

Device Modeling Bias in ReRAM-based Neural Network Simulations

Osama Yousuf, Imtiaz Hossen, Matthew W. Daniels, Martin Lueker-Boden, Andrew Dienstfrey, Gina C. Adam

Abstract—Data-driven modeling approaches such as jump tables are promising techniques to model populations of resistive random-access memory (ReRAM) or other emerging memory devices for hardware neural network simulations. As these tables rely on data interpolation, this work explores the open questions about their fidelity in relation to the stochastic device behavior they model. We study how various jump table device models impact the attained network performance estimates, a concept we define as modeling bias. Two methods of jump table device modeling – binning and Optuna-optimized binning are explored using synthetic data with known distributions for benchmarking purposes, as well as experimental data obtained from TiO_x ReRAM devices. Results on a multi-layer perceptron trained on MNIST show that device models based on binning can behave unpredictably particularly at low number of points in the device dataset, sometimes over-promising, sometimes under-promising target network accuracy. This paper also proposes device level metrics that indicate similar trends with the modeling bias metric at the network level. The proposed approach opens the possibility for future investigations into statistical device models with better performance, as well as experimentally verified modeling bias in different in-memory computing and neural network architectures.

Index Terms—Hardware Neural Networks, ReRAM, memristors, Device Modeling, Modeling Bias.

I. INTRODUCTION

ARTIFICIAL INTELLIGENCE is exploding and is resource-hungry, with advanced neural networks requiring hundreds of thousands of computing chips. At such large scales, traditional computing cores exhibit excessively high data movement and energy demands owing to the von-Neumann bottleneck, bringing us closer to hitting the “power wall” [1]. Thus, there is a growing need for investigating technologies with non-von Neumann architectures. Such alternatives can serve as building blocks for faster and more efficient neural network accelerators that exploit near- or in-memory computation to minimize data movement. Non-volatile resistive switching devices promise both dense storage and energy-efficient analog processing, making them suitable for in-memory computing for artificial intelligence applications [2], [3].

Hardware-based neural networks based on emerging non-volatile memory devices can exploit underlying physical phenomena to efficiently implement matrix-vector multiplication – a critical operation in neural networks. In particular, ReRAM or memristor devices can be used as highly energy-efficient physical implementations of artificial synaptic weights for neural networks owing to their non-volatility and fast switching characteristics [4], [5]. However, these devices exhibit complex multi-physics behavior leading to performance degradation in prototype networks. As with many other emerging technologies, the impact of these device non-idealities on the network performance needs to be further studied and resolved before they can be used to realize large-scale analog accelerators.

Despite significant progress in the past two decades [4], the study of hardware neural networks based on ReRAM devices is facing several major barriers. First, a purely experimental approach is unfeasible since commercial ReRAM tape-outs have long timelines and significant design and fabrication costs [6]–[8]. Secondly, before hardware prototyping can be practically motivated, results from hardware-aware simulations are needed to reliably predict the promising performance of these systems and optimize their end-to-end behavior across all levels – from underlying devices to neural networks. Therefore, suitable modeling of ReRAM devices is a key consideration for hardware neural networks investigations and in-memory computing systems based on ReRAM devices in general.

A broad range of models has been proposed. Atomistic models focus entirely on simulations from first principles in an effort to capture the multi-physics dynamics of filament formation and their conductance characteristics. These models may be used to provide insights needed for material and device design optimization [9]–[12]. Compact models needed for circuit design use physically inspired parametrizations which are tuned to match experimental current vs. voltage characteristics [13], [14]. While both classes of models are appropriate for their respective purposes, they rely on computationally expensive systems of equations and are therefore not well-suited for scaling to neural network simulations. Additionally, these models require investigating

This work was supported in part by NIST under grant 70NANB22H018, by Western Digital under grant ECNS21932N and the GW Cross-Disciplinary Research Fund. Corresponding author: Gina C. Adam (e-mail ginaadam@gwu.edu).

Osama Yousuf, Imtiaz Hossen, and Gina C. Adam are with the George Washington University, Electrical and Computer Engineering Department, Washington, DC, 20052. Matthew W. Daniels is with the National Institute of Standards and Technology, Gaithersburg, MD, 20899, USA and Andrew Dienstfrey is with the National Institute of Standards and Technology, Boulder, CO, 80305 USA. Martin Lueker-Boden is with the Western Digital Technologies, San Jose, CA, 95119, USA.

underlying physical phenomena such as the shape of the filament [15]–[17] for new devices, which is a time-consuming process that can delay algorithmic investigation.

Due to these limitations, for the following study we propose to investigate jump table models to characterize ReRAM device irregularities. Such models consist of lookup tables that specify the probability of moving from one conductance state to another as a function of present conductance state and applied pulse (and potentially other measurement parameters). In contrast to the physics-based models discussed above, these models are derived solely from experimental data without reference to underlying physical mechanisms. While these physical mechanisms are important for device design, they are not relevant in predicting neural network performance. It is known that variability in the switching of ReRAM devices leads to weight dispersion, which can potentially negatively impact the accuracy and performance of the overall system [18], [19]. Therefore, accurate device modeling is needed for providing a realistic estimate of the training characteristics of analog neuromorphic systems implemented with real ReRAM devices.

In this paper, we introduce and study *modeling bias* as a useful concept to evaluate the fitness of a device model in the context of neural network simulations. We investigate the impact of two methods for creating jump table-based device models: the traditional binning and interpolation method proposed in [20], and an iterative hyperparameter optimization method based on the Optuna framework [21]. For benchmarking purposes, we synthesize ReRAM jump tables using an assumed closed-form model for device switching and investigate the convergence of our interpolation methods, as detailed in Section II A. These distributions are inspired by the “Real Device” model from the NeuroSimV3.0 simulator [22]. Trends in network metrics are compared with trends in the goodness-of-fit for the device to systematically analyze the performance of the interpolation methods for device modeling in the context of neural network simulations. We also present the training performance and methods to estimate the modeling bias when *experimental jump tables* are used. These are jump tables based on experimental data obtained from TiO_x ReRAM devices, for which the underlying distributions for signal mean and switching noise are not known.

The remainder of this paper is organized as follows. Section II covers the methods used in our analysis including details on jump table modeling, device datasets, neural network structure and training routines, and metrics for inferring data interpolation quality. Section III presents comparative neural network results using (a) synthetic datasets, where we investigate the impacts of the number of input data points, device cycle-to-cycle variation, and device non-linearity on modeling bias, and (b) experimental datasets, where we present an iterative approach for jump table modeling based on hyperparameter optimization and show how the modeling bias performance could potentially be estimated without experimental apparatus. Section IV is a discussion highlighting the limitations and opportunities in emerging device modeling for large-scale neural network simulations. We wrap up with conclusions in Section V.

II. METHODS

A. Jump table modeling

A jump table is a set of cumulative distribution functions (CDFs) that define the stochastic change in device conductance per voltage pulse as a function of the current conductance state (i.e., the distribution of ΔG per pulse as a function of G) [18], [20], [23]. We define a jump table device model x as a series of distributions as follows:

$$x = \{X_{G_1}, X_{G_2}, \dots, X_{G_m}\}, \quad (1)$$

where each X_{G_i} is a normal distribution with mean μ_i and standard deviation σ_i , and m is the total number of distributions. The *mean profile* $\mu(G)$ refers to the ordered list of μ_i 's, and likewise, the *standard deviation profile* $\sigma(G)$ for σ_i 's.

Jump tables provide a single model to encompass both the stochastic nature of device programming and the non-uniform conductance response ΔG as a function of conductance G . The cycle-to-cycle variability of a device is represented by the standard deviation around the ΔG mean. An alternative representation in the resistance space (ΔR vs. R) [24] is also possible. Here, we opt to work in $(G, \Delta G)$ -space since conductances map naturally to matrix weights in a neural network implemented with real ReRAM devices.

During training, a desired weight update is determined by the gradient of the loss, and programming pulses are applied to each device to adjust its conductance according to this update [18]. Due to ReRAM physics, conductance increases and decreases are not always symmetrical. Thus, two jump tables are needed – one for potentiation (increase in G), and another for depression (decrease in G), referred to as the SET and RESET tables respectively.

Jump tables are constructed from sampled data. For example, a dataset could consist of ReRAM measurements of n experimentally sampled points $(G_1, \Delta G_1), \dots, (G_n, \Delta G_n)$, where for a given i , G_i is the initial device conductance and ΔG_i is the conductance change resulting from a single pulse at a fixed voltage. Alternatively, one may construct a synthetic dataset by sampling an arbitrary number of $(G_i, \Delta G_i)$ data points from a given mean and standard deviation profile. Given this discrete data, various interpolation methods can be used to create a continuous model of the underlying mean and standard deviation profiles.

