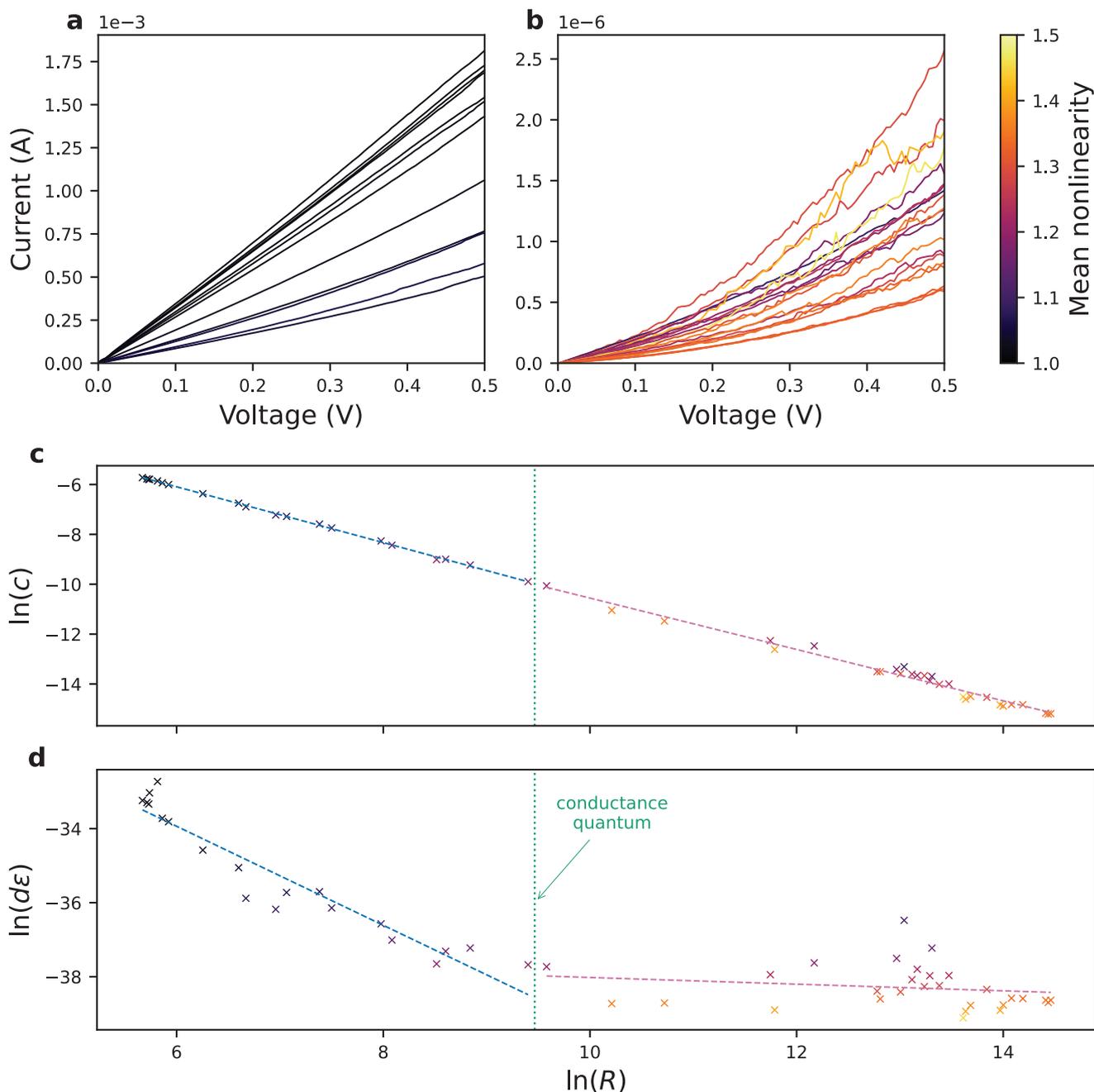
$$I = cV \exp\left(\frac{2e}{k_B T} \sqrt{\frac{eV}{4\pi d\epsilon}}\right) \quad (9)$$

where  $I$  is the current,  $V$  is the voltage,  $c$  is a constant (which has units of conductance),  $T$  is the temperature,  $d$  is assumed to be oxide thickness or effective thickness for partially oxidized filament, and  $\epsilon$  is the permittivity.

$T$  is room temperature, thus the following parameters were fit:  $c$  and the product of  $d$  and  $\epsilon$ . Different sets of fits were produced for above and below the conductance quantum  $G_0 = 2e^2/h$  because states below this value experienced different trends and/or amounts of variability. This is not surprising as memristive devices have been reported to exhibit different behavior in different resistance states, and that is often tied to the conductance quantum.<sup>[47,48]</sup>

The fits for the two sets of parameters are shown in Figure 2c,d. Slopes close to  $-1$  and intercepts close to 0 in Figure 2c demonstrate that  $c$  does indeed act similarly to conductance, i.e. reciprocal of resistance. In high-resistance states on the right, however, the variability is noticeably higher. In Figure 2d, product  $d\epsilon$  is indicative of the extent to which a given curve is nonlinear—the smaller this product, the more nonlinear the *I-V* curve is. For the low-resistance states,  $d\epsilon$  decreases with increasing resistance; however, for the first few states, this product is large—Equation (9) can approximate linear behavior (ohmic conduction); however, the values of  $d\epsilon$  are physically plausible only for the high resistance states ( $<G_0$ ). For those high-resistance states on the right side of Figure 2d, the product is, on average, lower but there is no *clear* trend *between* those states. It is also more difficult to predict  $d\epsilon$  from a given  $R$  as the deviations from the linear fit are very large. Thus, not only are highly resistive states more nonlinear, but there is also a significant level of variability. This is also indicated in the colors of the curves in Figure 2b where the last few states are shown—there is no obvious relationship between the resistance state and the color (which indicates nonlinearity). During nonideality-aware training, we take into account not only the nonlinearity but also the variability of *I-V* curves.

The simulations were performed by considering two different resistance ranges from our experimental data and by subsequently using fits from Figure 2c,d. Low-resistance group was constructed by interpolating between the lowest achieved resistance state of 289.8  $\Omega$  and five times that resistance. Similarly, to ensure the same dynamic range, high-resistance group was constructed by interpolating between the highest achieved resistance state of 1.905 M $\Omega$  and one fifth that resistance.<sup>[49]</sup> As hinted earlier, it is naive to simply use the aforementioned linear fits without taking into account significant deviations. In fact, the presence of uncertainty in the model is one of our main goals because we wish to demonstrate that nonideality-aware training can adapt not only to deviations from linear behavior but also to stochastic behavior. This makes nonideality-aware training

