Subthreshold Spintronic Stochastic Spiking Neural Networks with Probabilistic Hebbian Plasticity and Homeostasis

Steven D. Pyle, *Student Member, IEEE,* Ramtin Zand, *Student Member, IEEE,* Shadi Sheikhfaal, *Student Member, IEEE,* and Ronald F. DeMara, *Senior Member, IEEE*

***Abstract*—The Neural Sampling Core proposed herein offers a spintronic device based circuit and learning mechanism utilizing imprecise and stochastic components, similar to biological brains, to realize ultra-low-power neuromorphic computations at subthreshold voltages. Leveraging principles from Neural Sampling, a biologically-plausible theory from computational neuroscience, a spintronic stochastic spiking neuron with digital Post-Synaptic-Potentials is proposed in conjunction with low-precision spintronic synapses utilizing a new event-driven Probabilistic Hebbian Plasticity Rule, and a novel homeostasis mechanism that balances neural activity across multiple timescales and process variation effects. The primary computational operation, the summation of pre-synaptic potentials weighted by their corresponding synaptic efficacy and the neuron's homeostatic parameters, is performed in a parallel analog fashion using noisy and imprecise subthreshold components. It is demonstrated herein that the Neural Sampling Core is capable of learning orientation selectivity, much like the simple cells found in the visual cortex, in an unsupervised fashion at 311nW per neuron and 1.9-7.7 nW per active synapse using a 200mV supply voltage.**

***Index Terms*—neuromorphic, process variation, spintronic, unsupervised learning, homeostasis, subthreshold, neural sampling.**

I. Introduction

RECENT research into Spiking Neural Network (SNN) hardware has aimed to achieve computational capabilities akin to biological brains, such as inference, at efficiencies unachievable with Von-Neumann hardware [1]. Several custom SNN ASICs, such as IBM's TrueNorth [2] and SpiNNaker [3] have demonstrated impressive capabilities at very low power using standard CMOS technology. Furthermore, much work is being done on developing neuromorphic hardware utilizing emerging non-volatile devices, such as spintronics and memristors, to implement primarily synapses [4]–[12], but also neurons [8]–[13] in compact and energy-efficient designs when compared to CMOS-only approaches.

An intriguing observation is that biological brains and nanoscale electronic circuits share characteristics that can provide insights towards designing circuits and architectures that can be utilized for biologically-inspired computation with potential power efficiency comparable to brains. First, the primary computational structures of biological brains, neurons and synapses, are highly heterogeneous and imprecise [14], [15], which is akin to the fact that all manufactured nanodevices have behavioral variability arising from Process Variation (PV), especially CMOS devices operating at subthreshold voltages [16]. Perhaps by designing circuits and architectures that can adapt to, and even utilize, such heterogeneity while trying to aggressively lower supply voltages, even greater power efficiency could be achieved compared to adhering to the strict design margins and deterministic behaviors that VLSI circuits are typically designed to realize. Second, the fundamental mechanisms underlying neural activity, ion channel opening and closing, is a stochastic process, which leads to stochasticity throughout neural activity [17]. Coincidentally, a promising framework in computational neuroscience, Neural Sampling, has theoretically proven that a particular biologically-plausible model of stochastically spiking neurons in cortical circuit motifs represent samples from an underlying conditional distribution that can be used for probabilistic inference [18], [19]. Therefore, leveraging heterogeneity and stochasticity in neuromorphic architectures using emerging devices that are intrinsically stochastic, such as spintronics [8], [9], [13], could lead to more capable and efficient neuromorphic hardware.

The Neural Sampling Core (NSC) presented herein is motivated by the ultra-low-power and robust characteristics of biological neural networks, which utilize stochastic and heterogeneous components with local learning rules in competitive networks. The NSC is a thrust to mimic the underlying computational principles of the brain in nanoelectronic circuits that can realize self-adaptive and low-power neuromorphic hardware with noisy and imprecise CMOS and spintronic devices operating at subthreshold voltages.

The following contributions are provided:

1) a stochastic spiking neuron circuit with protracted digital post-synaptic-potentials realizing behaviors from Neural Sampling,
2) a low-precision hybrid spintronic-CMOS synapse circuit with a new event-based Probabilistic Hebbian Plasticity (PHP) unsupervised learning mechanism, and
3) a novel homeostasis mechanism regulating neural activity across multiple time-scales and process variations.

The above contributions are integrated into low-power neuromorphic hardware approach operating at subthreshold voltages, yet remaining robust to noisy and imprecise components.

The remainder of the manuscript is organized as follows. Section II delineates the requisite background information on spintronics, neuromorphic hardware, and Neural Sampling, which underlies the NSC approach. Section III presents the NSC and its associated circuits and algorithms. Section IV details the simulation results, analyzing the unsupervised learning capabilities, the power consumption of each circuit, and the effects of input noise. Section V concludes the paper

and provides future directions towards implementing and improving the NSC. The accompanying Supplementary Material details and justifies the modeling methodology, simulation framework, and provides the corresponding parameters which have been used.