The binning jump table device modelling algorithm is outlined in Algorithm 1. The inputs are a dataset D comprising of $(G_i, \Delta G_i)$ pairs, and the number of bins $G_{\text{bins}}, \Delta G_{\text{bins}}$. The outputs are the binning model's mean and standard deviation profiles $\mu(G)$ and $\sigma(G)$. In brief, the $(G_i, \Delta G_i)$ data is grouped into bins of specified count. From each G bin in this binned data, the mean and standard deviation information for the ΔG_i are computed. The mean and standard deviation values are then linearly interpolated across the G bins to obtain the jump table model.

It has been observed that binning and interpolation can introduce unwanted errors and artifacts in modeling such as edge effects, empty bins, excessive smoothing, and rugged predictions of the underlying mean and standard deviation [25]. To avoid such artifacts, parameters G_{bins} and ΔG_{bins} for Algorithm 1 are optimized for each input dataset individually, in line with our previous work [26]. This is done to ensure that

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

comparisons between different binning models remain fair relative to one another.

Algorithm 1. Binning Jump Table Device Modelling

```

1  procedure BinningModel (in D, Gbins, ΔGbins; out μ, σ)
2    hist, edges = Hist2D(D, Gbins, ΔGbins) // 2D binning
3    Gaxis, ΔGaxis = GetMean(edges) // average G, ΔG per bin
4    C = []
5    for i in range(hist) do
6      C[i] = ToCDF(hist[i]) // calculate CDF(ΔG) per G bin
7    end for
8    μ, σ = fit(C, Gaxis, ΔGaxis) // Gaussian fit, interpolated
9  end procedure

```

Details of the Optuna model are presented separately in Section II E. To summarize, both the binning and Optuna modelling schemes attempt to predict the underlying mean and standard deviation profile of the input conductance data based on data interpolation.

B. Device Datasets

Cycle-to-cycle switching variability present in ReRAM devices is a non-ideality that can significantly hamper the performance of neural networks based on ReRAM devices [27]. Significant efforts in the community focus on engineering devices with lower variability [28] and providing algorithmic approaches to manage this device variability at the neural network level [20], [29]. For this work, we are focused only on exploring the impact of cycle-to-cycle variability on the modeling bias at the network level. Different formalisms exist in the literature to quantify device non-idealities. For example, [22] expresses cycle-to-cycle switching variation as an overall percentage of the underlying conductance range, [30] expresses this variability in terms of the ratio between the standard deviation and device resistance, and [31] proposes the use of a stochastic translator which uses a bitwise-AND operation in place of multiplication to capture this switching noise. We model the cycle-to-cycle variability by a random variable with distribution given by the CDF within a jump table.

The underlying distribution and noise profile are unknown in experimental data. Therefore, for the purposes of testing the proposed methodology, we use synthetic target distributions. Synthetic Gaussian datasets are generated with pre-determined underlying mean and standard deviation profiles $\mu(G)$ and $\sigma(G)$ over a given conductance range $G \in [G_{\min}, G_{\max}]$ with fixed minimum and maximum bounds. A synthetic jump table can be created either from these target mean and standard deviation profiles, called a *target jump table*, or from interpolated predictions of these profiles obtained through the binning model, referred to as a *binning jump table*.

In this work, we focus on a family of synthetic distributions for which the random variable representing ΔG per pulse has a mean profile $\mu(G)$ that is linear in G , and the standard deviation profile $\sigma(G)$ is a constant. This was chosen to match the experimentally derived “*Real Device*” model from the MLP+NeuroSim V3.0 framework [22], which we refer to as the *analytical device model* hereafter.

The jump table for this analytical model was synthesized using the theoretical equations from NeuroSim; we first set the

cycle-to-cycle variation (and all other device variations) to zero, keeping only the non-linearity values for potentiation and depression as 2.4 and -4.88 respectively (which are the default values for Ag:Si-based ReRAM model in NeuroSim). The maximum number of conductance levels was set to 500. We use this model to extract $(G_i, \Delta G_i)$ datapoints for potentiation as well as depression. As briefly already highlighted, the mean of this device model followed a linear profile, with a constant standard deviation profile. Through data fitting, it was found that the cycle-to-cycle variation parameter of the analytical model, C2C, related to the constant standard deviation profile, σ , of the jump table model as follows:

$$\sigma(G) = C2C \cdot (G_{\max} - G_{\min}), \quad (2)$$

where G_{\min} and G_{\max} are physical minimum and maximum bounds on the device conductance values, fixed at 3 nS and 38 nS respectively. In essence, the standard deviation relates to the cycle-to-cycle switching variation as a percentage of the conductance switching range.

With the extracted mean profile $\mu(G)$ and the standard deviation $\sigma(G)$ using (2), we can sample ΔG vs. G datasets for any given analytical device model containing an arbitrary number of points at a specific cycle-to-cycle variation.

Algorithm 2. Synthetic Data Generation

```

1  procedure GenerateSyntheticData (in n, G, μ, σ; out D)
2    f = LinearInterpolate(x = G, y = μ)
3    g = LinearInterpolate(x = G, y = σ)
4    for i in range [1, n] do
5      Gi = RandomUniformFloat(min(G), max(G))
6      μi = f(Gi), σi = g(Gi)
7      ΔGi = RandomNormalFloat(μi, σi2) satisfying (3)
8      D[i] = (Gi, ΔGi)
9    end for
10 end procedure

```

The synthetic device data generation algorithm used is outlined in Algorithm 2. The inputs are the number of points to synthesize n , conductance dynamic range G , and the mean and standard deviation profiles $\mu(G)$ and $\sigma(G)$. The output is a synthetic dataset D comprising $(G_i, \Delta G_i)$ pairs. Due to physical device constraints, for all pairs $(G_i, \Delta G_i)$ recorded experimentally, we also require that

$$G_{\min} \leq G_i + \Delta G_i \leq G_{\max}, \quad (3)$$

that is, the jump table may not push a device beyond its minimal or maximal conductances. For any conductance update that exceeded these bounds, the conductance is clipped to the corresponding extreme value in order to eliminate non-physical device behavior. Additional constraints may be imposed to further reflect physical realism.

Fig. 1a depicts the one-to-one matching between the conductance trajectories for both SET and RESET phases of the analytical model and the corresponding target jump table model. The Ag:Si analytical ReRAM model from NeuroSim has a cycle-to-cycle variation of 3.5 %, whereas our equivalent target jump table model has a standard deviation of 1.2 nS. Fig. 1b shows the overlapping raw conductance data of the two models, indicating that our synthetic model is able to generate

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

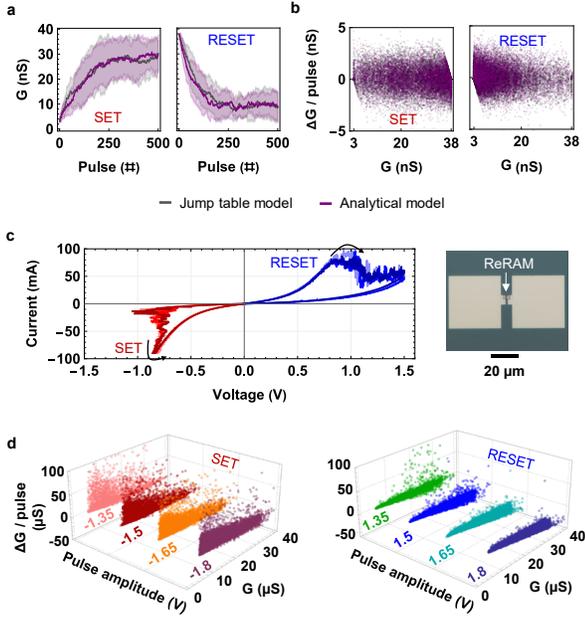


Fig. 1. Overview of device data used in our modelling. *Synthetic:* (a) Device conductance trajectories and (b) associated ΔG vs. G data of the analytical and target jump table device models at 3.5 % cycle-to-cycle variability and 1.2 nS standard deviation respectively. *Experimental:* (c) ReRAM devices used for experimental data gathering showing I-V characteristics, and (d) ΔG vs. G data at different pulse amplitudes.

$(G_i, \Delta G_i)$ data points in a dynamic range which matches the analytical model.

In this work, we also use experimental ΔG vs. G data from a 10 nm device with 2.5 nm Al_2O_3 / 15 nm TiO_x / 5 nm Ti / 30 nm Pt. The data were obtained utilizing a semiconductor parameter analyzer and pulse programming (Fig. 1c). Before programming, the devices were formed with monotonically increasing voltage pulses to reach a high conductance state. The data were collected after forming and then cycling between SET and RESET 30 times. A device is first programmed to a random conductance value within a given range. Then a write pulse voltage with 500 ns high time, 100 ns rise and fall time, and voltage amplitude randomly selected from among $\{\pm 1.35 \text{ V}, \pm 1.5 \text{ V}, \pm 1.65 \text{ V}, \pm 1.8 \text{ V}\}$ is applied. The conductance state is read via a subsequent pulse of 100 mV applied 100 μs after the write pulse; during the intermediary time, the device is held at 0 V. This process is repeated until either all pulses in the list have been applied or the device conductance exceeds some defined limit, then the algorithm restarts from a random conductance value. The end result is a separate $(G, \Delta G)$ dataset for every voltage amplitude (Fig. 1d). The exploration of pulse amplitudes is purposely random: typical approaches that use monotonically increasing or decreasing write pulse voltage steps may result in sparse data sets for high write voltages, since the device typically switches before the higher voltages are applied. By comparison, our resulting data sets for each write pulse voltage have roughly the same number of points

($\approx 10,000$). This is critical for the jump table modeling of this device data.