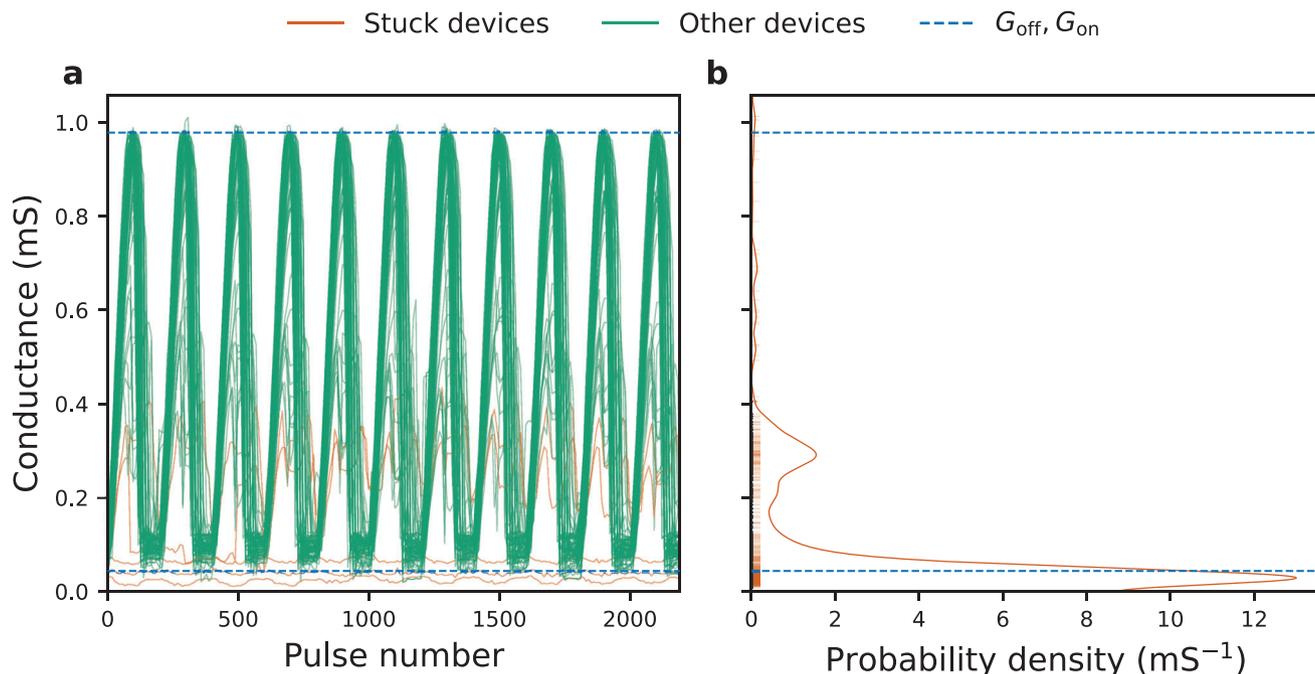


**Figure 2.**  $\text{SiO}_x$  data utilized in this work.  $I$ - $V$  sweeps of a  $\text{SiO}_x$  device are shown for **a**) a subset of the low-resistance region (with resistance at 0.1 V ranging from 289.8 to 1169  $\Omega$ ), and **b**) a subset of the high-resistance region (with resistance at 0.1 V ranging from 445.2 k $\Omega$  to 1.905 M $\Omega$ ). For all curves, only the range of voltages from 0.0 V to 0.5 V was considered. Nonlinearity was computed by dividing the conductance at voltage  $V$  by the conductance at voltage 0.5V. Poole–Frenkel fits in low- (left) and high-resistance (right) regions for **c**)  $c$  and **d**)  $d\epsilon$ . In both panels, to ensure dimensionless inputs to logarithms,  $R$ ,  $c$  and  $d\epsilon$  are the amounts of the corresponding quantities in SI units, e.g.,  $R$  is not the resistance, but rather the number of ohms of resistance measured at 0.1 V. Marker colors represent the mean nonlinearity of the fits, rather than the experimental  $I$ - $V$  curves.

approach generalizable because it does not require *exact* knowledge of device behavior—it improves the performance even when different hardware is used as will be demonstrated in Section 3.6.

During simulations involving nonlinear  $I$ - $V$  behavior, the output current of any given device was determined in the following way:

- 1)  $c$  and  $d\epsilon$  were interpolated from the fits in Figure 2c,d using resistance (parameter)  $R$
- 2)  $c$  and  $d\epsilon$  were disturbed using multivariate normal distribution with the covariance matrix determined using the residuals of the fits
- 3) current  $I$  was computed using Equation (9)



**Figure 3.** Experimental HfO<sub>2</sub> data utilized in this work. a) Eleven potentiation and depression cycles of 82 Ta/HfO<sub>2</sub> devices—data are shown for 1% of the devices in a 128 × 64 crossbar array. Any device whose maximum conductance range was less than half the median range was classified as “stuck.” The conductance of Ta/HfO<sub>2</sub> crossbar devices was modulated by controlling the voltage pulses applied to the gate of the selector transistor; more details are provided in Section 5 and Ref. [21]. b) Mean conductances of all stuck devices together with their estimated probability density function constructed using Gaussian kernel density estimation.

We note that such treatment, where only the linear relationship between the two sets of residuals is considered, likely *overestimates*, rather than underestimates, the amount of uncertainty for a given device. Additional information on heteroscedasticity, the correlation between the residuals of the two sets of parameters, and the justification for using normal distribution can be found in Section 5.

Linearity-nonpreserving nonidealities like *I-V* nonlinearity cannot be simulated using conventional noise injection methods, which simply disturb the conductance values. Instead, a forward propagation function must be defined reflecting the nonlinear relationship between inputs and outputs. One can express the procedure described earlier in the form of the aforementioned function *g* representing nonlinear behavior; this can be done by combining Equations (3), (7), and (9) leading to the form in Equation (10), which we implement using TensorFlow.

$$g(x, w_{\pm}) = \frac{1}{k_G} cx \exp\left(\frac{2e}{k_B T} \sqrt{\frac{ek_V x}{4\pi d\epsilon}}\right)$$

$$\text{where } \left[ \frac{c}{d\epsilon} \right] = \exp\left(\ln\left(\frac{1}{k_G w_{\pm} + G_{\text{off}}}\right) \mathbf{m} + \mathbf{b} + \mathbf{E}\right) \quad (10)$$

$$\text{and } \mathbf{E} \sim \mathcal{N}_2(\mathbf{0}, \Sigma)$$

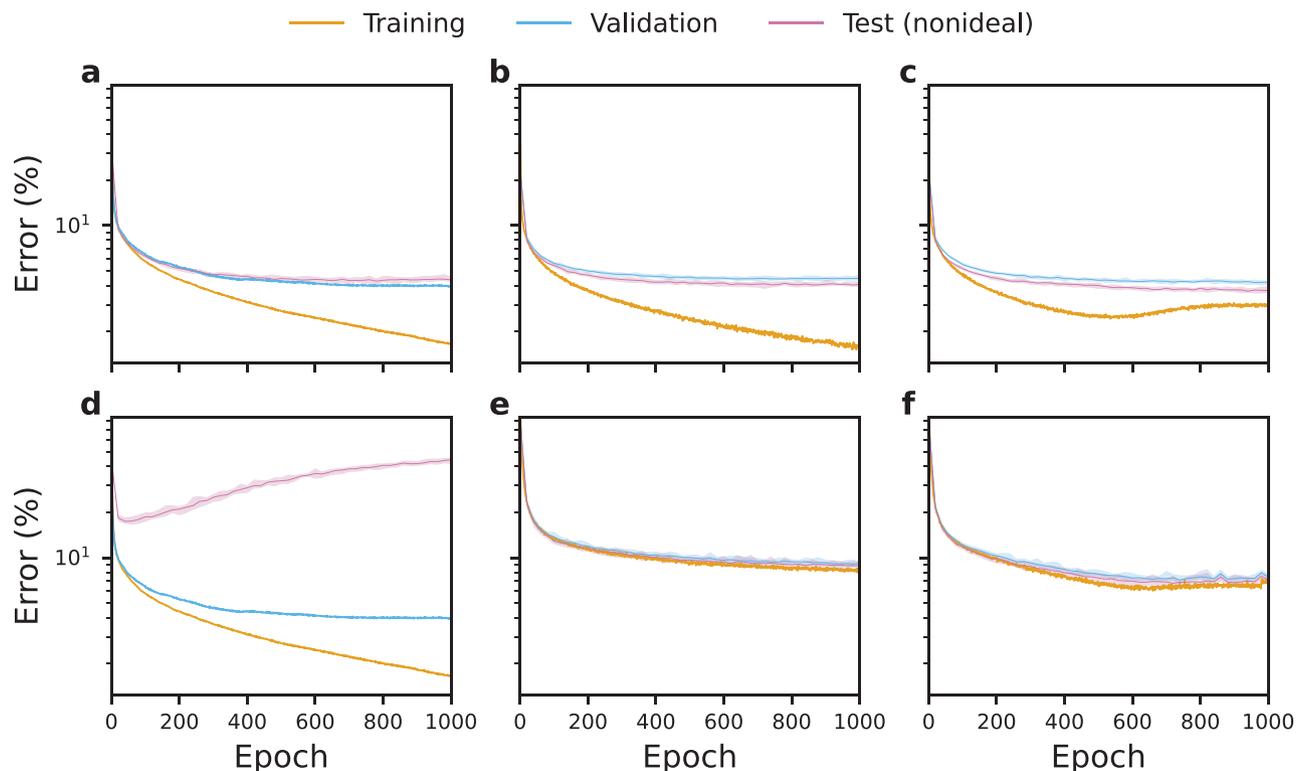
where **m** and **b** are slopes and intercepts, respectively, of the corresponding fits in Figure 2c,d, **Σ** is the covariance matrix of the residuals, and all inputs to logarithms or exponents are the amounts of quantities in SI units.

#### 2.4.2. Stuck Devices

Additionally, we explore linearity-preserving nonidealities, which can be simulated using noise injection into the conductances. One of such nonidealities is devices stuck in a particular state, which is a very common issue in memristors. The effect of stuck devices can be explored in isolation, but it also easily lends itself to being simulated along other nonidealities thus allowing to investigate the effectiveness of nonideality-aware training in more complex scenarios. In the modeling of this nonideality, we consider both real experimental data (where we draw the state in which devices may get stuck from a probability distribution) and a simplified model (where we assume devices can get stuck in only one state).

Data from 128 × 64 Ta/HfO<sub>2</sub> memristor crossbar array were analyzed for the purposes of modeling stuck devices' behavior. Figure 3a shows 11 potentiating and depression cycles (each consisting of a 100 voltage pulses) for a fraction of the devices. By considering the minimum and maximum conductance values achieved by each of the 8192 devices over 11 cycles, *G<sub>off</sub>* and *G<sub>on</sub>* were chosen as the median of these minimum and maximum values, respectively. Devices whose maximum range was less than half the median range (where median range had been defined as *G<sub>on</sub>* − *G<sub>off</sub>*) were classified as “stuck.”