II. Background

A. Spintronics

The field of spintronics aims to utilize the properties of nanoscaled magnetic structures to realize computational and non-volatile memory elements [20]–[22]. The most well-developed spintronic device is the Magnetic Tunnel Junction (MTJ) shown in Figure 1a, which consists of a thin tunneling oxide, typically MgO, sandwiched between two magnetic layers [21]. One magnetic layer is called the fixed layer, since its magnetic orientation remains unchanged during its operation, and the other is called the free layer, since its magnetic orientation is altered according to the physical behaviors underlying the device's switching mechanism. The most popular switching mechanism for MTJs is Spin-Transfer-Torque (STT), which works by passing a current of sufficient density and duration through the device [23]. The state of the MTJ is represented by the resistance of the device, which changes based on the orientations of the free layer relative to the fixed layer. The Anti-Parallel (AP) state results in a higher resistance than the parallel (P) state, and the relative resistance change between the two states is called the Tunneling Magnetoresistance Ratio (TMR) [21]. A relatively recent spintronic device, called the Spin-Hall Effect MTJ (SHE-MTJ), improves several aspects of the standard two-terminal MTJ by placing a heavy metal, such as Pt or β-Ta underneath the free layer, which decouples the read and write paths as shown in Figure 1b, and can improve the energy efficiency of the write process if properly designed [24]. For both the two-terminal MTJ and three-terminal SHE-MTJ, the switching process is a stochastic function of the current density and the pulse duration, whereby deterministic implementations require large current densities and pulse durations to ensure a very high probability of switching [25]. Alternatively, several works, including herein, utilize the intrinsic stochastic switching behavior, which allows for much less current density to be used during the switching process, and thus, less power [8], [9], [26]. A key parameter of spintronic devices is the energy barrier (Δ), which is a function of the magnetic material properties and the shape of the device, and it determines the retention time of the free layer and the current density needed to switch the device for a given pulse duration [9]. When used as a memory element, MTJs and SHE-MTJs typically have $\Delta \geq 40k_bT$, where k_b is Boltzmann's constant and T is the temperature in Kelvin. For $\Delta \ll 40k_bT$, thermal agitations can stochastically switch the device between parallel and anti-parallel states on timescales of seconds to picoseconds, which is practically unusable for non-volatile memory applications. However, recent work into Probabilistic Spintronic Logic has demonstrated that very low Δ spintronic devices are useful for realizing stochastic computations in compact circuits, which can be utilized for invertible logic and Boltzmann Machines [27], Restricted Boltzmann Machines [28], and stochastically spiking neural circuits [13]. One promising probabilistic spintronic device is the Embedded p-bit [29] shown in Figure 1c, which consists of a MTJ with a very low Δ, an NMOS transistor, and a CMOS inverter. The input to the inverter is essentially a voltage-divider between the stochastically switching MTJ and the NMOS, and therefore the *probability* of the output of the inverter being high will have a sigmoidal behavior based on the input voltage to the NMOS. This behavior is demonstrated in greater detail in [29] and the Supplementary Material.

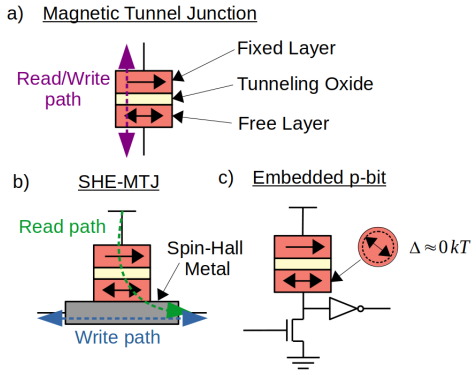

Fig. 1. An overview of relevant spintronic devices. a) the two-terminal MTJ illustrating its 3 primary layers with a shared read and write path. b) the three-terminal SHE-MTJ with a decoupled read path through the MTJ and the write path through the spin-hall metal. c) the embedded p-bit with a very low Δ MTJ and its associated CMOS circuitry.

The integration of multiple technologies on a single chip will always increase integration complexity. However, it is understood that the energy barrier can be manipulated by adjusting the volume of the device as well as the in-plane dimensions for in-plane devices [30], [31]. Thus, it is possible to integrate low-barrier and high-barrier devices on a chip by adjusting the length and width of the MTJs, which can be done on the same mask, adding a minimal increase in integration complexity.

B. Neural Sampling

Neural Sampling is a theory of brain computation from computational neuroscience that interprets the stochastic spiking behavior of biological neurons as stochastic samples of underlying conditional distributions [18], [19]. Particularly, it models the spiking behavior of neurons with an instantaneous stochastic spiking rate exponentially dependent upon the membrane potential, combined with a refractory period of duration τ and a commensurately prolonged rectangular Post-Synaptic-Potential (PSP), which approximates the PSPs found in-vivo. Combined with a Hebbian learning rule, such a model can realize a generative model of the input distribution [19]. This is in contrast to typical Leaky-Integrate and Fire (LIF) spiking neurons, which models spikes as impulses and neurons as a leaky integration of synaptically-weighted pre-synaptic spikes that fires if a threshold is reached and then reset [26]. For the rest of the paper, a spike means a rectangular pulse of τ clocks, as in Neural Sampling. Several cortically-inspired circuit motifs have been developed utilizing Neural Sampling that have demonstrated impressive results of unsupervised, and reward-based learning [19], [32]. Thus, Neural Sampling provides a theoretically-accomplished and biologically-relevant

TABLE I
COMPARISON TO PREVIOUS SPINTRONIC SPIKING NEURAL NETWORKS WITH UNSUPERVISED LEARNING

	Synapse Technology	Neuron Technology	Learning Rule	Homeostasis	Key Quantitative Findings
[12]	Compound MTJ	Stochastic switching MTJ	Simplified Stochastic STDP	None	91.27% classification accuracy on MNIST
[26]	LT-ST SHE-MTJs	LIF	Stochastic STDP	None	10.4 μJ to train network on MNIST
[8]	SHE-MTJ	Stochastic switching SHE-MTJ	Stochastic STDP	Spike count cutoff	682 nW per neuron
Herein	*Spin-CMOS*	*Embedded p-bit with PSP*	*PHP*	*Adaptive to fast and slow time-scales*	*311 nW per neuron 1.9-7.7 nW per synapse*

framework for using stochastic neural models to achieve brain-like computations, and we provide a detailed connection to this work in the Supplementary Material.

C. Stochastic Neuromorphic Hardware

Several recent works have leveraged the stochastic switching properties of spintronic devices to realize unsupervised learning in SNN neuromorphic hardware as delineated in Table I. The work developed by Zhang et al. [12] utilized multiple parallel MTJs to form a compound magnetoresistive synapse with a stochastic Spike-Timing-Dependent (STDP) learning rule in conjunction with a MTJ-based stochastic spiking neuron to realize a SNN able to achieve respectable accuracies on MNIST dataset. However, their work did not evaluate the power consumption of the design, which can be quite large for many parallel MTJs per synapse in a crossbar, nor the effect of process variation on the CMOS circuitry necessary for the neuron.

The long-term short-term stochastic synapse developed by Srinivasan et al. [26] utilizes two SHE-MTJs with distinct peripheral circuitry to realize various switching characteristics corresponding to different STDP sensitivities, enabling one SHE-MTJ to have sharper correlation sensitivity and greater synaptic strength then the other, which had moderate correlation sensitivity. They demonstrated that the scheme has faster training convergence, resulting in a reduction in total training energy consumption. However, the scheme was quite sensitive to STDP and circuit parameters, and they did not analyze the effect of process variations.