C. Neural Network Details

A 2-layer perceptron network with 324 neurons in the input layer, 50 neurons in the hidden layer, and 10 neurons in the output layer, with no biases, was used. The network was trained for image classification on a reduced version of the MNIST dataset, which consists of 20x20 pixel images (center-cropped from the original 28x28 pixel images) of handwritten digits labelled 0 through 9 [32]. Inputs to the network were first normalized by subtracting a constant (0.1307) and dividing the difference by a fixed scale factor (0.3801). This shift and scaling were determined as the mean and standard deviation of all pixels in the 60,000 images of the original, uncropped MNIST training set. A sigmoidal activation function was used for each hidden neuron, and a softmax activation was used for the output layer. Mean squared error was used as the loss function for training. The weights were initialized uniformly at random from $-\sqrt{k}$ to \sqrt{k} , where k is the reciprocal of the number of input features for a given layer (thus $k = \frac{1}{400}$ and $k = \frac{1}{10}$ for the input and output layers respectively.) Mini-batch gradient descent was used based on previous investigations [29], [33], [34] of its performance benefits when training a network with non-ideal synaptic weights, and the batch size was fixed at 4096. Furthermore, all network-level quantities such as layer weights, gradients, activations and errors were quantized to 6-bit fixed-point precision using the QPyTorch library [35]. The choice of network architecture, loss function, optimizer, and the quantization scheme are all motivated by the specifications of our future hardware prototype [36].

Fig. 2 presents background on how a feed forward neural network can be implemented using ReRAM devices in hardware, and on their modelling via jump tables. A single ReRAM device behaves like an artificial synapse, and ReRAM device crossbars can physically implement densely connected synaptic weight matrices, with weights w_{ij} of a network layer represented as device conductances G_{ij} in the memristive crossbar array. For inference, pixels of an input image are converted to voltages on the input wires of the ReRAM crossbar, as shown in Fig. 2a. As a result of resistive circuit physics (Ohm's and Kirchhoff's Laws), these voltages induce currents on the layer-output wires whose values are equal to the matrix-vector product of synaptic weights times input voltages. These currents are then converted into numerical values using an analog-to-digital conversion. The nonlinear activation function is computed using digital logic. This result is fed as a voltage applied to the inputs of the following network layer, and so on until the final output layer is reached.

For backpropagation, traditional minibatch gradient descent can be used to update the conductances of the crossbar array. For a conductance G_{ij} at minibatch iteration k , firstly the total number of pulses p_{ij} to be applied is determined by:

$$p_{ij} = p_{\max} \cdot \left(\nabla_w^{(k)} L \right)_{ij}, \quad (4)$$

where $\nabla_w^{(k)} L$ is the gradient of the loss function averaged over all minibatch inputs at iteration k , and p_{\max} is the device dynamic range, defined as the total number of pulses required

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

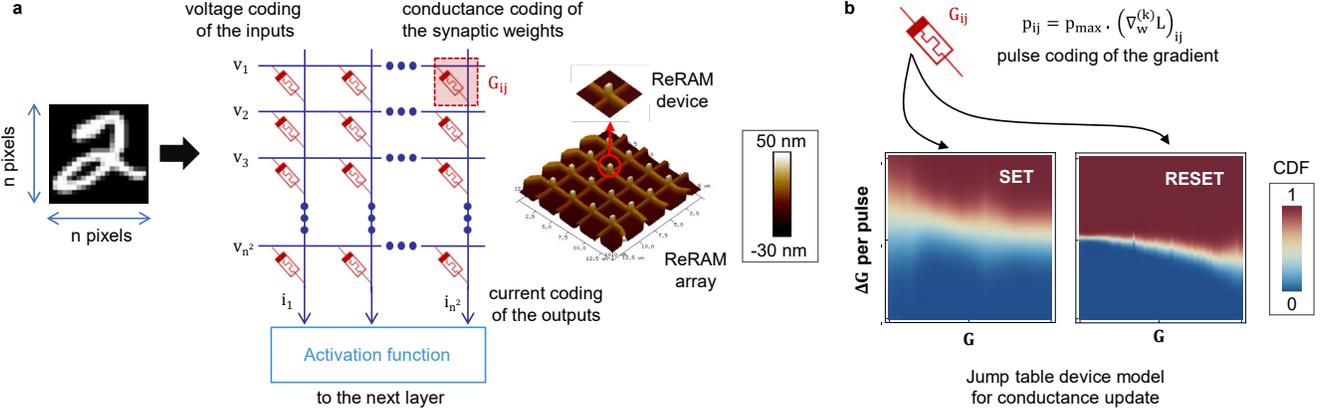


Fig. 2. Inference and training in a ReRAM-based multi-layer perceptron network. (a) *Inference*: ReRAM conductances implement matrix-vector multiplication in hardware as a consequence of electrical circuit physics (Ohm’s and Kirchoff’s Laws). Atomic Force Microscopy picture of a representative crossbar matrix of individual ReRAM devices is also shown. (b) *Training*: For all ReRAM devices in the network, the derivative of the loss function with respect to the synapse is converted to the number of voltage pulses to be applied to the corresponding device. The change in conductance resulting from these pulses is modelled by two jump tables, one for SET and one for RESET, samples of which are also shown.

to either SET the device from G_{\min} to G_{\max} , or alternatively, RESET the device from G_{\max} to G_{\min} . In hardware, the p_{ij} pulses would physically be applied to the conductance G_{ij} directly. However, in our jump table training regime, the following algorithm is employed to update G_{ij} due to the train of applied pulses p_{ij} : A random number u_k is generated between 0 and 1. The corresponding conductance change for a pulse is the inverse of the CDF for a device at conductance G_{ij} applied to u , $\Delta G = \text{CDF}^{-1}(u)$. The device conductance is updated to $G_{ij} + \Delta G$ and this cycle is repeated for all pulses. These jump table conductance updates intrinsically quantize and model the conductance update behavior of real devices.

We implemented jump table-based neural network training in PyTorch [37]. In our simulations, the synaptic weights for each network layer are represented in the conductance space as a crossbar array via the CDF given by a pair of SET and RESET jump tables (Fig. 2b). While the scope of our work is modeling neural networks based on ReRAM devices, our framework can easily be extended to model other emerging analog memory technologies such as phase change memories, magnetic tunnel junctions, or photonic modulators [18], [38]. Additionally, this work is entirely focused on training and investigating modeling bias using device models during training. Once programmed, the ReRAM device weight behaves like a fixed and relatively stable circuit element, so any small read noise and relaxation effects were considered negligible for the inference phase.

For all target models, a learning rate optimization was implemented using a grid-search. In brief, training was carried out using learning rates sampled on a logarithmic scale covering a wide range. The optimal learning rate was defined as that which gave the highest test accuracy averaged over the first ten training epochs. Subsequent network training simulations were computed over 100 epochs with the learning rate set to this value in order to isolate the effects of jump table models on network performance.

D. Metrics for Evaluating Modeling Quality

In order to quantify the quality of the device model in the context of the neural network training, the following metrics are employed at the network and device levels.

D.1. Modeling Bias

Given that ReRAM conductance updates have a stochastic character, any single training simulation and its attendant accuracy should be viewed as an instance of a random variable. Different models for the ReRAM devices have different distributions in accuracy, and it is sensible to compare their means. We define modeling bias (MB) as the amount that the network’s average modelled test accuracy ($\text{Mean Acc.}_{\text{model}}$) – the accuracy when the synaptic weights are represented by ReRAM devices using an interpolated jump table model – differs from the average target test accuracy ($\text{Mean Acc.}_{\text{target}}$). These averages are computed for each epoch as shown in (5). Note that the target test accuracy can have a different meaning based on the type of data we used for modeling. For synthetic data, it is the accuracy when the synaptic weights are modelled by ReRAM devices using the synthesized target jump table model. For experimental data, it is the experimental test accuracy that would be obtained in a physical implementation based on the real behavior of the device population used for creating the jump table models.

Since modeling bias is a function of the training epoch, a good device model is one that has lower modeling bias throughout training. Formally, we define

$$\text{MB}(i) = \text{Mean Acc.}_{\text{model}}(i) - \text{Mean Acc.}_{\text{target}}(i) \quad (5)$$

as a function of epoch number i . The bias can either be positive, when the modelled accuracy is too optimistic and overpredicts the target accuracy, or negative, when the modelled accuracy underpredicts it. For the results shown below, mean network accuracies required by (5) were determined by averaging over 20 instances of network training for a given device model.