Stuck devices were simulated using a probabilistic model. Using the aforementioned stuck devices' definition, 10.1% of the devices were classified as such. To simplify modeling, these devices were simulated to be *fully* stuck, meaning their conductance could not be changed even by a small amount.<sup>[50]</sup> The average values of all stuck devices are denoted by markers



**Figure 4.** Training results for standard and nonideality-aware schemes when exposed to  $I$ - $V$  nonlinearities. **a–c)** Low  $I$ - $V$  nonlinearity (data from Figure 2a), **d–f)** high  $I$ - $V$  nonlinearity (data from Figure 2b); **a, b, e)** standard training (same training and validation curves), **c, f)** nonideality-aware training with regularization. The panels show curves for one of five sets of trained networks. Networks were trained on MNIST dataset. Where error was computed multiple times, the curves show the median value, as well as the region bounded by the minimum and maximum values.

on the  $y$  axis of Figure 3b. The probability density function of these average values was constructed using kernel density estimation with truncated normal distributions. Scott's rule<sup>[51]</sup> was used for bandwidth estimation, while mirror reflections of the distributions were employed to correct for bias at the 0 S clipping boundary.<sup>[52]</sup> When simulating the effect of the nonideality, each device could be set to any conductance state between  $G_{\text{off}}$  and  $G_{\text{on}}$ ; however, every device also had a 10.1% probability of getting stuck. When a device was classified as stuck, its conductance would be drawn from the probability distribution constructed using kernel density estimation in Figure 3b.

Additionally, a simple model of devices getting stuck in  $G_{\text{off}}$  or  $G_{\text{on}}$  was considered. This allowed to combine nonidealities together, specifically with  $\text{SiO}_x$   $I$ - $V$  nonlinearities. For both states, devices were simulated to have 5% probability of getting stuck. As with experimental stuck devices' model, this was simulated as noise injection into the conductances. Both models of stuck device behavior were used to test the robustness of nonideality-aware training.

#### 2.4.3. Device-to-Device Variability

We also consider D2D variability arising from inaccuracies during device programming. During mapping onto conductances, one may end up with different values than intended; in some memristors, these (resistance) deviations are modeled using log-

normal distribution.<sup>[53]</sup> As with stuck devices, this can be incorporated into training by disturbing the values in each iteration—in this case, by drawing from lognormal distribution. We use this nonideality mostly to explore

- 1) double weights (see Section 3.4)
- 2) the effects of incredibly stochastic nonidealities (see Section 3.6)

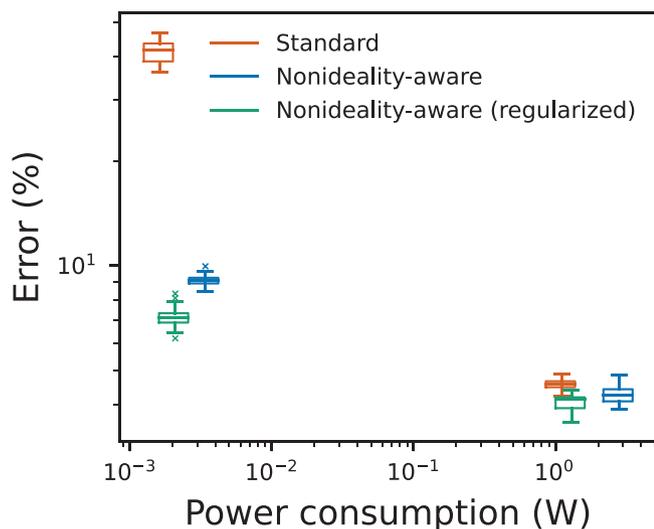
For the lognormal modeling, we linearly interpolate the standard deviation of the natural logarithm of resistances from the following values meant to represent different device behaviors:

- 1) 0.25 for  $R_{\text{off}}$  and  $R_{\text{on}}$  (more uniform D2D variability)
- 2) 0.5 for  $R_{\text{off}}$  and 0.05 for  $R_{\text{on}}$  (less uniform D2D variability)
- 3) 0.5 for  $R_{\text{off}}$  and  $R_{\text{on}}$  (high-magnitude D2D variability)

## 3. Results and Discussion

### 3.1. Nature of Training

Figure 4 contains training curves for MNNs trained on MNIST dataset and exposed to  $I$ - $V$  nonlinearities. Although not used to affect any aspect of the training process, error curves for the test subset<sup>[54]</sup> are included as well. Testing on nonideal configuration during training can help to understand the differences between standard and nonideality-aware training.



**Figure 5.** Inference results of using nonideality-aware training to deal with  $I$ - $V$  nonlinearities. The three box plots on the right refer to memristive neural networks that used data from Figure 2a, and the three on the left—to memristive neural networks that used data from Figure 2b. Networks were trained on MNIST dataset. The variability in power consumption for the data points of each group is small thus the average power consumption is used for the horizontal position of each box plot.

Figure 4a–c explores the effect of low  $I$ - $V$  nonlinearity. In Figure 4a, the validation curve (which is computed assuming digital implementation of the ANN) is closely coupled with the test curve (which assumes that nonideal effects *are* present)—this suggests negligible effect of low  $I$ - $V$  nonlinearity. Consequently, nonideality-aware training produces similar results on the test set both without (Figure 4b) and with (Figure 4c) regularization.

Figure 4d–f explores the effect of high  $I$ - $V$  nonlinearity. In Figure 4d, where the results of standard training are presented, we notice that validation and test curves are detached from one another. Not only that, but the global minimum of the (nonideal) test curve occurs very early in the training, while the (ideal) validation error keeps decreasing. This indicates that without taking nonidealities into account during training, a highly suboptimal version of the ANN may be chosen for inference stage with nonidealities. Nonideality-aware training without (Figure 4e) and with (Figure 4f) regularization is much more effective—the validation and test curves are closely coupled together and the test error decreases to lower values.

### 3.2. Performance Improvement

Inference results for  $I$ - $V$  nonlinearity simulations are summarized in Figure 5. Because the simulated memristors are nonlinear, power consumption was computed using  $P = IV$ , instead of  $P = I^2R$ , for each of the individual devices. Apart from crossbar arrays, MNN implementations require additional circuitry,<sup>[55]</sup> which suggests power consumption due to passive elements dominates in the  $\mu\text{S}$  range and above (i.e., the conductance range of the device investigated in this work); only at lower conductances does the relative impact of other energy components become significant.

Box plots representing low-resistance (and, by extension, low-nonlinearity) devices are presented on the right side of Figure 5. With such devices, nonideality-aware training without regularization achieves median error of 4.3%. This is close to the error—4.6%—of (digital) ANNs of the same size trained using standard procedure; the small difference is indicative of the relatively small effect of low  $I$ - $V$  nonlinearity (and its uncertainty). However, regularization not only decreases the power consumption by more than half, but also helps to achieve median error rate of 4.1%.

As shown on the left side of Figure 5, MNNs implemented using high-resistance devices achieve almost three orders of magnitude lower power consumption. However, MNNs trained using the standard procedure have median error of 41.6%, which would be unacceptable in most scenarios. Fortunately, adjusted training results in much lower error rate, while maintaining low power consumption (compared to low-nonlinearity devices). In the non-regularized case, the median error is 9.1%, and in regularized MNNs it is 7.1%. Thus, nonideality-aware training makes it feasible to use orders of magnitude more power-efficient high-resistance devices while maintaining error rates similar to those achieved with low-resistance devices.

One may also compare estimated absolute energy efficiency in both cases. By assuming the values used in<sup>[14]</sup>—read pulses of 50 ns, and two operations per synaptic weight (multiplication and accumulation)—one can calculate energy efficiency in  $\text{OPs}^{-1}\text{W}^{-1}$  using Equation (11).

$$\text{Energy efficiency} = \frac{2 \times n}{50 \times 10^{-9} \times P_{\text{avg}}} \quad (11)$$

where  $n$  is the number of synaptic weights and  $P_{\text{avg}}$  is the average power consumption.