The all-spin stochastic SNN developed in [8] leverages one-bit SHE-MTJ synapses with a stochastic-STDP learning rule and SHE-MTJ based stochastic spiking neurons with a homeostasis mechanism to realize a low-energy SNN with online learning. However, the SHE-MTJ neuron requires write-read-reset cycling, which adds additional timing and energy overheads, the stochastic-STDP learning rule requires precision between the spike timing, switching probability, and write current, the homeostasis mechanism is rather coarse since it simply cuts off neurons that reach a certain spike count during learning, and the effect of process variations are not analyzed.

Additionally, the spintronic stochastic spiking neurons in [8], [12], [30], [33] demonstrate the utilization of high-barrier spintronic devices for stochastic spike generation by applying an input current pulse, which may or may not have switched the device, then reading the state of the SHE-MTJ to determine if it spiked, then applying a strong reset pulse. This three-phase write-read-reset scheme requires additional timing and power overheads that are not experienced with low-barrier p-bits. The work in [30] also explored utilizing low-barrer telegraphic SHE-MTJs to implement stochastic spiking neurons without the write-read-reset overheads, but their approach utilizes the direct output of an inverter without any event generation or synchronization mechanism, so it does not resemble spiking or PSP behavior, which can make it challenging for event-based probabilistic Hebbian learning rules.

Thus, the NSC developed herein extends beyond these promising works by developing a robust subthreshold stochastic SNN approach utilizing a 3-bit hybrid spin-cmos synapse with series and parallel SHE-MTJs, a flexible and adaptive homeostasis mechanism, and a p-bit stochastic spiking neuron with digital PSPs implementing neural sampling and enabling a simple and robust event-driven unsupervised learning mechanism, all developed and analyzed with the effect of process variations in both the spintronic and CMOS devices.

III. NEURAL SAMPLING CORE

The Section delineates the constituent circuits of the NSC, such as the stochastically spiking neuron with a refractory period and prolonged digital PSPs congruent to those utilized in Neural Sampling's theoretical modeling, a three-bit synapse with event-driven probabilistic Hebbian learning rules, and a novel homeostasis mechanism. Since an important premise of this work is that the NSC should be able to adapt and utilize the heterogeneity of components that emerges from PV, we model PV in both the spintronic and CMOS devices as described in the Supplementary Material at all stages of development and analysis. This section is organized by first detailing the operational principles of each circuit and then integrating them into a cohesive mixed-signal architecture with discussions. Although detailed later, it is worth mentioning here that there are two reciprocating phases based on the state of the clock; the read-phase occurs when the clock is low, and the update-phase occurs when the clock is high.

A. Stochastic Spiking Neuron

The Stochastic Spiking Neuron circuit shown in Figure 2a consists of an embedded p-bit and a digital PSP circuit that operates as follows. Based on the voltage applied to IN and the state of the stochastically switching MTJ in the embedded p-bit, $p-bit_{OUT}$ will either be high or low.

If $p-bit_{OUT}$ is high at the positive edge of a 100 MhZ CLK, then the output of the PSP circuit, $Neuron_{OUT}$, will go high and hold it for eight clocks, which corresponds to a τ of eight clocks. The waveforms shown in Figure 2b is an illustrative snapshot that shows the relevant circuit signals obtained from SPICE simulations for the parameters given in the Supplementary Material. A more detailed analysis of the sigmoidal probabilistic circuit behavior with PV is also detailed in the Supplementary Material.

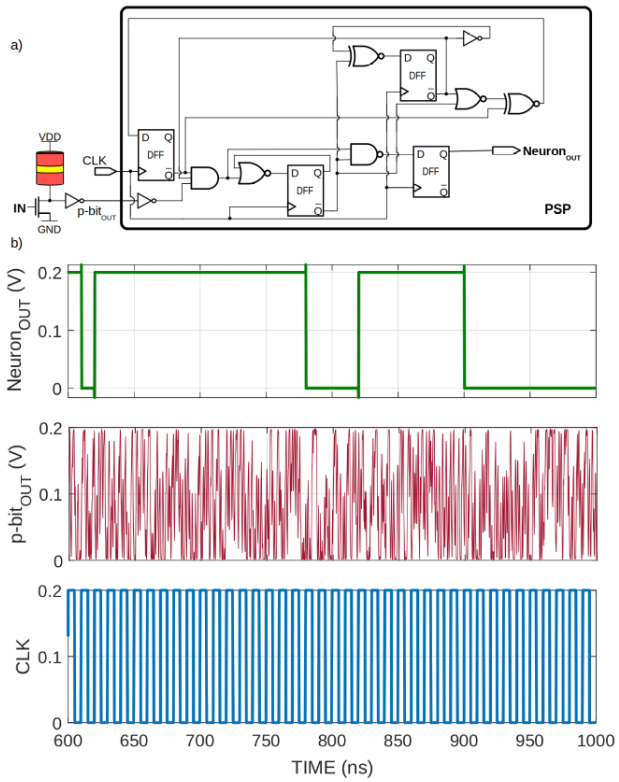

Fig. 2. The stochastic spiking neuron developed herein. a) the neuron utilizes an embedded p-bit to compute the sigmoidal probability of spiking based on the voltage at IN and the PSP circuit senses the state of the p-bit at the positive clock edge and holds $Neuron_{OUT}$ high for 8 clocks for each spike. b) the operational waveforms for $IN = 144mV$, which corresponds to a spiking probability of ~ 0.5.

B. Hybrid Synapse with Probabilistic Hebbian Plasticity

The hybrid spintronic-CMOS synapse shown in Figure 3 and the PHP learning rule were co-designed to take advantage of the prolonged PSP signals with the stochastic switching behavior of spintronic devices. After extensive investigations and examinations with alternative learning rules, PHP was found to yield the best results for circuits with PV as shown in Section IV. The circuit operates as follows. The synapse in Figure 3 uses three SHE-MTJs (S1-S3) to store the synaptic weight, one PMOS transistor ($M1$) that operates as a voltage-controlled current source since the circuit is at subthreshold, and two NMOS transistors ($M2-M3$) that are used when updating the synapse. The circuit operates during the read phase as follows. If the pre-synaptic neuron is not active, or has not spiked within the last τ clocks, then $\overline{IN}$ will be at VDD, N will be at VDD, and no current will flow through $M1$ onto SUM. If the pre-synaptic neuron has spiked within the previous τ clocks, then $\overline{IN}$ will be at GND, causing a voltage-divider between $S1$ and $S2-S3$, which determines the voltage at N, which then controls the current through $M1$ into SUM. The synaptic weights determined by the P or AP states of S1-S3 are shown in Table 2 where $W0 < W1 < W2 \sim W3 < W4 < W5$ and detailed in the Supplementary Material. It is also worth noting that this scheme can be amended to have a greater range of possible weight values by increasing the number of SHE-MTJs in series or parallel with S1-S3 for additional area overhead.