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

D.2. Switching Sign Discrepancy and Overlapping Error

To estimate the quality of a device model, we propose the use of two device-level error metrics: switching sign discrepancy (SSD) and overlapping error (OVLE). At the network level, the switching direction of ReRAM devices plays a pivotal role in effective training. As an example, suppose the target mean values for a given jump table model were all positive across the entire range of the G parameter with minimal variance, as is the case with some of our SET tables. If an interpolated model regularly predicts negative ΔG updates for many of the corresponding G indices, the result would be a weight that is changed in the opposite direction of what is prescribed by the corresponding gradient. These updates with inverted signs would almost certainly lead to a degradation in training performance. The same argument applies to the RESET tables. To measure the presence of this failure mechanism, we define the switching sign discrepancy (SSD) as:

$$\text{SSD}(x, y) = \frac{1}{n} \sum_{i=1}^n |P(X_{G_i} < 0) - P(Y_{G_i} < 0)|, \quad (6)$$

where X_{G_i} and Y_{G_i} are distributions of device models x and y respectively, n is the total number of distributions within these models (assumed to be the same for both models), and $P(X_{G_i} < 0)$ is the probability that the i -th distribution of model x produces a negative variate. For our jump table models with Gaussian distributions, this is computed as:

$$P(X_{G_i} < 0) = \int_{-\infty}^0 \text{PDF}_{X_{G_i}} dG = \frac{1}{2} \text{erfc}\left(\frac{\mu_i}{\sigma_i \sqrt{2}}\right), \quad (7)$$

where $\text{PDF}_{X_{G_i}}$ is the probability density function of X_{G_i} , $\text{erfc}(\cdot)$ is the complementary error function, and μ_i, σ_i are the mean and standard deviation of X_{G_i} . The SSD quantifies the average signed agreement between two device models in terms of their switching behavior. The SSD of a device model with itself is 0 since the agreement is maximum and is 1 when there is no signed switching agreement between two models.

Along similar lines, the overall overlap between device model distributions is also of importance. For this, we make use of the complement of the overlapping coefficient from [39] and define the overlapping error (OVLE) between two device models x and y as follows:

$$\text{OVLE}(x, y) = \frac{1}{2n} \sum_{i=1}^n \int_{-\infty}^{+\infty} |(\text{PDF}_{X_{G_i}} - \text{PDF}_{Y_{G_i}})| dG. \quad (8)$$

The OVLE of a device model with itself is 0 since the overlap is maximum and is 1 in the case of no overlap between two models.

These device metrics can potentially present insights about the modeling bias performance of a given device model, prior to carrying out expensive neural network experiments, and can thus save significant time. Additionally, these metrics can also serve as a tool for device designers to quickly quantify disagreement (or agreement) between different devices at the level of device switching. As an example, different jump table models could be extracted from experimental $(G_i, \Delta G_i)$ datasets obtained from different devices, and these error metrics could be computed and compared across device populations.

D.3. Kolmogorov-Smirnov test

For experimental data where the underlying statistics are unknown, the SSD and OVLE cannot be directly calculated. The two-sample Kolmogorov-Smirnov test (or a similar non-parametric two sample test) can be used instead. The Kolmogorov-Smirnov test can provide a probabilistic estimate of two datasets being drawn from the same underlying distribution. The test value reports the largest absolute difference D between two distribution functions $I(x)$ and $J(x)$ as given by (9).

$$D = \sup_x |I(x) - J(x)|, \quad (9)$$

where \sup denotes the supremum. A lower test value corresponds to a better fit, meaning that the probability of the two datasets having been drawn from the same distribution is higher. Since our experimental device data is two-dimensional $(G_i, \Delta G_i)$, we use a two-dimensional variant of the Kolmogorov-Smirnov test, as specified in [40]. Our methodology is as follows: First, the experimental data is transformed to reflect a compliance current to limit switching between G_{\min} and G_{\max} using (3). Second, the data is randomly split into non-overlapping modeling and testing subsets. The modelling subset is used to generate binning and Optuna jump table models. We then use these models to synthesize a $(G_j, \Delta G_j)$ dataset. Finally, we apply the Kolmogorov-Smirnov test on the initial testing subset and each of the two synthesized datasets. As a baseline for comparison, we additionally report the Kolmogorov-Smirnov test values between the two experimental subsets in our results.

E. Optuna modelling

We propose a method to explore the network modeling bias in the absence of experimental apparatus and/or experimental network accuracies. Ideally, the experimental data includes a complementary set of experimental device $(G_i, \Delta G_i)$ datapoints and the corresponding network train / test accuracies obtained from a prototype network physically implemented with these devices. In such case, different device modeling methods, their device- and network-level metrics, and respective network simulation results could be tested against the experimental reality directly to determine if the device model needs to be further optimized. However, prototyping neural networks in emerging hardware is challenging, expensive and time consuming. Therefore, for a given emerging device technology, only experimental device $(G_i, \Delta G_i)$ data might be available to benchmark both the device model and the network simulation results. Nevertheless, the question remains if the device model utilized is sufficient to obtain a realistic network accuracy or if there is space for optimization in the device model.

To address this challenge, this paper presents a device model optimization methodology based on Optuna [21], an automated hyperparameter optimization framework. The algorithm employs a derivative-free optimization strategy to push the limits of device modeling towards experimental realism. As an optimization metric, the Kolmogorov-Smirnov test is used. Algorithm 3 details our developed Optuna algorithm. The input is an experimental dataset D comprising of $(G_i, \Delta G_i)$ pairs, conductance range G , polynomial degrees m and n to use for modelling mean and standard deviation profiles respectively, objective function f , and a termination value e to

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

be optimized. In our case, the objective function is the two-dimensional two sample Kolmogorov-Smirnov error between the testing subset of the experimental dataset, and synthesized Optuna dataset. However, this can easily be replaced with an alternate metric as needed. e is chosen as the binning test error, which is the Kolmogorov-Smirnov error between the testing subset of the experimental dataset and a synthesized binning dataset. The outputs are Optuna mean and standard deviation profiles $\mu'(G)$ and $\sigma'(G)$, the test error of which is better than the corresponding binning model by iterative optimization.

Algorithm 3. Optuna Jump Table Device Modelling

```

1  procedure OptunaModel (in D, G, m, n, f, e; out  $\mu'$ ,  $\sigma'$ )
2     $e' = \infty$  (Initialize Optuna error as inf.)
3     $\mu_{\text{params}} \leftarrow$  Initialize m + 1 coefficients for  $\mu(G)$  profile
4     $\sigma_{\text{params}} \leftarrow$  Initialize n + 1 coefficients for  $\sigma(G)$  profile
5    while  $e' \geq e$  do
6       $\mu' = \text{EvaluatePolynomial}(G, \mu_{\text{params}})$ 
7       $\sigma' = \text{EvaluatePolynomial}(G, \sigma_{\text{params}})$ 
8      GenerateSyntheticData(in length(D), G,  $\mu'$ ,  $\sigma'$ ; out D')
9       $e' = f(D', D)$ 
10      $\mu_{\text{params}} \leftarrow$  Update m + 1 coefficients for  $\mu(G)$  profile
11      $\sigma_{\text{params}} \leftarrow$  Update n + 1 coefficients for  $\sigma(G)$  profile
12   end while
13 end procedure

```

In essence, Optuna iteratively learns the mean and standard deviation profiles of an input experimental dataset consisting of $(G_i, \Delta G_i)$ points recorded at a certain voltage by fitting two separate polynomials (of degrees m and n) on the input dataset – one for the mean profile over G , and another for the standard deviation profile over G . The $m + 1$ coefficients for the mean profile, and the $n + 1$ coefficients for the standard deviation profile serve as hyperparameters for optimization. Since the

algorithm is iterative, there is no guaranteed upper-bound on its time complexity. In comparison with binning, it is fairly more time-consuming in terms of runtime because of having to sample from a mn -dimensional space a large number of times.

For our datasets, hyperparameters were initialized based on direct polynomial fitting over binning mean and standard deviation profiles. Successive sampling was done by the tree-structured parzen estimator – a Bayesian optimization strategy. For all RESET voltage pulse amplitudes from our experimental datasets, the termination criterion was achieved using parabolic modelling of the mean and standard deviation profiles, and for all SET voltages, it was achieved using linear modelling of the mean and standard deviation profiles.

III. RESULTS

A. Impact of number of data points

Identifying the optimal number of $(G, \Delta G)$ data points for jump table modeling is of significant interest to provide a highly representative simulation result at the network level, while keeping the device measurement time to a minimum.