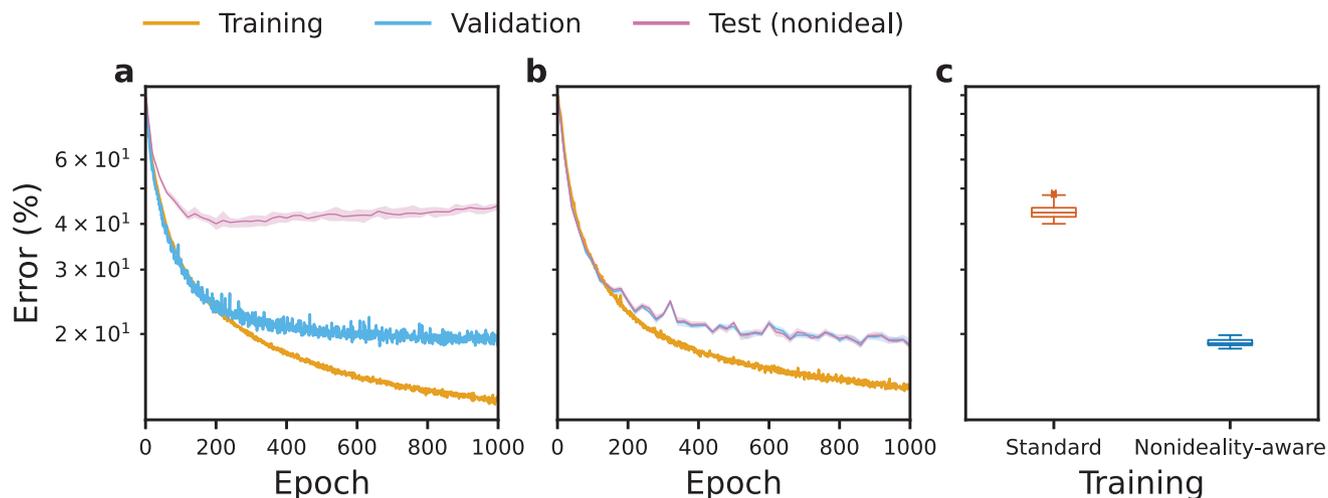
Using these assumptions, standard training using low-resistance devices achieves energy efficiency of  $0.715 \text{ TOPs}^{-1}\text{W}^{-1}$ , while nonideality-aware training using high-resistance devices achieves energy efficiency of  $234 \text{ TOPs}^{-1}\text{W}^{-1}$  in the nonregularized case and of  $381 \text{ TOPs}^{-1}\text{W}^{-1}$  in the regularized case. As explained earlier, these estimates incorporate power consumption only on crossbar arrays.

### 3.3. More Complex Architectures and Datasets

To understand how well nonideality-aware training performs on more complex tasks, we employed CIFAR-10 dataset. For this, we trained CNNs assuming that their convolutional layers would be implemented digitally, and their fully connected layers—using memristive crossbar arrays suffering from high  $I$ - $V$  nonlinearity. Standard training is explored in Figure 6a, and nonideality-aware training in Figure 6b. As with MNNs trained on MNIST, there is a much greater coupling between validation and test curves when nonidealities are taken into account. As shown in Figure 6c, nonideality-aware approach reduces the median inference error from 43.0% to 18.9%.

### 3.4. Importance of Weight Implementation

Double weights can be advantageous because they expose conductances to the training process in a more direct way. To



**Figure 6.** Results for standard and nonideality-aware schemes when employed by convolutional neural networks. **a)** Standard training, **b)** nonideality-aware training, and **c)** inference error comparing the approaches in **(a)** and **(b)**. Networks were trained on CIFAR-10 dataset and their fully connected layers were exposed to high  $I$ - $V$  nonlinearity (data from Figure 2b) during inference. Panels **(a)** and **(b)** show curves for one of five sets of trained networks.

investigate the utility of this modified weight implementation, we utilized lognormal D2D variability of two kinds:

- 1) more uniform variability where the relative magnitude of deviations is the same throughout [ $G_{\text{off}}, G_{\text{on}}$ ]
- 2) less uniform variability where the relative magnitude of deviations is much greater near  $G_{\text{off}}$  compared to  $G_{\text{on}}$ —similar to what is experienced in real devices when trying to program them<sup>[53]</sup>

This allowed to

- 1) evaluate the performance of double weight implementation when the severity of nonideality does not depend on the conductance value
- 2) test whether double weight implementation would outperform standard weights when exposed to nonidealities whose severity depended on the conductance value

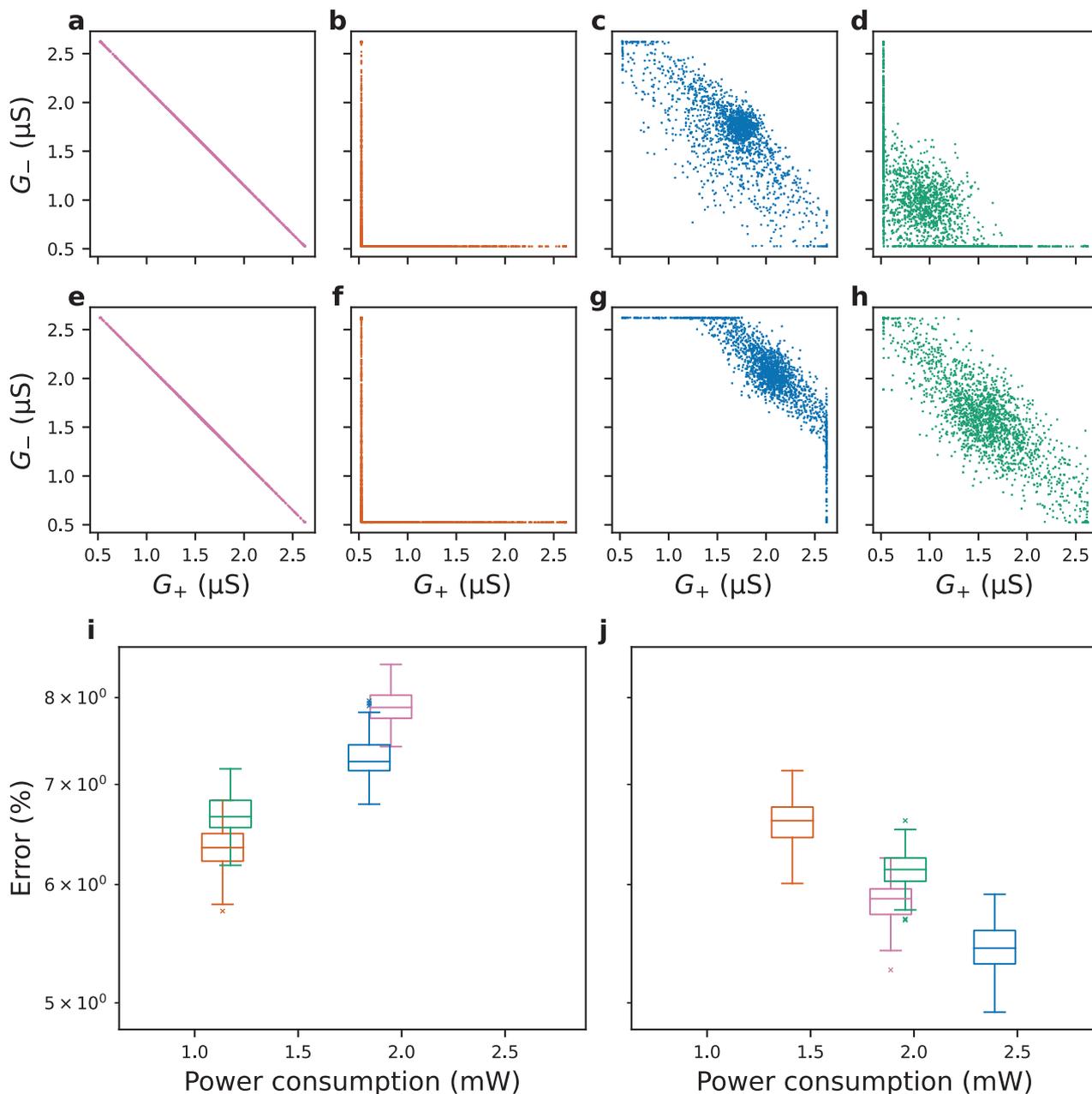
Both standard weights with different mapping schemes and double weights without and with regularization are investigated in Figure 7.

First, we consider weight implementations in the context of D2D variability with relatively high uniformity across the conductance range. Figure 7a,b shows conductances resulting from mapping conventional weights using rules in Equations (4) and (5), respectively. In both cases, the conductances are all mapped along either one or two line segments. In contrast, Figure 7c,d shows conductances obtained from double weight implementation and training without and with regularization, respectively. In the nonregularized case (Figure 7c), the conductances are distributed mostly around the diagonal, though they are spread out, unlike in Figure 7a. Regularization results in more data points in the bottom left corner of the diagram representing low conductance values, as shown in Figure 7d.

Figure 7e–h demonstrates the utility of double weights. Here, the MNNs were trained to deal with less uniform D2D variability where the disturbances were much greater at low conductance values compared to high conductance values. As demonstrated in Figure 7g, the training naturally results in most of the pairs concentrated in the top right corner representing high conductance values. Like before, regularization in Figure 7h results in conductances with lower values.