TABLE II
SYNAPSE WEIGHTS

S1	S2	S3	Weight
P	AP	AP	W0
P	P	AP	W1
P	AP	P	W1
P	P	P	W2
AP	AP	AP	W3
AP	P	AP	W4
AP	AP	P	W4
AP	P	P	W5

PHP modifies the synapses during the update phase in an event-driven fashion as follows. If the post-synaptic neuron, $POST$, has spiked during the previous τ clocks, then both M2 and M3 are turned on, allowing current to flow through the write paths of S1-S3 based on the voltages applied to PRE and $\overline{IN}$. If the pre-synaptic neuron has spiked within the previous τ clocks as well, then the synapse will update according to a *synaptic potentiation* event, that is, different voltages will be applied to PRE and $\overline{IN}$ for a given pulse duration such that S1 has a probability of switching to its anti-parallel state and S2-S3 have a probability of switching to their parallel states, which all have the effect of lowering the voltage at N and increasing the current through M1 during the read-phase. If the pre-synaptic neuron has not spiked within the previous τ clocks, then the synapse will update according to a *synaptic depression* event, that is, voltages will be applied to PRE and $\overline{IN}$ for a given pulse duration such that S1 has a probability of switching to its parallel state and S2-S3 have a probability of switching to their anti-parallel states, which all have the effect of increasing the voltage at N and decreasing the current through M1 during the read-phase. Therefore, each time a post-synaptic neuron spikes, all associated synapses are probabilistically updated for τ clocks, and more coincident pre-synaptic spiking will have a higher chance of strengthening the synapse, while non-spiking pre-synaptic neurons will have a chance of being depressed. More details can be found in the Supplementary Material.

C. Homeostasis Mechanism

The homeostasis mechanism acts to increase the activity of under-active neurons and decrease the activity of over-active neurons, is implemented with a number of the homeostatic synapses shown in Figure 4 connected to the input of each neuron. The two homeostatic synapse designs shown in Figure 4 utilize alternative mechanisms for implementing

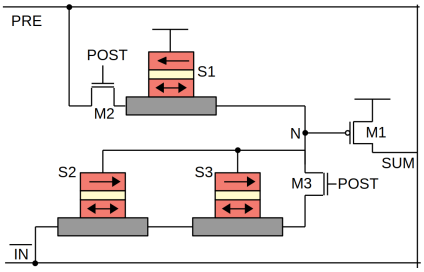

Fig. 3. The three-bit hybrid spin-cmos synapse circuit developed herein.

homeostasis on both fast and slow time-scales, where S1 has a higher probability of switching compared to S2, and therefore adapts on a faster time-scale. The positive-feedback effect of synaptic plasticity needs a fast homeostasis mechanism to balance network activity [34], [35] while a slower homeostasis mechanism is beneficial for balancing the neuron's excitability in the presence of its intrinsic heterogeneity arising from PV. Both of the designs operate similar to the regular synapse during the read phase as follows. During the read phase $\overline{BOT}$ is pulled to GND, causing a voltage divider between S1 and S2, which determines the voltage at N, which then determines the current through M1 into SUM. The weight values are akin to the regular synapses described previously in that if S1 is AP and/or S2 is P, then the homeostatic synapse has a higher effective weight than vice versa, and is detailed in the Supplementary Material. The two designs differ during the update phase as follows. The circuit in 4a requires S1 to have a lower Δ than S2, which causes it to have a higher probability of switching for the same current and pulse duration. The circuit in 4b does not require S1 and S2 to have different Δs, but requires more overhead with an additional NMOS and two horizontal wires to isolate the two devices during the update phase, allowing different voltages and/or pulse durations to switch the two devices with different probabilities such that S1 switches with a higher probability than S2. During the update phase, $UPDATE$ goes high and different voltages are applied to TOP and $\overline{BOT}$ for Figure 4a, or TOP, $\overline{TOP}$, BOT, and $\overline{BOT}$ for Figure 4b, depending on the state of the connected neuron - if the neuron is active, then a *homeostatic depression* event occurs, and if the neuron is inactive, then a *homeostatic potentiation* event occurs.

D. Inhibition Mechanism

Inhibitory feedback is a mechanism to ensure that only a small number of output neurons are active at a time by decreasing the input strength of the others, and therefore their chances of spiking, each time one has spiked. This enforces competition between the neurons, which enforces selectivity [36]. Without it, it is likely for all neurons to become receptive to all input patterns, and therefore there is no information from the network that can be used to discern the input patterns from one another, which is key for unsupervised learning and probabilistic inference [36]. The exact inhibitory mechanisms that the brain utilizes is still an active area of research, but many SNN models utilize a fixed inhibition model such that every time a neuron spikes, a fixed decrease in input strength is

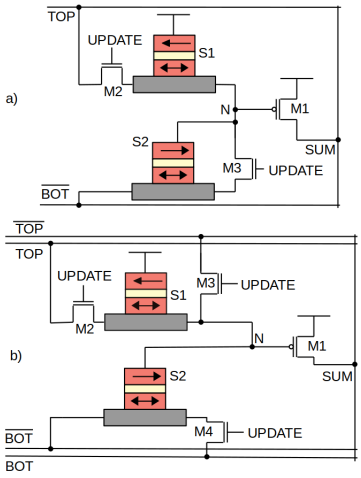

Fig. 4. Two alternative implementations of the homeostatic synapse. a) this implementation requires the energy barrier of S1 to be different than S2 such that they have different switching probabilities for the same current. b) this implementation does not require S1 to have a different energy barrier than S2, but requires additional update circuitry to provide different current pulses to each SHE-MTJ.

applied to all other neurons [19], [37], and the same is used for the NSC. In order to minimize area overhead, the inhibition mechanism is implemented with a single NMOS connected to the SUM wire and GND. The input voltage to that NMOS is chosen such that the effect on SUM is equivalent to the negative of the strongest synaptic weight, $W5$, and its associated distribution according to PV, as discussed in more detail in the Supplementary Material.