To study this, we generated series of jump tables following the linear mean, constant standard deviation profiles from Fig. 2 (equivalent to 3.5% cycle-to-cycle variability of the analytical device model) with an increasing number of points in the raw dataset, ranging from 40 to a total of 10,000 data points. Neural network simulations using these jump table models were run at their respective optimized hyperparameters and the modeling bias as well as SSD and OVLE metrics were determined.

The results in Fig. 3 show a potential correlation between the quality of the device model and the quality of the network simulation based on this device model. As the number of data points increases, conductance trajectories of the binning models align more with the target model (Fig. 3a, b). Additionally, device-level metrics SSD and OVLE of the binning models

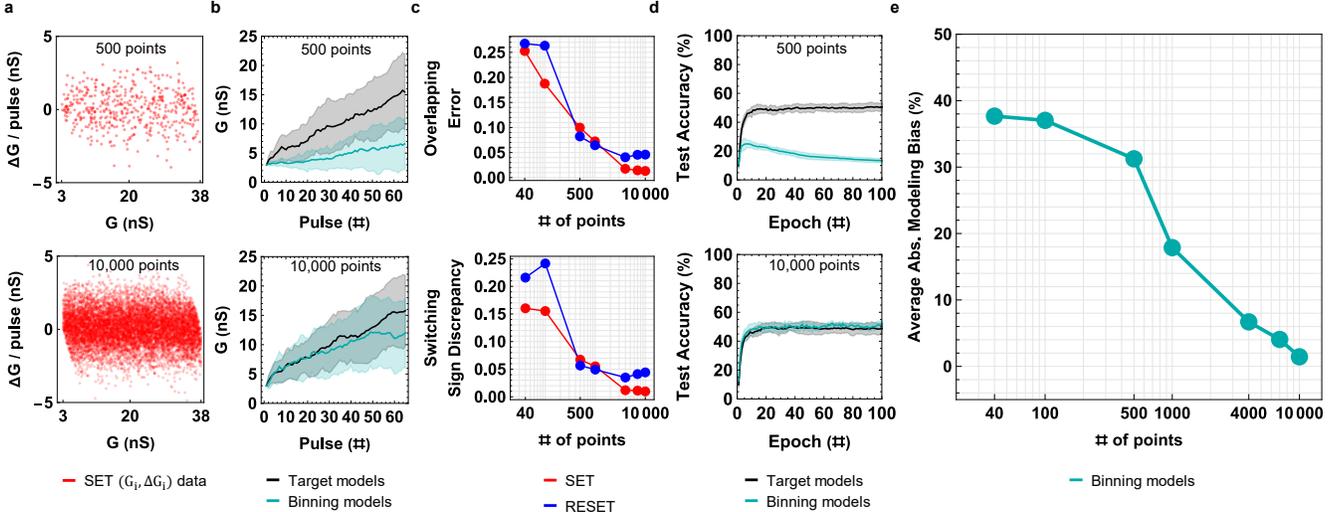


Fig. 3. Impact of the number of points for device modelling and respective network simulations. (a) Synthesized $(G, \Delta G)$ datasets for SET modelling with 500 and 10,000 points, and (b) corresponding conductance trajectories of the target device model and interpolated binning model for up to 64 applied pulses. (c) Overlapping Error and Switching Sign Discrepancy of different binning jump table models. (d) Network training curves of the 500 point and 10,000-point models, and (e) average absolute modeling bias of the different binning jump table models, indicating that the modeling bias performance improves as the number of points increase. The batch size used is 4096 and optimized network learning rate is 0.1.

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

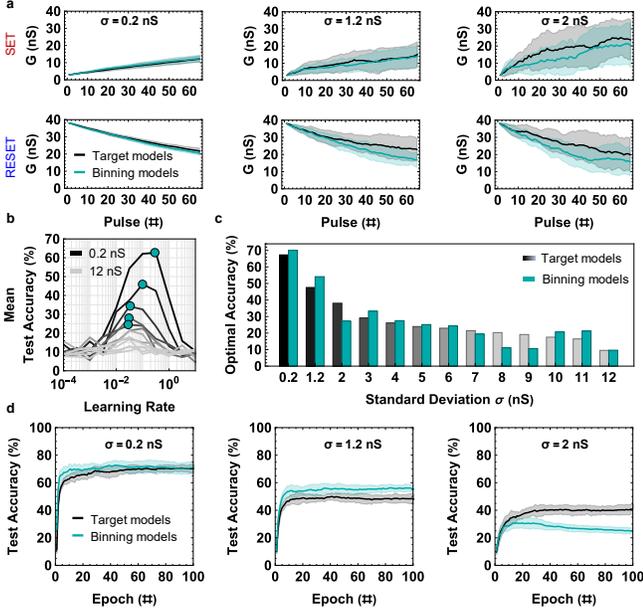


Fig. 4. Impact of the cycle-to-cycle variability on device modelling and respective network simulations. (a) Conductance trajectories of target and binning jump table models at different levels of standard deviation for up to 64 applied pulses. (b) Training windows of target models. Cyan markers indicate optimal learning rates. (c) Comparative optimal accuracies of target and binning models. Traces in (b) correspond to target models of the same shade in (c). (d) Convergence curves corresponding to the device models in (a).

both decrease, and so does the modeling bias of the corresponding binning jump table models (Fig. 3c, d). This occurs because the binning algorithm is able to more effectively capture the underlying target mean and standard deviation profiles given rich data, leading to overall good network-level estimation of the target models (Fig. 3e).

There is minimal benefit in terms of modeling bias upon increasing the points from 4,000 to 10,000. The remainder of the jump tables generated in this work are thus based on 4,000 samples of $(G_i, \Delta G_i)$ points, where these samples are drawn from target Gaussian distributions with known mean and standard deviation profiles in the synthetic case and are experimental measurements when working with real devices.

B. Impact of Cycle-to-Cycle Variability

To investigate how the device cycle-to-cycle variability impacts the device modeling and network simulation respectively, we synthesized variants based on the analytical device model at increasing values of the underlying standard deviation, sampled in a range from 0.2 nS to 12 nS, corresponding roughly to cycle-to-cycle variations from 0.5 % to 34 % of the underlying conductance range. The mean profile was not altered, and the number of $(G, \Delta G)$ data points was fixed at 4,000.

Fig. 4 presents an analysis of the overall performance of these models. As the cycle-to-cycle variability of target devices increases, the target conductance trajectories become noisier. As a consequence, the corresponding binning models find it

harder to predict the underlying mean and standard deviation profiles. This manifests in binning conductance trajectories deviating from target trajectories (Fig. 4a).

At the network level, the increase in cycle-to-cycle variability translates to an overall degradation of the training capability of target devices, evident from narrower learning rate training windows in Fig. 4b and the decline in optimal target accuracies in Fig. 4c. Additionally, it can be observed that the optimal learning rate values – shown in cyan markers in Fig. 4b – inversely decrease with the increase in device cycle-to-cycle variability. This suggests that in order to optimally train a ReRAM-based hardware network, smaller gradient steps proportional to the cycle-to-cycle variability of the underlying ReRAM devices may be required. Fig. 4a also highlights the importance of performing a learning rate optimization. This is because modeling bias results at unoptimized learning rates can be potentially misleading.

Finally, network simulations of binning models show that the network modeling bias tends to increase with the increase in noise, as captured in Fig. 4d. It can also be observed that the network results are largely unreliable in terms of the sign of modeling bias – with some cases over-estimating target accuracy, and others under-estimating target accuracy. Overall, we conclude that the binning algorithm fails to effectively capture target mean and standard deviation profiles when the standard deviation in the input dataset is high given a fixed number of data points, i.e., beyond 2 nS given 4,000 $(G, \Delta G)$ data points.

C. Impact of Non-Linearity

The impact of the device non-linearity on modeling bias is investigated using variants of the analytical model at different device non-linearity values. The chosen parametrization for non-linearity is formulated as follows: given a base jump table model with mean profile $\mu(G)$, the corresponding model at non-linearity k – where k is a positive integer – is the model with mean profile $k\mu(G) = \{k\mu_1, k\mu_2, \dots, k\mu_n\}$. Note that the standard deviation profile is unaltered, though the multiplication of the mean profile by a constant can lead to a change in the net coefficient of variation (defined as dispersion around the mean). Fig. 5 summarizes our analysis on device non-linearity.

Fig. 5a shows the conductance trajectories of target and binning models at different non-linearities. Contrary to Fig. 4a where target devices get noisier, here we see that target devices show less noise under our chosen scheme of increasing non-linearity. This is due to the net coefficient of variation decreasing with the increase in device non-linearity. Additionally, it can be observed that binning modelling is able to better model target devices with high non-linearity and low cycle-to-cycle variability, compared to target devices that have low non-linearity and high cycle-to-cycle variability. This improvement is manifested in the OVLE and SSD device metrics shown in Fig. 5b, which are consistently significantly lower compared to devices from Fig. 3c. As the non-linearity increases, we make the following two observations at the network level. Firstly, the optimal accuracy of target models increases. Secondly, the modeling bias performance of corresponding binning models improves, as captured in the thinning gray region in Fig. 5c. The convergence curves in Fig.