Figure 7i shows the inference error for various weight implementations of MNNs encountering more uniform D2D variability. Conventional weights with mapping rule in Equation (5) result in the lowest median error of 6.3%; in comparison, double weights achieve median error of 7.2% without regularization and median error of 6.7% with regularization. Conventional weights with mapping rule in Equation (4) achieve the highest median error of 7.9%, indicating the unpredictability of the performance of mapping methods. Even so, it is important to point out that double weights do not achieve optimal performance in this case. We hypothesize that in scenarios where the dependence of the severity of the nonideality on the conductance values is not strong, double weights might struggle to find optimal configuration—indefinitely many pairs may result in the same behavior, making it a more computationally difficult problem.

Figure 7j shows the inference error for equivalent weight implementations of MNNs dealing with less uniform D2D variability. In this case, the advantage of double weights is much more apparent—the median error in nonregularized case is 5.4% compared to 5.9% and 6.6% resulting from conventional weights with mapping rules in Equations (4) and (5), respectively. When double weights are trained by employing regularization, they take on lower values thus decreasing power consumption but also increasing the error—in the case of this specific nonideality, there is a tradeoff between energy efficiency and accuracy. However, double weights together with regularization provide a straightforward way of specifying to what extent low power consumption should be prioritized at the expense of accuracy.



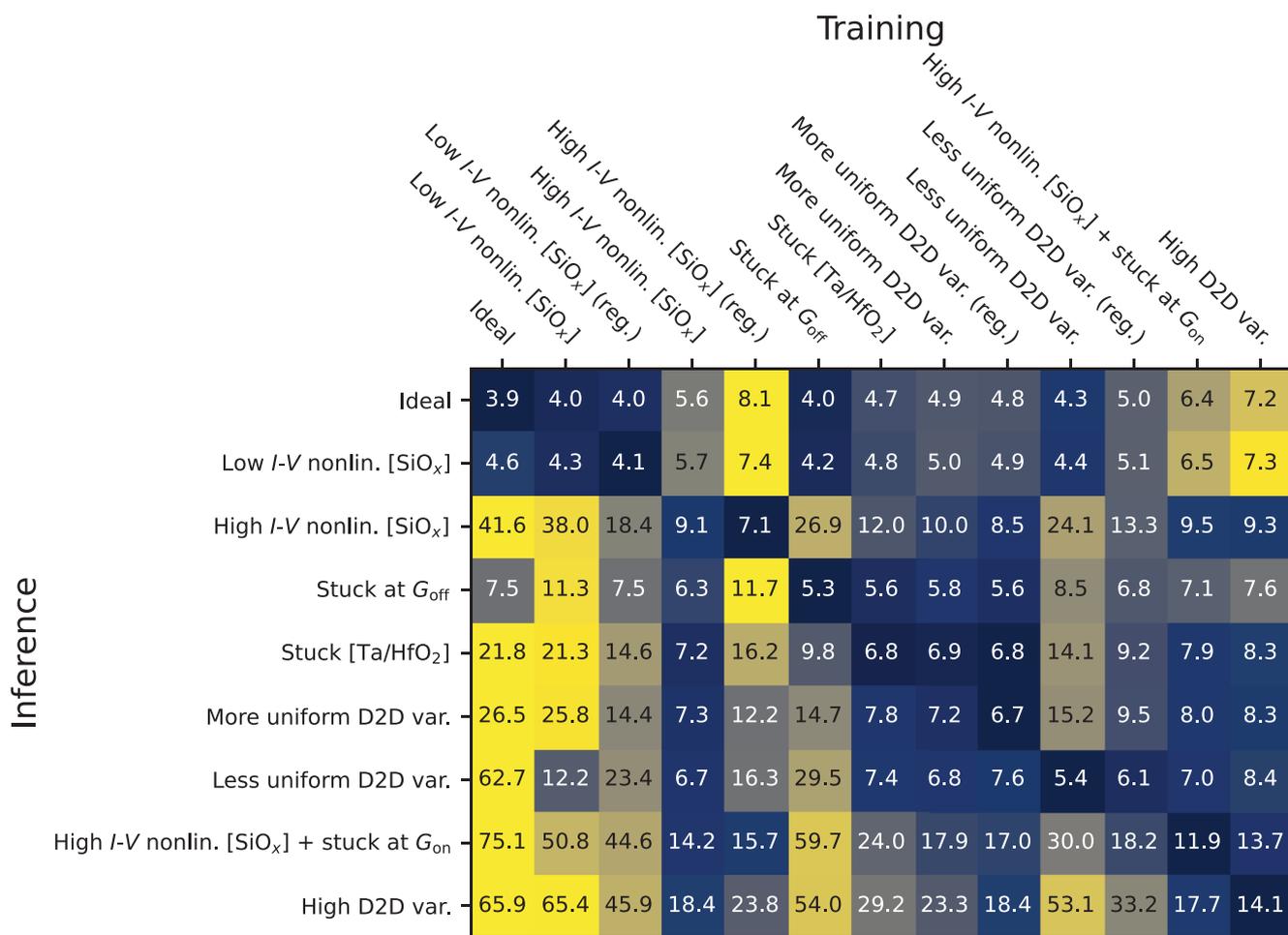
**Figure 7.** Comparison of weight implementations. Conductance distributions in the first synaptic layer of a memristive neural network that deals with more uniform device-to-device variability and utilizes **a**) conventional weights mapped symmetrically around average conductance, **b**) conventional weights mapped onto devices by preferring lowest total conductance, **c**) double weights, **d**) double weights with regularization; and **e–h**) corresponding conductance distributions for a memristive neural network dealing with *less* uniform device-to-device variability. Scatter plots contain conductance data from 10% of the devices for one of five sets of networks. Inference error for different weight implementations of memristive neural networks that deal with **i**) more and **j**) less uniform device-to-device variability, where the box plot colors correspond to different weight implementations in **a–d** and **e–h**, respectively

### 3.5. Memristive Validation

As explained in Section 2.3, many memristive nonidealities are nondeterministic, therefore it might be advantageous to compute an aggregate metric for use in validation. During training, with each batch, we simulate nonidealities separately, e.g., parameters for  $I$ - $V$  nonlinearity are drawn from a probability distribution or the exact devices that get stuck are picked randomly

each time. As a result, we believe that memristive validation can provide more reliable estimates of performance during training. Although we hypothesize that *in aggregate* this method will achieve only marginally better performance, it should help to avoid choosing a highly suboptimal version of the weights, which might yield higher error in a *small number* of cases.

Of course, memristive validation parameters may have to be optimized individually for each training configuration. For



**Figure 8.** Extent of nonideality agnosticism of nonideality-aware training. Median inference error in percent is shown for various training setups. Compared to standard (ideal) training, nonideality-aware approach usually results in lower error rate during inference, even if the nonideality encountered is different from the one the networks were exposed to during training. This is especially the case with more severe nonidealities. Networks were trained on MNIST dataset. Each row in the heatmap uses a separate instance of the logarithmically scaled colormap. Additional information on training and inference setups can be found in Table S4 (Supporting Information).

example, in the memristive CNN training in Figure 6b, the variability of validation error is usually lower at *any given checkpoint* than *between checkpoints*. In that case, one may increase the frequency of checkpoints by either decreasing the number of repeats at each checkpoint (and thus increasing uncertainty) or keeping it the same (and thus increasing computation time). At the extreme—if one can afford additional training time—validation error may be computed every epoch multiple times.

### 3.6. Nonideality Agnosticism

Accurate modeling of nonidealities for nonideality-aware ex situ training is a significant challenge. First, the nature of nonidealities encountered in practice may be different than what was modeled for the purposes of training. For example, the existence of D2D variability of SiO<sub>x</sub> memristors means that the behavior of any individual device is not perfectly representative of the nature of other devices. Therefore, to hedge against fitting the model to the behavior of any specific device, we assumed that Poole-

Frenkel parameters are inferred using a linear fit (determined by a trend in the experimental data) and disturbed by drawing random deviations from a probability distribution. Even so, in different devices, these trends and amounts of deviations may be different. Second, in the real world, one may encounter completely different types of nonidealities. If the training takes into account the effects of only, say, *I-V* nonlinearities, MNNs could still suffer from, for example, stuck devices when deployed. Therefore, it is important to find out how robust the MNNs employing nonideality-aware training are.

To investigate this, we utilized networks trained either by assuming no nonidealities or by being exposed to one of the eight different combinations of nonidealities. In the case of both types of *I-V* nonlinearity and two types of D2D variability, networks were additionally trained using regularization. During inference, each group of networks was then exposed to the

- 1) setup that they were trained on
- 2) setups of the other groups of networks

In total, this produced  $(1 + 8 + 2 + 2) \times (1 + 8) = 117$  scenarios; median inference error for each is presented in the heatmap in **Figure 8**.

The heatmap shows that, in most cases, the lowest error for a given nonideality during inference is achieved by the network that was exposed to it during training. A few exceptions exist, including regularized networks that were exposed to more uniform D2D variability during training—when exposed to high  $I$ - $V$  nonlinearity during inference, they achieve lower error of 8.5% compared to the error of 9.1% of networks that were trained (without regularization) on this exact nonideality. This may suggest that the nature of some nonidealities might often overlap and, when additional techniques like regularization are employed, the performance might be increased beyond even what can be achieved with the knowledge of a particular nonideality.