E. Architectural Discussion

Figure 5 shows all of the core components of the NSC integrated into a single layer feed-forward SNN. During the read phase, which is when CLK is low, if $POST$ is also low, and therefore M_{read} is on, all of the synapses with spiked pre-synaptic neurons, all of the homeostatic synapses, and all of the inhibitory feedback with active $POST$ signals will source and sink current, generating a voltage at SUM due to the resistance of R_{sum} and M_{read}, which is the resulting parallel analog computation of weighted pre-synaptic spikes plus the cumulative effect of the homeostatic synapses minus any active inhibition, and is applied to the input of the stochastic spiking neuron circuit. When CLK goes high, the post-synaptic neuron may or may not have spiked, M_{read} turns off to prevent wasted current flow, all inhibitory feedback turns off for the same reason, and the synapse and homeostatic update mechanisms occur according to Algorithm 1.

The event-based nature of the NSC with its non-volatile parameters affords flexibility to its operational and greater architectural needs. For instance, the NSC is described herein with two phases corresponding to different states of the clock for simplicity, but in principle, many other operations could intermix with the two main phases, such as routing algorithms for intra- and inter-chip communications or monitoring processes. Additionally, the clock rate could be adjusted based on application needs, using a slower clock when idle and a faster clock as needed. The clock rate could also be adjusted based

Algorithm 1: Update Phase

Input neuron state: $\mathbf{y}(t) \in [0, 1]$
Output neuron state: $\mathbf{z}(t) \in [0, 1]$
for $z_i(t)$ ***in*** $\mathbf{z}(t)$ **do**
  // Update homeostatic synapses
  if $z_i(t) == 1$ **then**
    homeostatic-depression(i);
  else
    homeostatic-potentiation(i);
  end
  // Update input-output synapses
  if $z_i(t) == 1$ **then**
    for $y_j(t)$ ***in*** $\mathbf{y}(t)$ **do**
      if $y_j(t) == 1$ **then**
        synaptic-potentiation(i,j);
      else
        synaptic-depression(i,j);
      end
    end
  end
end

on as-manufactured timing considerations. Another beneficial aspect of the non-volatile nature of the NSC is that the more power-intensive update phases are only required during training and/or re-training. Once a desired capability is achieved, the update phases can be much more dispersed or stopped altogether, saving considerable power.

Another benefit of the NSC is that it is able to learn patterns of different dimensions, as is described in the following Section, using all the same constituent circuits and devices with just an alteration of the number of homeostatic synapses - smaller dimensional inputs require more homeostatic synapses. Therefore, fixed NSC networks of a certain size could be fabricated and then inputs and homeostatic synapses could be turned on or off depending on the application needs. Also, this could provide redundancy in the case of unusable components, providing a higher potential yield.

IV. Results

This Section describes the simulation results of the NSC. The circuits of the NSC were simulated and analyzed using SPICE and then modeled in Brian2, a SNN simulation framework [38], to obtain the unsupervised learning results. Readers are strongly recommended to refer to the Supplementary Material regarding the details about Monte Carlo simulations for process variation in both synapses and neurons.

A. Unsupervised Learning

The emergent unsupervised learning capabilities of the NSC are demonstrated by learning a cortically-inspired behavior, orientation selectivity, within a feed-forward SNN of 50 output neurons with 60 homeostatic synapses each and 900 Poisson spiking input neurons that each correspond to a pixel in a 30x30 stimulus window. The input pattern distribution consist of 180 28x2 bars centered and rotated in the stimulus window such that they cover the complete 180 degrees of rotation. The synaptic weights are initialized with S1-S3 randomly distributed and the homeostatic synapses are initialized with

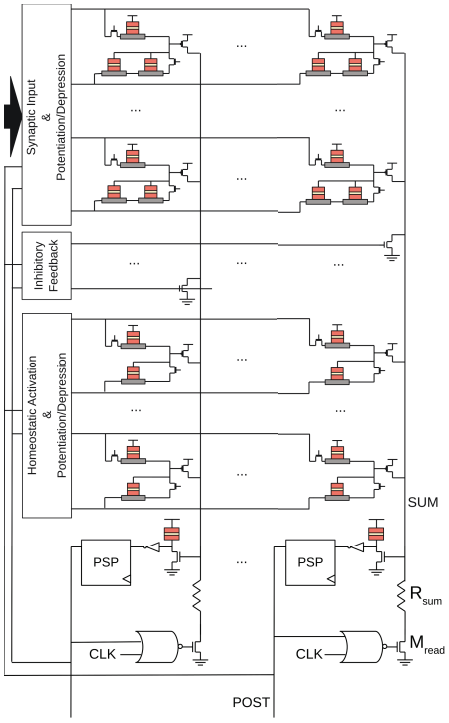

Fig. 5. The structure of NSC illustrating the integration of the synapses, inhibitory feedback, homeostasis mechanism, and stochastic spiking neuron.

S1 in AP state and S2 in P state. Up to 10,000 randomly chosen samples from the input distribution are presented to the network for 100 clocks where each input neuron that the randomly chosen bar corresponds to has a Poisson spike rate of 75 spikes per 1,000 clocks and all others have a spike rate of 1 spike per 1,000 clocks. In between each sample is a brief period of 20 clocks whereby all input neurons have a spike rate of 1 spikes per 1,000 clocks. Figure 6a shows the temporal evolution of a random selection of output neuron's receptive fields, that is, the strength of their 900 synapses shaped into a 30x30 window corresponding to the stimulus window, where a lighter color indicates a stronger synaptic strength, illustrating the emergent specialization of each neuron to a particular input pattern. Figure 6b illustrates the emergent orientation selectivity in another way, where all synapses were fixed and each input pattern was presented to the network for 100 clocks and the spikes of all output neurons were counted and shown for a random selection of 5 output neurons. It can be seen that the spike counts closely resemble the tuning curves for simple cells in V1 cortex [39]. Additionally, the Supplementary Material shows the tuning curves for all output neurons, and it can be seen that the entire range of possible orientations are well represented by the collection of output neurons. The NSC was also tested using a smaller stimulus window of 20x20 and bars of 18x2, and the only needed change was an increase in the number of homeostatic synapses to 90.