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

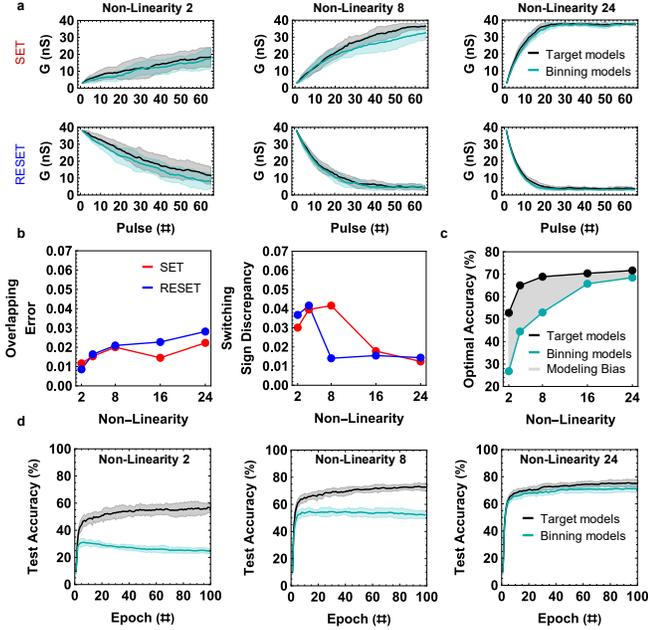


Fig. 5. Impact of device non-linearity on device modelling and respective network simulations. (a) Conductance trajectories of target and binning jump table models with increasing non-linearity for up to 64 applied pulses. At higher non-linearity, devices saturate faster. (b) Overlapping error and switching sign discrepancy of the binning models. (c) Comparative optimal accuracies at different levels of device non-linearity. (d) Corresponding optimized network training curves at different levels of non-linearity.

5d reinforce these observations. Our device metrics do not seem to fully capture modeling bias characteristics, contrary to Fig. 3 where target profiles were unaltered. We hypothesize that this happens because of significant changes to switching behavior from one non-linearity model to another, and acknowledge that our device metrics need to be further refined to better capture device switching intricacies.

D. Experimental Model Verification

While prior sections investigated synthetic datasets, this section summarizes the modeling results on experimental data. We use our experimental datasets (Fig. 1c, d) to derive jump table models using the approach described in Section II D.3.

Fig. 6a shows Kolmogorov-Smirnov test error values for the different SET and RESET datasets for the binning and Optuna models. To illustrate the room for optimization in these models, we also plot the test error between the experimental model and experimental test sub-sets. These values are the lowest as expected, as the two are indeed drawn from exactly the same distribution. The subsampling of the experimental datasets, the binning modelling, and the error calculations were repeated for a total of 20 iterations for maintaining statistical significance. An error bar indicates one standard deviation.

There is a performance gap between what the binning models can do to approximate the experimental data, as observed by the gap between the cyan and gray points in Fig. 6a. This is line with the results from Fig. 4. Our proposed Optuna-based algorithm for optimization produces $(G, \Delta G)$ data is significantly closer to the experimental reality compared to

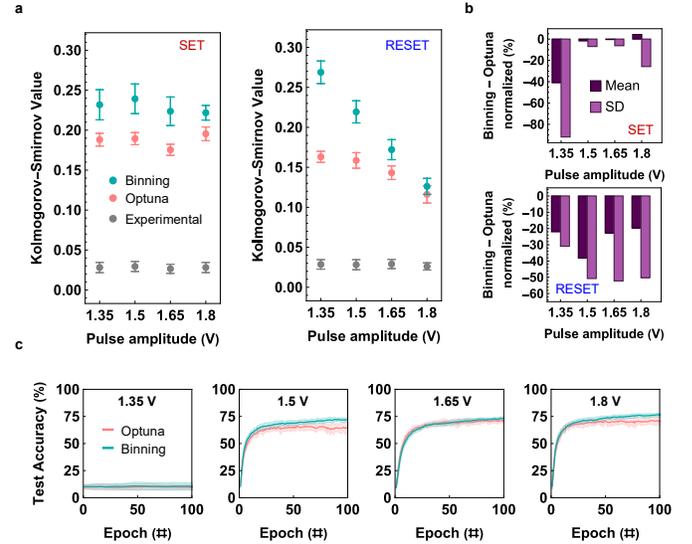


Fig. 6. Experimental Model Verification. (a) Kolmogorov-Smirnov test errors of the experimental test dataset with the experimental model dataset and synthesized datasets from the binning and the Optuna-optimized binning models. (b) Normalized difference (%) between mean and standard deviation profiles of the binning models with the Optuna models. (c) Optimized network training curves investigated pulse amplitudes.

the binning models in all investigated cases, with lower K-S values. Since Optuna models still exhibit a performance gap from the experimental data, better interpolation methods still need to be explored. Nevertheless, network simulations of the Optuna models seem to represent experimental reality closer than corresponding binning models. End-to-end experimental verification is left for future work.

Fig. 6b shows the normalized difference of the mean and standard deviation profiles between binning and Optuna models. In all cases, it can be observed that differences in the standard deviation profile are higher than the differences in the mean profile. A negative bar links to a binning profile underestimating the corresponding Optuna profile, and it can be seen that binning tends to generally underestimate the standard deviation. This underestimation of the device noise directly manifests in the network convergence curves shown in Fig. 6c, where binning models over-promise the test accuracy that would be obtained by a ReRAM network implemented using our devices compared to Optuna models.

IV. DISCUSSION

These results point to the need to understand the goodness of a device model in connection with the experimental reality of ReRAM hardware for in-memory compute and neural network accelerators. In hardware implementations, the obtained accuracy depends on the underlying distribution of the device population used for implementation. While this underlying distribution is unknown and can only be approximated indirectly via electrical measurements, a good statistical model of the device based on these measurements should be able to predict the target accuracy that a hardware prototype would

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

achieve – if available. The value of the accuracy is not of critical consideration here per se; the modeling bias, defined as the difference between simulated accuracy and target accuracy, is more relevant. For example, a software simulation based on a given device model may achieve an accuracy higher than the target model, supposing such a hardware model exists for comparison. This outcome would be undesirable since the simulation results overpromise on the particular device technology proposed. For our investigation at the network level, we are only interested in how well a statistical model can approximate the underlying target distribution of the device population, which does not have to be the model that achieves the highest network training accuracy.

To support these co-design efforts in the absence of network experimental data, this work has proposed the use of automatic optimizers such as Optuna to estimate network convergence as the benchmark for estimating the modeling bias performance of corresponding experimentally derived device models. While this approach does not replace the need of experimentally obtaining network performance estimates and studying exact device performance, it could be used as a short-term solution to support iterative device and network co-optimization, particularly when experimental results of network training might be difficult to obtain.

We have shown that there are limitations to binning models in approximating synthetic target model convergence. Specifically, binning models exhibit poor device modeling and a high modeling bias when a) the input dataset is scarce – having fewer than 4,000 data points, and b) when the device cycle-to-cycle variability is high – beyond 2 nS in our chosen network training regime. These limitations mean that exploring other interpolation methods in the context of jump table-based device modelling is promising. Our prior work in device modeling with Gaussian Process Regression methods, such as ordinary Kriging [26] hints at a potential method that will be investigated in the future for network training.

It is important to point out that this initial investigation is restricted to Gaussian interpolation methods applied mostly to linear mean and constant standard deviation profiles. In reality, emerging devices can follow various profiles, e.g. piecewise linear profile [20], parabolic mean and constant standard deviation [19], etc. Moreover, these Gaussian interpolation methods can be insufficient if the underlying data is skewed due to physical bounds and limitations based on the parasitic resistance of the device. More general interpolation methods, e.g., based on skew normal distribution, could better predict complex device behavior and possibly neural network behavior more accurately and consistently. We plan to explore in future work how various interpolation modeling methods compare to one another in terms of corresponding network modeling bias where the underlying distributions are complex and non-trivial.

Since our device metrics and modeling bias results both monotonically decrease (e.g., Fig. 3), we hypothesize that these metrics could be the first step towards having a mechanism to predict modeling bias performance of different device models relative to one another. However, before such a link could be established, these metrics need to be further investigated and refined to fully understand their connection with modeling bias. Network parameters such as the learning rate, loss function, and

network dimensionality would have to be studied in conjunction with device training statistics such as the number of times SET and RESET tables are used over training, number of SET or RESET pulses applied, etc. This is left for future work.