Importantly, **Figure 8** demonstrates that nonideality-aware training makes networks robust to the effects a wide array of nonidealities. Except for the ideal case and low-severity nonidealities (low  $I$ - $V$  nonlinearity and devices stuck at  $G_{\text{off}}$ ), nonideality-aware training results in lower median error compared to conventional training for all MNN groups—even when the networks encounter different nonidealities during inference. The heatmap also suggests that the effect of regularization on robustness is not always the same. In the case of high  $I$ - $V$  nonlinearity, regularization usually results in higher error (compared to nonregularized case) when encountering different nonidealities. On the other hand, regularization in networks dealing with D2D variability produces more robust behavior—when encountering different nonidealities, they usually achieve lower error.

#### 4. Conclusion

In this work, design a novel nonideality-aware training scheme that can improve the performance of MNNs. We demonstrate the importance of taking nonideal device behavior into account during training—without that, training performance may not be indicative of how well the networks will perform during inference. Importantly, we show the utility of nonideality-aware training in dealing with linearity-nonpreserving nonidealities, specifically  $I$ - $V$  nonlinearity. Our simulations show that the dichotomy between stable device behavior and power efficiency can become less relevant if the training stage is adjusted. Indeed, if nonideality-aware training is used, high-nonlinearity devices may achieve similar error rate, while also having almost three orders of magnitude better energy efficiency of  $381 \text{ TOPs}^{-1} \text{ W}^{-1}$  (with regularization) compared to  $0.715 \text{ TOPs}^{-1} \text{ W}^{-1}$  with the conventional training scheme and low-resistance devices.

We additionally explore a number of adjacent factors that are worth considering while dealing with MNN training. For example, we explore the use of double weights as a way to control conductance values of MNNs more directly through methods like  $\ell_1$  regularization. We explore the need to consider how validation is performed during training when nonidealities are stochastic and thus the outputs of an MNN are nondeterministic. In contrast to many previous works, we also investigate how robust nonideality-aware training is—we find that, compared to conventional training, it can usually achieve much lower error when encountering nonidealities that it was not trained to deal with. Besides, nonideality-aware training can be applied not just to one

single type of nonideality—it can deal with multiple nonidealities at once, for example  $I$ - $V$  nonlinearity and stuck devices, as shown in **Figure 8**.

Nonideality-aware training schemes are critical to making memristor-based ANNs feasible. Such schemes are key to ensuring low error rate and high energy efficiency. Using experimental data and several novel optimization techniques, we demonstrate that our training design deals with a wide range of nonidealities, importantly linearity-nonpreserving nonidealities, which have not been addressed during ex situ training before. Further, we demonstrate that high-resistance operational ranges can be used to reduce power consumption by almost three orders of magnitude without significant accuracy loss, despite the high level of associated nonidealities.

#### 5. Experimental Section

**Fabrication of  $\text{SiO}_x$ :**  $\text{SiO}_x$  RRAM devices were fabricated on a Si substrate with  $1 \mu\text{m}$  of thermal oxide on top. A 100 nm of Mo was deposited on top of the Si/ $\text{SiO}_2$  substrate that served as the bottom electrode. The  $\text{SiO}_x$  layer was sandwiched between the bottom and top electrodes and was deposited by reactive sputtering. The top electrodes consisted of 5 nm Ti wetting layer followed by a 100 nm of Au; their pattern was defined using a shadow mask. The device sizes ranged from  $200 \mu\text{m} \times 200 \mu\text{m}$  to  $800 \mu\text{m} \times 800 \mu\text{m}$ .

**Characterization of  $\text{SiO}_x$ :** Electrical characterization of a  $400 \mu\text{m} \times 400 \mu\text{m}$  device was performed using Keithley 4200A-SCS. The signals were applied to the top electrode (Au), while the bottom electrode (Mo) was connected to the ground. The device required an initial electroforming step before stable resistive switching could be achieved. The forming process was carried out by a negative voltage sweep, which stopped when the current had reached the limit of 3 mA. Subsequently, 18 voltage sweeps were performed to guarantee proper device performance: the voltage was ramped from 0.0 V, to  $\pm 2.5$  V, and back to 0.0 V using a 3 mA current compliance.

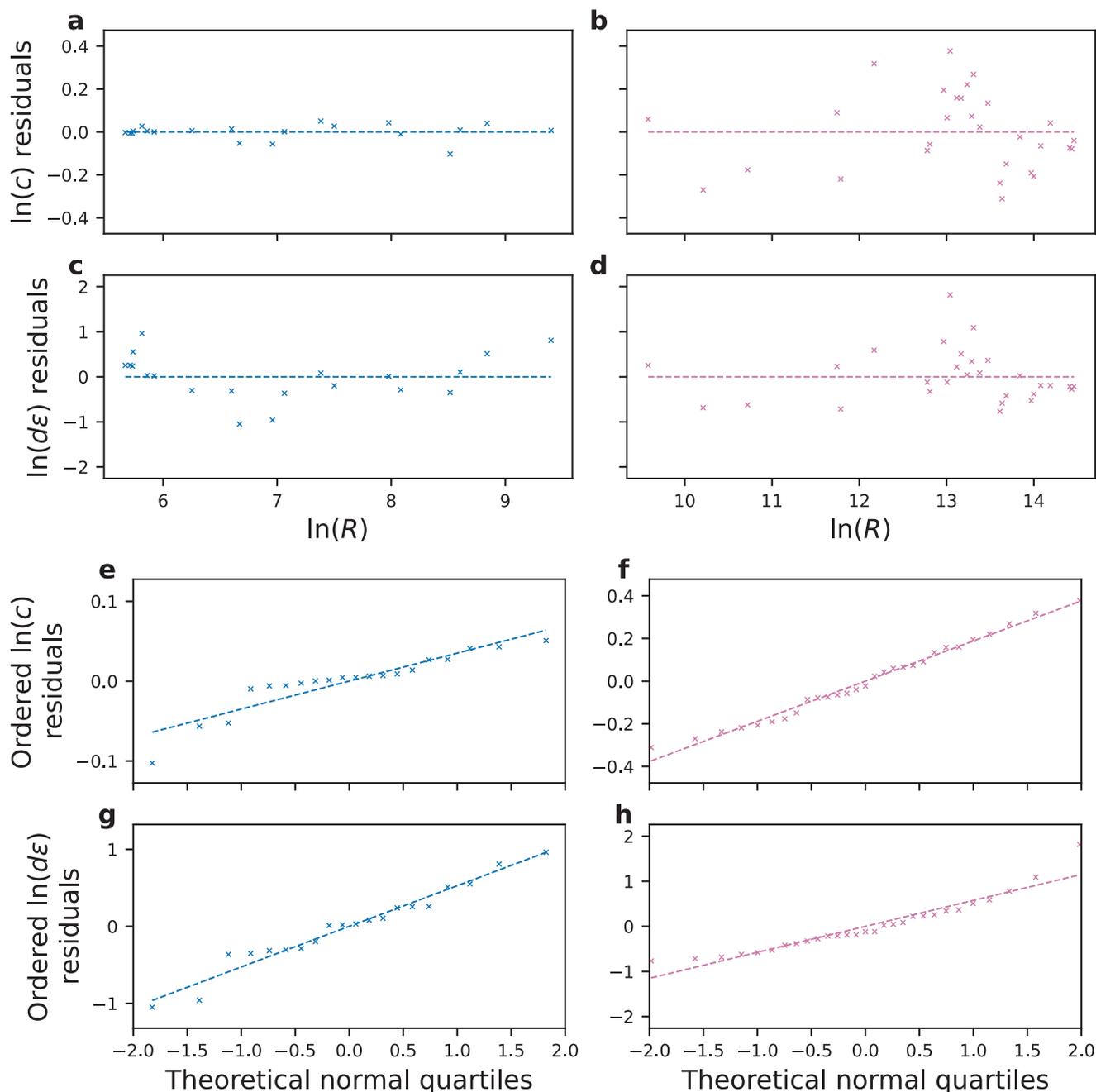
After this, to achieve a wide range of resistances, incremental positive sweeps were applied to the sample, starting from 0.5 V and increasing by 0.05 V in each run. This was being repeated until there was no further resistance change, i.e., the filament had returned to its initial (post-forming) state. The obtained  $I$ - $V$  curves are shown in (**Figure S2**, Supporting Information), while a subset of curves utilized in this work are shown in **Figure 2a,b**.