B. Noise Analysis

The NSC is also quite robust to input noise, and is actually able to utilize such noise for some benefit. This was tested

by adding a uniformly distributed random spike rate between 0 and 7.5 spikes per 1000 clocks to each input neuron for each pixel in the stimulus window as described previously. The noise had a regulating effect, decreasing the amount of homeostatic synapses required to just 30 for a 30x30 stimulus window. The NSC was still able to learn orientation selectivity with the noise, although the receptive field was qualitatively more noisy and the tuning curves were on average a bit broader, as shown in the Supplementary Material.

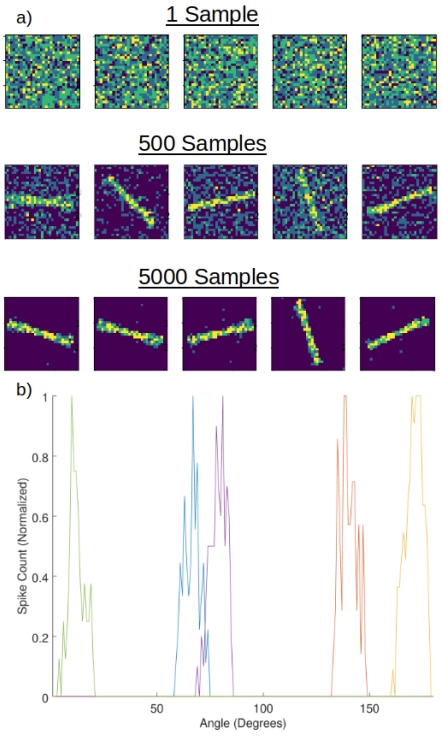

Fig. 6. Unsupervised learning results for the NSC. a) the emergence of orientation selectivity is illustrated here in the receptive fields for a random selection of five neurons. Brighter colors correspond to a higher synaptic weight. b) the tuning curves for a random selection of five neurons are shown, illustrating their response to a narrow range of orientations. Each color represents a different neuron.

C. Power Analysis

The average power consumption for each of the NSC circuits was found using SPICE simulations as described in the Supplementary Material and was found to be 310nW for the stochastic spiking neuron with PSP circuit, 1.9-7.7nW for each of the input synapses, depending on the synaptic strength, and 1-3.4nW for each of the homeostatic synapses, depending on its strength. The average power consumption of the network during the read phase for the neurons, homeostasis mechanism, and active synapses for the 20x20, 30x30, and 30x30 with noise test cases are shown in Figure 7. The inhibitory mechanism was found to be negligible since very few output neurons are ever active at one time. As shown, the power consumption due to the output neurons are all equal since the number of neurons does not change. The power consumption due to the synapses increases from the 20x20 case to the 30x30 case since there are more inputs and synapses, and the noise increases the synaptic power consumption due to there being more active synapses as well as a higher number of higher strength synapses. The homeostasis power consumption is highest for the 20x20 case since it has the fewest active input synapses, and therefore needs on average more homeostatic input synapses to drive the neurons to spike, and is lowest for the 30x30 with noise for the exact opposite reason. The power consumption of the update phase is not considered due to the NSC requiring updates only during training or re-training, and is thus a very small fraction of the total lifetime energy usage. Additionally, the power consumption of the update phase depends heavily upon the materials, dimensions, and technology of the devices used.

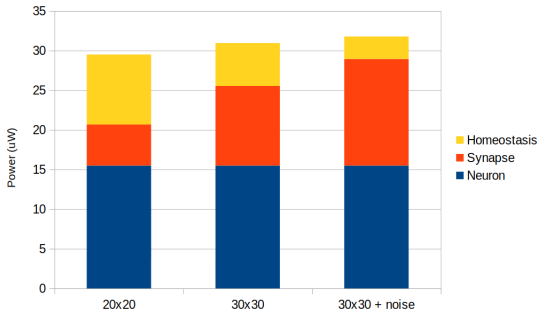

Fig. 7. The average power consumption for each component of the NSC during the presentation of all 180 degrees of possible orientations for the 20x20, 30x30, and 30x30 with noise cases.

V. CONCLUSIONS

The NSC described herein provides several intriguing insights to realizing ultra-low-power neuromorphic circuits and architectures. Future directions for extending the NSC could be to implement recurrent connections and migrate the inhibitory mechanism to a population of inhibitory neurons, which would more closely resemble cortical network motifs, to explore how networks of NSCs could be connected together in deep or hierarchical fashions to realize greater computational ability, or to develop methodologies that can implement supervised or reinforcement learning capabilities.

ACKNOWLEDGMENT

This work was supported in part by the Center for Probabilistic Spin Logic for Low-Energy Boolean and Non-Boolean Computing (CAPSL), one of the Nanoelectronic Computing Research (nCORE) Centers as task 2759.006, a Semiconductor Research Corporation (SRC) program sponsored by the NSF through CCF 1739635.

REFERENCES

[1] S. Yu, "Neuro-inspired computing with emerging nonvolatile memorys," *Proceedings of the IEEE*, vol. 106, no. 2, pp. 260–285, 2018.

[2] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

[3] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The spinnaker project," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 652–665, 2014.

[4] E. Covi, R. George, J. Frascaroli, S. Brivio, C. Mayr, H. Mostafa, G. Indiveri, and S. Spiga, "Spike-driven threshold-based learning with memristive synapses and neuromorphic silicon neurons," *Journal of Physics D: Applied Physics*, vol. 51, no. 34, p. 344003, 2018.

[5] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. Nolfo, S. Sidler, M. Giordano, M. Bodini, N. C. Farinha *et al.*, "Equivalent-accuracy accelerated neural-network training using analogue memory," *Nature*, vol. 558, no. 7708, p. 60, 2018.

[6] G. Pedretti, V. Milo, S. Ambrogio, R. Carboni, S. Bianchi, A. Calderoni, N. Ramaswamy, A. S. Spinelli, and D. Ielmini, "Stochastic learning in neuromorphic hardware via spike timing dependent plasticity with rram synapses," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 1, pp. 77–85, 2018.

[7] P. M. Sheridan, F. Cai, C. Du, W. Ma, Z. Zhang, and W. D. Lu, "Sparse coding with memristor networks," *Nature nanotechnology*, vol. 12, no. 8, p. 784, 2017.

[8] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic tunnel junction enabled all-spin stochastic spiking neural network," in *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2017, pp. 530–535.