V. CONCLUSIONS

This paper proposes the concept of “modeling bias” as a useful metric to quantify the goodness of a device model at the neural network level. To exemplify this concept, the binning interpolation method was used for modeling ReRAM device jump tables and its applicability in predicting neural network convergence behavior was investigated. For testing, a wide range of synthetic Gaussian datasets with linear mean and standard deviation profiles as well as experimentally obtained datasets was used, in conjunction with metrics at the device and network level were explored. The results show that device models based on binning can lead to unreliable modeling bias behavior, sometimes over-promising and sometimes under-promising the network accuracy. Better interpolation methods that have lower bias than binning, particularly at high device switching noise and low number of data points will be investigated in the future. Promising device metrics seem to have a similar trend with modeling bias at the network level, but additional investigations are needed for more complex device switching profiles and more difficult image classification datasets. Additionally, simulated accuracy results will be experimentally verified. Finally, the proposed Optuna algorithm for device modelling will be further refined by exploring other, more application-friendly error metrics beyond the Kolmogorov-Smirnov test. This work highlights the need for additional correlated efforts in device / network modeling and prototyping for in-memory compute and neural networks.

ACKNOWLEDGMENT

We thank Mark Anders and Lin Wang for prior useful discussions. The authors acknowledge the use of high-performance computing clusters, advanced support from the research technology services, and IT support at The George Washington University and NIST.

REFERENCES

- [1] P. Bose, “Power Wall,” in *Encyclopedia of Parallel Computing*, D. Padua, Ed. Boston, MA: Springer US, 2011, pp. 1593–1608. doi: 10.1007/978-0-387-09766-4_499.
- [2] S. Channamadhavuni, S. Thijssen, S. K. Jha, and R. Ewetz, “Accelerating AI Applications using Analog In-Memory Computing: Challenges and Opportunities,” in *Proceedings of the 2021 on Great Lakes Symposium on VLSI*, Virtual Event USA, Jun. 2021, pp. 379–384. doi: 10.1145/3453688.3461746.
- [3] N. Verma, H. Jia, H. Valavi, Y. Tang, M. Ozatay, L.-Y. Chen, B. Zhang, and P. Deaville, “In-Memory Computing: Advances and Prospects,” *IEEE Solid-State Circuits Magazine*, vol. 11, no. 3, pp. 43–55, 2019, doi: 10.1109/MSSC.2019.2922889.
- [4] Y. Chen, “ReRAM: History, Status, and Future,” *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1420–1433, Apr. 2020, doi: 10.1109/TED.2019.2961505.
- [5] M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, Q. Xia, and J. P. Strachan, “Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine,” *Advanced Materials*, vol. 30, no. 9, p. 1705914, 2018, doi: 10.1002/adma.201705914.

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

- [6] C.-X. Xue, T.-Y. Huang, J.-S. Liu, T.-W. Chang, H.-Y. Kao, J.-H. Wang, T.-W. Liu, S.-Y. Wei, S.-P. Huang, W.-C. Wei, Y.-R. Chen, T.-H. Hsu, Y.-K. Chen, Y.-C. Lo, T.-H. Wen, C.-C. Lo, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, "15.4 A 22nm 2Mb ReRAM Compute-in-Memory Macro with 121-28TOPS/W for Multibit MAC Computing for Tiny AI Edge Devices," in *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, Feb. 2020, pp. 244–246. doi: 10.1109/ISSCC19947.2020.9063078.
- [7] F. Cai, J. M. Correll, S. H. Lee, Y. Lim, V. Bothra, Z. Zhang, M. P. Flynn, and W. D. Lu, "A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations," *Nat Electron*, vol. 2, no. 7, pp. 290–299, Jul. 2019, doi: 10.1038/s41928-019-0270-x.
- [8] A. Valentian, F. Rummens, E. Vianello, T. Mesquida, C. L.-M. de Boissac, O. Bichler, and C. Reita, "Fully Integrated Spiking Neural Network with Analog Neurons and RRAM Synapses," in *2019 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2019, p. 14.3.1-14.3.4. doi: 10.1109/IEDM19573.2019.8993431.
- [9] S. Menzel, "Comprehensive modeling of electrochemical metallization memory cells," *J Comput Electron*, vol. 16, no. 4, pp. 1017–1037, Dec. 2017, doi: 10.1007/s10825-017-1051-2.
- [10] S. Aldana, P. García-Fernández, R. Romero-Zalaz, M. B. González, F. Jiménez-Molinos, F. Gómez-Campos, F. Campabadal, and J. B. Roldán, "Resistive switching in HfO₂ based valence change memories, a comprehensive 3D kinetic Monte Carlo approach," *J. Phys. D: Appl. Phys.*, vol. 53, no. 22, p. 225106, Apr. 2020, doi: 10.1088/1361-6463/ab7bb6.
- [11] E. Abbaspour, S. Menzel, and C. Jungemann, "Studying the switching variability in redox-based resistive switching devices," *J Comput Electron*, vol. 19, no. 4, pp. 1426–1432, Dec. 2020, doi: 10.1007/s10825-020-01537-y.
- [12] A. Padovani, D. Z. Gao, A. L. Shluger, and L. Larcher, "A microscopic mechanism of dielectric breakdown in SiO₂ films: An insight from multi-scale modeling," *Journal of Applied Physics*, vol. 121, no. 15, p. 155101, Apr. 2017, doi: 10.1063/1.4979915.
- [13] C. Bengel, A. Siemon, F. Cüppers, S. Hoffmann-Eifert, A. Hardtdegen, M. von Witzleben, L. Hellmich, R. Waser, and S. Menzel, "Variability-Aware Modeling of Filamentary Oxide-Based Bipolar Resistive Switching Cells Using SPICE Level Compact Models," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 12, pp. 4618–4630, Dec. 2020, doi: 10.1109/TCSI.2020.3018502.
- [14] C. Yakopcic, T. M. Taha, G. Subramanyam, and R. E. Pino, "Memristor SPICE model and crossbar simulation based on devices with nanosecond switching time," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, Aug. 2013, pp. 1–7. doi: 10.1109/IJCNN.2013.6706773.
- [15] G. Bersuker, D. C. Gilmer, D. Veksler, P. Kirsch, L. Vandelli, A. Padovani, L. Larcher, K. McKenna, A. Shluger, V. Iglesias, M. Porti, and M. Nafria, "Metal oxide resistive memory switching mechanism based on conductive filament properties," *Journal of Applied Physics*, vol. 110, no. 12, p. 124518, Dec. 2011, doi: 10.1063/1.3671565.
- [16] D. Ielmini, F. Nardi, and C. Cagli, "Physical models of size-dependent nanofilament formation and rupture in NiO resistive switching memories," *Nanotechnology*, vol. 22, no. 25, p. 254022, May 2011, doi: 10.1088/0957-4484/22/25/254022.
- [17] G. González-Cordero, J. B. Roldan, F. Jiménez-Molinos, J. Suñé, S. Long, and M. Liu, "A new compact model for bipolar RRAMs based on truncated-cone conductive filaments—a Verilog-A approach," *Semicond. Sci. Technol.*, vol. 31, no. 11, p. 115013, Oct. 2016, doi: 10.1088/0268-1242/31/11/115013.
- [18] S. Sidler, I. Boybat, R. M. Shelby, P. Narayanan, J. Jang, A. Fumarola, K. Moon, Y. Leblebici, H. Hwang, and G. W. Burr, "Large-scale neural networks implemented with Non-Volatile Memory as the synaptic weight element: Impact of conductance response," in *2016 46th European Solid-State Device Research Conference (ESSDERC)*, Lausanne, Switzerland, Sep. 2016, pp. 440–443. doi: 10.1109/ESSDERC.2016.7599680.
- [19] M. J. Marinella, S. Agarwal, A. Hsia, I. Richter, R. Jacobs-Gedrim, J. Niroula, S. J. Plimpton, E. Ipek, and C. D. James, "Multiscale Co-Design Analysis of Energy, Latency, Area, and Accuracy of a ReRAM Analog Neural Training Accelerator," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 1, pp. 86–101, Mar. 2018, doi: 10.1109/JETCAS.2018.2796379.
- [20] G. W. Burr, R. M. Shelby, S. Sidler, C. di Nolfo, J. Jang, I. Boybat, R. S. Shenoy, P. Narayanan, K. Virwani, E. U. Giacometti, B. N. Kurdi, and H. Hwang, "Experimental Demonstration and Tolerancing of a Large-Scale Neural Network (165 000 Synapses) Using Phase-Change Memory as the Synaptic Weight Element," *IEEE Transactions on Electron Devices*, vol. 62, no. 11, pp. 3498–3507, Nov. 2015, doi: 10.1109/TED.2015.2439635.
- [21] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, Jul. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.
- [22] Y. Luo, X. Peng, and S. Yu, "MLP+NeuroSimV3.0: Improving On-chip Learning Performance with Device to Algorithm Optimizations," in *Proceedings of the International Conference on Neuromorphic Systems*, New York, NY, USA, Jul. 2019, pp. 1–7. doi: 10.1145/3354265.3354266.
- [23] J. Niroula, S. Agarwal, R. Jacobs-Gedrim, R. L. Schiek, D. Hughart, A. Hsia, C. D. James, and M. J. Marinella, "Piecewise empirical model (PEM) of resistive memory for pulsed analog and neuromorphic applications," *J Comput Electron*, vol. 16, no. 4, pp. 1144–1153, Dec. 2017, doi: 10.1007/s10825-017-1107-3.
- [24] S. Stathopoulos, A. Serb, A. Khiat, M. Ogorzałek, and T. Prodromakis, "A Memristive Switching Uncertainty Model," *IEEE Transactions on Electron Devices*, vol. 66, no. 7, pp. 2946–2953, Jul. 2019, doi: 10.1109/TED.2019.2918102.
- [25] J. van Leeuwen, J. B. J. Smeets, and A. V. Belopolsky, "Forget binning and get SMART: Getting more out of the time-course of response data," *Atten Percept Psychophys*, vol. 81, no. 8, pp. 2956–2967, Nov. 2019, doi: 10.3758/s13414-019-01788-3.
- [26] I. Hossen, M. A. Anders, L. Wang, and G. C. Adam, "Data-driven RRAM device models using Kriging interpolation," *Sci Rep*, vol. 12, no. 1, Art. no. 1, Apr. 2022, doi: 10.1038/s41598-022-09556-4.
- [27] T. Dalgaty, N. Castellani, C. Turck, K.-E. Harabi, D. Querlioz, and E. Vianello, "In situ learning using intrinsic memristor variability via Markov chain Monte Carlo sampling," *Nat Electron*, vol. 4, no. 2, Art. no. 2, Feb. 2021, doi: 10.1038/s41928-020-00523-3.
- [28] S.-O. Park, H. Jeong, J. Park, J. Bae, and S. Choi, "Experimental demonstration of highly reliable dynamic memristor for artificial neuron and neuromorphic computing," *Nat Commun*, vol. 13, no. 1, Art. no. 1, Jun. 2022, doi: 10.1038/s41467-022-30539-6.
- [29] Y. Gao, S. Wu, and G. C. Adam, "Batch Training for Neuromorphic Systems with Device Non-idealities," in *International Conference on Neuromorphic Systems 2020*, Oak Ridge TN USA, Jul. 2020, pp. 1–4. doi: 10.1145/3407197.3407208.
- [30] E. Ambrosi, A. Bricalli, M. Laudato, and D. Ielmini, "Impact of oxide and electrode materials on the switching characteristics of oxide ReRAM devices," *Faraday Discuss.*, vol. 213, pp. 87–98, 2019, doi: 10.1039/C8FD00106E.
- [31] T. Gokmen and Y. Vlasov, "Acceleration of Deep Neural Network Training with Resistive Cross-Point Devices: Design Considerations," *Frontiers in Neuroscience*, vol. 10, 2016, Accessed: Mar. 21, 2022. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fnins.2016.00333>
- [32] L. Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, Nov. 2012, doi: 10.1109/MSP.2012.2211477.
- [33] J. Zhao, S. Huang, O. Yousuf, Y. Gao, B. D. Hoskins, and G. C. Adam, "Gradient Decomposition Methods for Training Neural Networks With Non-ideal Synaptic Devices," *Frontiers in Neuroscience*, vol. 15, p. 1524, 2021, doi: 10.3389/fnins.2021.749811.
- [34] A. Laborieux, M. Bocquet, T. Hirtzlin, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal, and D. Querlioz, "Implementation of Ternary Weights with Resistive RAM Using a Single Sense Operation per Synapse," arXiv, arXiv:2007.14234, Oct. 2020. doi: 10.48550/arXiv.2007.14234.
- [35] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep Learning with Limited Numerical Precision," in *Proceedings of the 32nd International Conference on Machine Learning*, Jun. 2015, pp. 1737–1746. Accessed: Dec. 13, 2021. [Online]. Available: <https://proceedings.mlr.press/v37/gupta15.html>
- [36] B. Hoskins, W. Ma, M. Fream, O. Yousuf, M. Daniels, J. Goodwill, A. Madhavan, H. Tung, M. Branstad, M. Liu, R. Madsen, J. McLelland, G. Adam, and M. Lueker-Boden, "A System for Validating Resistive Neural Network Prototypes," in *International Conference on Neuromorphic Systems 2021*, Knoxville TN USA, Jul. 2021, pp. 1–5. doi: 10.1145/3477145.3477260.