**Fabrication of  $\text{Ta}/\text{HfO}_2$ :**  $\text{Ta}/\text{HfO}_2$  1T1R array contained NMOS transistors (with feature size of 2 nm) and  $\text{Pt}/\text{HfO}_2/\text{Ta}$  RRAM devices. The bottom electrode was deposited by evaporating 20 nm Pt layer on top of a 2 nm Ta adhesive layer. A 5 nm  $\text{HfO}_2$  switching layer was deposited by atomic layer deposition using water and tetrakis(dimethylamido)hafnium as precursors at  $250^\circ\text{C}$ . 50 nm Ta sputtered layer followed by 10 nm Pd served as the top electrode.<sup>[29]</sup> Fabrication process is described in more detail in Ref. [56].

**Characterization of  $\text{Ta}/\text{HfO}_2$ :** Device conductance was being increased using SET pulses ( $500 \mu\text{s}$  @ 2.5 V and gate voltage linearly increasing from 0.6 V to 1.6 V). After each 100-pulse cycle, RESET pulses ( $5 \mu\text{s}$  @ 0.9 V linearly increasing to 2.2 V and gate voltage of 5 V) were used to reduce the conductance. More information can be found in Ref. [21].

**Simulations:** The following architectures were employed:

- 1) fully connected ANNs (trained on MNIST<sup>[57]</sup>) containing
  - a) fully connected layer with 25 hidden neurons and logistic activation function
  - b) fully connected layer with 10 output neurons and softmax activation function
- 2) CNNs (trained on CIFAR-10<sup>[58]</sup>) containing
  - a) convolutional layer with 32 output filters,  $3 \times 3$  kernel size and ReLU activation function
  - b) pooling layer with  $2 \times 2$  pool size



**Figure 9.** Residuals from the trends of Poole–Frenkel parameters. Residuals of  $\ln(c)$  for a) low- and b) high-resistance states. Residuals of  $\ln(d\varepsilon)$  for c) low- and d) high-resistance states. Normal probability plots of the residuals of  $\ln(c)$  for e) low- and f) high-resistance states. Normal probability plots of the residuals of  $\ln(d\varepsilon)$  for g) low- and h) high-resistance states. In all panels, the inputs to logarithms are made dimensionless by using the amounts of the corresponding quantities in SI units.

- c) convolutional layer with 64 output filters,  $3 \times 3$  kernel size and ReLU activation function
- d) pooling layer with  $2 \times 2$  pool size
- e) convolutional layer with 64 output filters,  $3 \times 3$  kernel size and ReLU activation function with maximum value of 1
- f) fully connected layer with 25 hidden neurons and logistic activation function
- g) fully connected layer with 10 output neurons and softmax activation function

To account for high variability of nonidealities (which were nondeterministic), five networks were trained for each configuration. Each trained network went through 25 inference runs, totaling  $5 \times 25 = 125$  runs for each configuration.

*Networks used 4:* 1 training-validation split. All networks were trained for a 1000 epochs with batch size of 64. Where  $\ell_1$  regularization had been employed, regularization factor of  $10^{-4}$  was used. To ensure double weights stayed nonnegative, NonNeg weight constraint provided by the Keras machine learning library was utilized. For any given *batch* (whose

size was 64 during training, as mentioned before, and 100 during inference), conductances were disturbed *once* in the case of linearity-preserving nonidealities and Poole-Frenkel parameters associated with individual devices were drawn from a probability distribution *once* in the case of *I-V* non-linearity.

#### Statistical Analysis:

- 1) In all box plots, the maximum whisker length was set to  $1.5 \times \text{IQR}$ .
- 2) In Figure 4 and equivalent plots, curves with semitransparent regions consist of two parts summarizing 20 inference repeats at certain epochs: opaque curve representing the median values and semitransparent region bounded by the minimum and maximum values.
- 3) To avoid large file size, only a subset of all data points is presented in Figures 3 and 7; these subsets were chosen randomly using NumPy.

Fifty-three  $\text{SiO}_x$  resistance states were achieved using the procedure described earlier but several (four) of them were excluded from the analysis. Specifically, states where there were abrupt changes in current were not considered. This was done by excluding the curves where the maximum ratio of the second derivative of current (with respect to voltage) to average current exceeded a threshold of  $0.1 \text{ V}^{-2}$ .

The analysis of the residuals of  $\ln(c)$  and  $\ln(dc)$  is provided in Figure 9. One of the issues, which was evident, was that these two sets of residuals correlated to some extent, especially at higher resistance states as can be seen in Figure 9b,d. If these deviations were simulated independently, the amount of uncertainty would be significantly overestimated, which is why covariance matrix of the residuals is used in Equation (10). The rationale for simulating the deviations using normal distribution is provided in the normal probability plots in Figure 9e–h.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

A.M. acknowledges funding from the Royal Academy of Engineering under the Research Fellowship scheme, G.A.C. acknowledges funding from the Engineering and Physical Sciences Research Council (EP/S030069/1 and EP/L016796/1), A.J.K. acknowledges funding from the Engineering and Physical Sciences Research Council (EP/P013503/1) and the Leverhulme Trust (RPG-2016-135), D.J. acknowledges studentship funding from the Engineering and Physical Sciences Research Council (ref. 2094654). E.W. did the work while at Imperial College London. The authors thank Professor Qiangfei Xia of University of Massachusetts Amherst for sharing data on the switching of  $\text{Ta}/\text{HfO}_2$  devices.

## Conflict of Interest

A.M. and A.J.K. are co-founders and W.H.N. is an employee of Intrinsic, a company developing memristor technology.

## Author Contributions

D.J. and E.W. contributed equally to this work, as did A.M. and G.A.C.

## Data Availability Statement

The source code for all the simulations in the paper can be found at <https://github.com/joksas/nonideality-aware-mnn-training/releases/tag/2022-04-22>.  $\text{Ta}/\text{HfO}_2$  data that support the findings of this study are available from D. J. upon request.  $\text{SiO}_x$  data have been made publicly available;<sup>[59]</sup> file `excelDataCombined.mat` was utilized in the simulations of this work. All data generated by the simulations have been made publicly available.<sup>[60]</sup>

## Keywords

hardware acceleration, memristors, neural networks

Received: December 13, 2021

Revised: March 18, 2022

Published online:

- [1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, et al., *Language Models are Few-Shot Learners in Advances in Neural Information Processing Systems* (Eds: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin), vol. 33, Curran Associates, Inc. **2020**, pp. 1877–1901. arXiv: 2005.14165 [cs.CL].
- [2] A. R. Fayjie, S. Hossain, D. Oualid, D.-J. Lee, in *2018 15th International Conference on Ubiquitous Robots (UR)*, IEEE **2018**, pp. 896–901. <https://doi.org/10.1109/URAI.2018.8441797>.
- [3] E. Strubell, A. Ganesh, A. McCallum, *Energy and Policy Considerations for Deep Learning in NLP* **2019**. arXiv: 1906.02243 [cs.CL].
- [4] S. Han, H. Mao, W. J. Dally, in *International Conference on Learning Representations* **2016**. arXiv: 1510.00149 [cs.CV].
- [5] B. Wu, F. Iandola, P. H. Jin, K. Keutzer, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* **2017**, pp. 446–454. <https://doi.org/10.1109/CVPRW.2017.60>.
- [6] A. Mehonic, A. J. Kenyon, arXiv: 679 2104.14517 [cs.ET], **2021**.
- [7] S. A. McKee, in *Proceedings of the 1st Conference on Computing Frontiers*, Association for Computing Machinery, Ischia, Italy, ISBN 1581137419 **2004**, p. 162. <https://doi.org/10.1145/977091.977115>.
- [8] J. J. Yang, D. B. Strukov, D. R. Stewart, *Nat. Nanotechnol.* **2013**, *8*, 13.
- [9] D. Ielmini, Z. Wang, Y. Liu, *APL Mater.* **2021**, *9*, 050702.
- [10] A. Mehonic, D. Joksas, W. H. Ng, M. Buckwell, A. J. Kenyon, *Frontiers in Neuroscience* **2019**, *13*, 593.
- [11] J. J. Yang, M.-X. Zhang, M. D. Pickett, F. Miao, J. P. Strachan, W.-D. Li, W. Yi, D. A. A. Ohlberg, B. J. Choi, W. Wu, et al., *Appl. Phys. Lett.* **2012**, *100*, 113501.
- [12] A. Amirsoleimani, F. Alibart, V. Yon, J. Xu, M. R. Pazhouhandeh, S. Ecoffey, Y. Beilliard, R. Genov, D. Drouin, *Advanced Intelligent Systems* **2020**, *2*, 2000115.
- [13] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, W. D. Lu, *Nat. Electron.* **2019**, *2*, 290.
- [14] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. J. Yang, H. Qian, *Nature* **2020**, *577*, 641.
- [15] L. Chen, J. Li, Y. Chen, Q. Deng, J. Shen, X. Liang, L. Jiang, in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2017, IEEE **2017**, pp. 19–24. <https://doi.org/10.23919/DATE.2017.7926952>.
- [16] S. Jain, A. Raghunathan, *ACM Transactions on Embedded Computing Systems (TECS)* **2019**, *18*, 1.
- [17] Y. Wang, S. Wu, L. Tian, L. Shi, *Neurocomputing* **2020**, *407*, 270.
- [18] C. Liu, M. Hu, J. P. Strachan, H. Li, in *2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*, IEEE **2017**, pp. 1–6. <https://doi.org/10.1145/3061639.3062310>.
- [19] B. Li, B. Yan, C. Liu, H. Li, in *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, **2019**, pp. 224–229. <https://doi.org/10.1145/3287624.3288744>.
- [20] Z. Wang, C. Li, P. Lin, M. Rao, Y. Nie, W. Song, Q. Qiu, Y. Li, P. Yan, J. P. Strachan, N. Ge, N. McDonald, Q. Wu, M. Hu, H. Wu, R. S. Williams, Q. Xia, J. J. Yang, *Nature Machine Intelligence* **2019**, *1*, 434.
- [21] Z. Wang, C. Li, W. Song, M. Rao, D. Belkin, Y. Li, P. Yan, H. Jiang, P. Lin, M. Hu, J. P. Strachan, N. Ge, M. Barnell, Q. Wu, A. G. Barto, Q. Qiu, R. S. Williams, Q. Xia, J. J. Yang, *Nat. Electron.* **2019**, *2*, 115.