[9] A. Sengupta and K. Roy, "Encoding neural and synaptic functionalities in electron spin: A pathway to efficient neuromorphic computing," *Applied Physics Reviews*, vol. 4, no. 4, p. 041105, 2017.

[10] P. Wijesinghe, A. Ankit, A. Sengupta, and K. Roy, "An all-memristor deep spiking neural computing system: A step toward realizing the low-power stochastic brain," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 5, pp. 345–358, 2018.

[11] Z. Wang, S. Joshi, S. Savelev, W. Song, R. Midya, Y. Li, M. Rao, P. Yan, S. Asapu, Y. Zhuo *et al.*, "Fully memristive neural networks for pattern classification with unsupervised learning," *Nature Electronics*, vol. 1, no. 2, p. 137, 2018.

[12] D. Zhang, L. Zeng, Y. Zhang, W. Zhao, and J. O. Klein, "Stochastic spintronic device based synapses and spiking neurons for neuromorphic computation," in *Nanoscale Architectures (NANOARCH), 2016 IEEE/ACM International Symposium on*. IEEE, 2016, pp. 173–178.

[13] S. D. Pyle, K. Y. Camsari, and R. F. DeMara, "Hybrid spin-cmos stochastic spiking neuron for high-speed emulation of in vivo neuron dynamics," *IET Computers & Digital Techniques*, 2018.

[14] L. E. Dobrunz and C. F. Stevens, "Heterogeneity of release probability, facilitation, and depletion at central synapses," *Neuron*, vol. 18, no. 6, pp. 995–1008, 1997.

[15] F. Baroni and A. Mazzoni, "Heterogeneity of heterogeneities in neuronal networks," *Frontiers in computational neuroscience*, vol. 8, p. 161, 2014.

[16] S. Fisher, A. Teman, D. Vaysman, A. Gertsman, O. Yadid-Pecht, and A. Fish, "Digital subthreshold logic design-motivation and challenges," in *Electrical and Electronics Engineers in Israel, 2008. IEEEI 2008. IEEE 25th Convention of*. IEEE, 2008, pp. 702–706.

[17] G. Deco, E. T. Rolls, and R. Romo, "Stochastic dynamics as a principle of brain function," *Progress in neurobiology*, vol. 88, no. 1, pp. 1–16, 2009.

[18] L. Buesing, J. Bill, B. Nessler, and W. Maass, "Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons," *PLoS computational biology*, vol. 7, no. 11, p. e1002211, 2011.

[19] R. Legenstein, Z. Jonke, S. Habenschuss, and W. Maass, "A probabilistic model for learning in cortical microcircuit motifs with data-based divisive inhibition," *arXiv preprint arXiv:1707.05182*, 2017.

[20] B. Behin-Aein, J.-P. Wang, and R. Wiesendanger, "Computing with spins and magnets," *MRS Bulletin*, vol. 39, no. 8, pp. 696–702, 2014.

[21] D. E. Nikonov and I. A. Young, "Overview of beyond-cmos devices and a uniform methodology for their benchmarking," *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2498–2533, 2013.

[22] R. L. Stamps, S. Breitkreutz, J. Åkerman, A. V. Chumak, Y. Otani, G. E. Bauer, J.-U. Thiele, M. Bowen, S. A. Majetich, M. Kläui *et al.*, "The 2014 magnetism roadmap," *Journal of Physics D: Applied Physics*, vol. 47, no. 33, p. 333001, 2014.

[23] A. Brataas, A. D. Kent, and H. Ohno, "Current-induced torques in magnetic materials," *Nature materials*, vol. 11, no. 5, p. 372, 2012.

[24] S. Manipatruni, D. E. Nikonov, and I. A. Young, "Energy-delay performance of giant spin hall effect switching for dense magnetic memory," *Applied Physics Express*, vol. 7, no. 10, p. 103001, 2014.

[25] T. Devolder, J. Hayakawa, K. Ito, H. Takahashi, S. Ikeda, P. Crozat, N. Zerounian, J.-V. Kim, C. Chappert, and H. Ohno, "Single-shot time-resolved measurements of nanosecond-scale spin-transfer induced switching: Stochastic versus deterministic aspects," *Physical review letters*, vol. 100, no. 5, p. 057206, 2008.

[26] G. Srinivasan, A. Sengupta, and K. Roy, "Magnetic tunnel junction based long-term short-term stochastic synapse for a spiking neural network with on-chip stdp learning," *Scientific reports*, vol. 6, p. 29545, 2016.

[27] K. Y. Camsari, R. Faria, B. M. Sutton, and S. Datta, "Stochastic p-bits for invertible logic," *Physical Review X*, vol. 7, no. 3, p. 031014, 2017.

[28] R. Zand, K. Y. Camsari, S. D. Pyle, I. Ahmed, C. H. Kim, and R. F. DeMara, "Low-energy deep belief networks using intrinsic sigmoidal spintronic-based probabilistic neurons," in *Proceedings of the 2018 on Great Lakes Symposium on VLSI*. ACM, 2018, pp. 15–20.

[29] K. Y. Camsari, S. Salahuddin, and S. Datta, "Implementing p-bits with embedded mtj," *IEEE Electron Device Letters*, vol. 38, no. 12, pp. 1767–1770, 2017.

[30] C. M. Liyanagedera, A. Sengupta, A. Jaiswal, and K. Roy, "Stochastic spiking neural networks enabled by magnetic tunnel junctions: From nontelegraphic to telegraphic switching regimes," *Physical Review Applied*, vol. 8, no. 6, p. 064017, 2017.

[31] P. Debashis, R. Faria, K. Y. Camsari, and Z. Chen, "Design of stochastic nanomagnets for probabilistic spin logic," *IEEE Magnetics Letters*, vol. 9, pp. 1–5, 2018.

[32] D. Kappel, R. Legenstein, S. Habenschuss, M. Hsieh, and W. Maass, "A dynamic connectome supports the emergence of stable computational function of neural circuits through reward-based learning," *eNeuro*, vol. 5, no. 2, pp. ENEURO–0301, 2018.

[33] A. Sengupta, G. Srinivasan, D. Roy, and K. Roy, "Stochastic inference and learning enabled by magnetic tunnel junctions," in *2018 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2018, pp. 15–6.