> REPLACE THIS LINE WITH YOUR MANUSCRIPT ID NUMBER (DOUBLE-CLICK HERE TO EDIT) <

- [37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*, 2019, vol. 32. Accessed: Nov. 06, 2021. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html>
- [38] N. C. Harris, J. Carolan, D. Bunandar, M. Prabhu, M. Hochberg, T. Baehr-Jones, M. L. Fanto, A. M. Smith, C. C. Tison, P. M. Alsing, and D. Englund, "Linear programmable nanophotonic processors," *Optica*, *OPTICA*, vol. 5, no. 12, pp. 1623–1631, Dec. 2018, doi: 10.1364/OPTICA.5.001623.
- [39] H. F. Inman and E. L. Bradley, "The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities," *Communications in Statistics - Theory and Methods*, vol. 18, no. 10, pp. 3851–3874, Jan. 1989, doi: 10.1080/03610928908830127.
- [40] J. A. Peacock, "Two-dimensional goodness-of-fit testing in astronomy," *Monthly Notices of the Royal Astronomical Society*, vol. 202, no. 3, pp. 615–627, Mar. 1983, doi: 10.1093/mnras/202.3.615.



Osama Yousuf (Student Member, IEEE) received a B.S. degree in computer science with a minor in mathematics from Habib University, Karachi, Pakistan in 2020. He is currently pursuing a Ph.D. degree in computer engineering at the George Washington University, Washington, DC, USA in the Adaptive Devices And Microsystems (ADAM) group as a research assistant. He also works as a research associate at the National Institute of Standards and Technology (NIST), Gaithersburg, MD. His research interests include robust neural network training algorithms, modelling emerging non-volatile memory devices, and prototyping ReRAM-based accelerators.



Imtiaz Hossen (Student Member, IEEE) received the B.Sc. degree in electrical and electronic engineering from the University of Dhaka, Dhaka, Bangladesh, in 2016, and the M.Sc. degree in electrical and computer engineering from Marquette University, Milwaukee, WI, USA, in 2020. He is currently pursuing a Ph.D. degree in electrical engineering at the George Washington University, Washington, DC, USA in the Adaptive Devices And Microsystems (ADAM) group as a research assistant. His research interests are nano-scale fabrication and testing of novel memory devices, device modeling, heterogeneous integration of ReRAM in CMOS foundry chips, ultra-high resolution temperature sensors and RF/microwave sensors.



Matthew W. Daniels is a research physicist in the Alternative Computing Group at the National Institute of Standards and Technology (NIST), Gaithersburg, MD. He received a B.S. in physics with a minor in mathematics from Clemson University in 2012 and a Ph.D. in physics from Carnegie Mellon University in 2017. His dissertation work was on topological physics and theoretical antiferromagnetic spintronics. His current research program

seeks to build a theory of hardware neural networks and neuromorphic systems. He currently works on understanding how time and stochasticity can be effectively utilized as primitive encodings for computing systems and on designing digital architectures that use the spectrally-biased, low-rank properties of neural network gradients to accelerate their training in hardware.



Martin Lueker-Boden is an engineering research director for Western Digital Corporation (WDC). He earned a Ph.D. in physics from the University of California at Berkeley in 2010. After earning his doctorate, he worked as a post-doctoral scholar at the California Institute of Technology. He started working at WDC in 2014 with a research focus on developing novel applications for non-volatile memory devices, including development of new cell technologies, methods for media management techniques, security considerations, I/O frameworks, software optimizations and in-memory/neuromorphic computing. His work prior to joining WDC focused on the development of superconducting low-noise microwave radiation sensors, and the statistical analysis of extra-galactic radiometry data.



Andrew Dienstfrey received the Ph.D. in Applied Mathematics from the Courant Institute of Mathematical Sciences at New York University. Dienstfrey joined the Mathematical and Computational Sciences Division at the National Institute of Standards and Technology (NIST), Boulder, CO in 2000. His research interests include applications of mathematics to computational physics and numerical analysis.



Gina C. Adam (Senior Member, IEEE) received the B.Sc. degree in applied electronics from the University Politehnica of Bucharest, Bucharest, Romania, in 2010, and the Ph.D. degree in electrical and computer engineering from the University of California at Santa Barbara, CA, USA, in 2015. From 2016 to 2018, she was a Research Scientist with the National Institute for Research and Development in Microtechnologies, Voluntari, Romania, and a Visiting Scholar with the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. She is currently an Assistant Professor of electrical and computer engineering with the School of Engineering and Applied Science, George Washington University, Washington, DC, USA. Her current research interests include resistive switching devices and their use in memory storage, computing, and communications applications.