- [22] Y. Fang, Z. Yu, Z. Wang, T. Zhang, Y. Yang, Y. Cai, R. Huang, *IEEE Electron Device Lett.* **2018**, *39*, 819.
- [23] W. Wu, H. Wu, B. Gao, P. Yao, X. Zhang, X. Peng, S. Yu, H. Qian, in *Symposium on VLSI Technology*, IEEE **2018**, <https://doi.org/10.1109/VLSIT.2018.8510690>.
- [24] J. Woo, K. Moon, J. Song, S. Lee, M. Kwak, J. Park, H. Hwang, *IEEE Electron Device Lett.* **2016**, *37*, 994.
- [25] C. Li, M. Hu, Y. Li, H. Jiang, N. Ge, E. Montgomery, J. Zhang, W. Song, N. Dávila, C. E. Graves, Z. Li, J. P. Strachan, P. Lin, Z. Wang, M. Barnell, Q. Wu, S. Williams, J. Yang, Q. Xia, *Nat. Electron.* **2018**, *1*, 52.
- [26] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. D. Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. P. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, G. W. Burr, *Nature* **2018**, *558*, 60.
- [27] L. Xia, W. Huangfu, T. Tang, X. Yin, K. Chakrabarty, Y. Xie, Y. Wang, H. Yang, *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* **2017**, *8*, 102.
- [28] M. Hu, J. P. Strachan, Z. Li, S. R. Williams, in *17th International Symposium on Quality Electronic Design* **2016**, <https://doi.org/10.1109/ISQED.2016.7479230>.
- [29] D. Joksas, P. Freitas, Z. Chai, W. H. Ng, M. Buckwell, C. Li, W. D. Zhang, Q. Xia, A. J. Kenyon, A. Mehonic, *Nat. Commun.* **2020**, *11*, 4273.
- [30] Y. Zhu, G. L. Zhang, T. Wang, B. Li, Y. Shi, T.-Y. Ho, U. Schlichtmann, in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, IEEE **2020**, pp. 1590–1593. <https://doi.org/10.23919/DATE48585.2020.9116244>.
- [31] V. Joshi, M. Le Gallo, S. Haefeli, I. Boybat, S. R. Nandakumar, C. Piveteau, M. Dazzi, B. Rajendran, A. Sebastian, E. Eleftheriou, *Nat. Commun.* **2020**, *11*, 1.
- [32] B. Liu, H. Li, Y. Chen, X. Li, Q. Wu, T. Huang, in *Proceedings of the 52nd Annual Design Automation Conference* **2015**, pp. 1–6. <https://doi.org/10.1145/2744769.2744930>.
- [33] J. A. Zavatone-Veth, C. Pehlevan, *Phys. Rev. E* **2021**, *103*, L020301.
- [34] A. Mehonic, M. S. Munde, W. H. Ng, M. Buckwell, L. Montesi, M. Bosman, A. L. Shluger, A. J. Kenyon, *Microelectron. Eng.* **2017**, *178*, 98.
- [35] The number of memristive devices remains the same as with conventional weight implementation schemes.
- [36] If the inputs are digital, they may be applied in the form of voltage using digital-to-analog converters (DACs) or, for example, by utilizing binarized inputs[37], which eliminate the need for such converters altogether.
- [37] C.-X. Xue, Y.-C. Chiu, T.-W. Liu, T.-Y. Huang, J.-S. Liu, T.-W. Chang, H.-Y. Kao, J.-H. Wang, S.-Y. Wei, C.-Y. Lee, et al., *Nat. Electron.* **2021**, *4*, 81.
- [38] H. Kim, M. R. Mahmoodi, H. Nili, D. B. Strukov, *Nat. Commun.* **2021**, *12*, 1.
- [39] An argument could be made, for example, that such a mapping scheme would produce very few conductances with values near  $G_{off}$  or  $G_{on}$ , which may be more difficult to achieve in real devices.
- [40] J. Kendall, R. Pantone, K. Manickavasagam, Y. Bengio, B. Scellier, *arXiv: 2006.01981* **2020**.
- [41] S. Han, J. Pool, J. Tran, W. J. Dally, in *International Conference on Neural Information Processing Systems* **2015**, pp. 1135–1143. arXiv: 1506.02626 [cs.NE].
- [42] This may manifest itself as avoiding overfitting, for example.
- [43] Y. Shao, G. N. Taff, S. J. Walsh, *IEEE Geoscience and Remote Sensing Letters* **2010**, *8*, 113.
- [44] C. Sung, S. Lim, H. Kim, T. Kim, K. Moon, J. Song, J.-J. Kim, H. Hwang, *Nanotechnology* **2018**, *29*, 115203.
- [45] F. Lentz, B. Roesgen, V. Rana, D. J. Wouters, R. Waser, *IEEE Electron Device Lett.* **2013**, *34*, 996.
- [46] S. M. Sze, Y. Li, K. K. Ng, *Physics of Semiconductor Devices*, John Wiley & Sons **2006**.
- [47] W. Yi, S. E. Savel'Ev, G. Medeiros-Ribeiro, F. Miao, M.-X. Zhang, J. Yang, A. M. Bratkovsky, R. S. Williams, *Nat. Commun.* **2016**, *7*, 1.
- [48] A. Mehonic, M. Buckwell, L. Montesi, L. Garnett, S. Hudziak, S. Fearn, R. Chater, D. McPhail, A. J. Kenyon, *J. Appl. Phys.* **2015**, *117*, 124505.
- [49] The curves of resistance states falling in either of these two intervals are the ones depicted in Figure 2a,b.
- [50] Similar to uncertainty in  $\text{SiO}_x$  devices, this assumption overestimates the effect of stuck  $\text{HfO}_2$  devices, not underestimates it.
- [51] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*, Wiley, New York, US **1992**.
- [52] To ensure numerical stability, mirror reflections were only included if the area under the curve of the underlying normal distribution below 0 S was more than  $10^{-8}$ .
- [53] K. M. Kim, J. J. Yang, J. P. Strachan, E. M. Grafals, N. Ge, N. D. Melendez, Z. Li, R. S. Williams, *Sci. Rep.* **2016**, *6*, 1.
- [54] Test set errors at training checkpoints were computed in the same way as memristive validation curves, see Section 2.3.
- [55] H.-Y. Chang, P. Narayanan, S. C. Lewis, N. C. P. Farinha, K. Hosokawa, C. Mackin, H. Tsai, S. Ambrogio, A. Chen, G. W. Burr, *IBM J. Res. Dev.* **2019**, *63*, 8:1.
- [56] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, W. Song, J. P. Strachan, M. Barnell, Q. Wu, R. S. Williams, J. J. Yang, Q. Xia, *Nat. Commun.* **2018**, *9*, 2385.
- [57] Y. LeCun, C. Cortes, C. J. C. Burges, *The MNIST database of handwritten digits*, **2010**, <http://yann.lecun.com/exdb/mnist>.
- [58] A. Krizhevsky, Learning multiple layers of features from tiny images 2009, <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>
- [59] N. Barmapsalos, A. Mehonic, *SiO<sub>x</sub> memristor structure - gradual RESET conductance modulation* **2021**, <https://doi.org/10.5281/zenodo.5728039>.
- [60] D. Joksas, E. Wang, N. Barmapsalos, W. H. Ng, A. J. Kenyon, G. A. Constantinides, A. Mehonic, *Generated data for the manuscript "Nonideality-aware training for accurate and robust low-power memristive neural networks"* **2021**, <https://doi.org/10.5281/zenodo.5775044>.