[34] F. Zenke, G. Hennequin, and W. Gerstner, "Synaptic plasticity in neural networks needs homeostasis with a fast rate detector," *PLoS computational biology*, vol. 9, no. 11, p. e1003330, 2013.

[35] G. G. Turrigiano and S. B. Nelson, "Hebb and homeostasis in neuronal plasticity," *Current opinion in neurobiology*, vol. 10, no. 3, pp. 358–364, 2000.

[36] F. Jug, "On competition and learning in cortical structures," Ph.D. dissertation, ETH Zurich, 2012.

[37] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in computational neuroscience*, vol. 9, p. 99, 2015.

[38] M. Stimberg, D. F. Goodman, V. Benichoux, and R. Brette, "Equation-oriented specification of neural models for simulations," *Frontiers in neuroinformatics*, vol. 8, p. 6, 2014.

[39] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of physiology*, vol. 148, no. 3, pp. 574–591, 1959.

Steven D. Pyle received the M.Sc. degree in Electrical Engineering at the University of Central Florida and is currently a Ph.D. candidate in Computer Engineering at the University of Central Florida. His research interest lies in bridging the multidisciplinary fields of neuromorphic hardware with emerging devices, machine learning, computational neuroscience, and neurophysiology to realize brain-inspired computational hardware at ultra-low-power.

Ramtin Zand received B.Sc. degree in Electrical Engineering in 2010 from IKIU, Iran. He received his M.Sc. degree in Digital Electronics from Sharif University of Technology, Tehran, Iran, in 2012. He is a Ph.D. Candidate in Computer Engineering at the University of Central Florida (UCF), Orlando, FL. His research interests include Machine Learning and Neuromorphic Computing, Emerging Nanoscale Electronics including Spin-based Devices, Reconfigurable and Adaptive Computer Architectures, and Low-Power and Reliability-Aware VLSI Circuits.

Shadi Sheikhfaal received her B.Sc. degree in computer engineering from Azad University, Ardebil, Iran, in 2012 and her M.Sc. degree in computer engineering and computer systems architecture from Science and Research Branch of Azad University, Tehran, Iran, in 2014. She is currently pursuing her Ph.D. degree in computer engineering at University of Central Florida, Orlando, FL, USA. Her current research interests include brain inspired computing and spin-based computing.

Ronald F. DeMara (S87-M93-SM05) has been a full-time faculty member at the University of Central Florida since 1993. His interests are in computer architecture, post-CMOS devices, and reconfigurable fabrics with applications to intelligent and neuromorphic systems, on which he has published 250 articles and holds one patent. He is a Senior Member of IEEE and Topical Editor of IEEE Transactions on Computers and has been Keynote Speaker of IEEE RAW and IEEE ReConFig conferences, and Guest Editor of IEEE Transactions on Emerging Topics in Computing and IEEE Transactions on Computers 2017 Special Section on Innovation in Reconfigurable Fabrics and 2019 Special Section on Non-Volatile Memories. He received the Joseph M. Bidenbach Outstanding Engineering Educator Award from IEEE in 2008.

Supplementary Material

Steven D. Pyle, *Student Member, IEEE,* Ramtin Zand, *Student Member, IEEE,* Shadi Sheikhfaal, *Student Member, IEEE,* and Ronald F. DeMara, *Senior Member, IEEE*

***Abstract*—Detailed results, modeling, and simulation framework for the Neural Sampling Core is described. The SPICE circuits simulations in the presence of process variation, evolution of receptive fields, whole-system orientation selectivity, homeostatic behavior, and average spike rate results are presented. The simulation framework combining SPICE and a spiking neural network simulator is delineated.**

I. Introduction

THIS provides results, details, and justifications to the results achieved for the Neural Sampling Core (NSC) and how they were modeled and simulated.

The rest of the paper is organized as follows. Section II illustrates the results. Section III provides an additional discussions of the update phase.

II. Results

This section delineates the SPICE simulation parameters and results for the stochastic spiking neuron circuit, synapse circuit, and homeostatic synapse circuits, as well as the architectural simulation results from Brian2, a Spiking Neural Network Simulator [1]. All MOSFET models used 7nm high performance PTM FinFET models [2] with threshold voltages modified by a Gaussian distribution, $\mathcal{N}(0mV, 75mV)$, where $\mathcal{N}(\mu, \sigma)$ is a Gaussian distributed random variable with mean μ and standard deviation σ, to model effects of Process Variation (PV). All of the resistances of Magnetic Tunnel Junctions (MTJ) were modified from their ideal values with a Gaussian distribution with a mean of the ideal value and a standard deviation of 20% to model PV effects. The supply voltage was set to 200mV.

A. Stochastic Spiking Neuron

The stochastic spiking neuron as shown in Figure 2 of the main paper was first modeled by using SPICE simulations to obtain the spiking probabilities for all input voltages and then that behavior was modeled in Brian2 simulations. The low-energy barrier MTJ can be modeled by the stochastic Landau-Lifshitz-Gilbert (s-LLG) equation below [3].

$$\begin{aligned}(1+\alpha^2)d\hat{m}/dt = &-|\gamma|\hat{m}\times\vec{H} - \alpha|\gamma|(\hat{m}\times\hat{m}\times\vec{H}) \\ &+ 1/qN(\hat{m}\times\vec{I_S}\times\hat{m}) \\ &+ \Big(\alpha/qN(\hat{m}\times\vec{I_S})\Big)\end{aligned} \tag{1}$$

where α is the damping coefficient of the nanomagnet, γ is the electron gyromagnetic ratio, q is the electron charge, and $\vec{I_S}$ is the spin current applied to the free layer. The spin currents polarization, P, is equivalent to the polarization of the fixed layer, which is $\hat{z}$, and its amplitude is given by $\vec{I_S} = PI_c\hat{z}$, where I_c is the charge current flowing through the MTJ. N is the number of spins in the free layer, which is given by $N = M_s\text{Vol.}/\mu_B$, where M_s is the saturation magnetization, μ_B is the Bohr magneton, and Vol. is the volume of the nanomagnet. The effective field for the monodomain circular magnet used for the free layer is $\vec{H}$ is given as $-4\pi M_s m_x\hat{x} + \vec{H}_n$, $\hat{x}$ being the out-of-plane direction of the magnet. $\vec{H}_n$ is the isotropic thermal noise field, uncorrelated in three directions: $(H_n^{x,y,z})^2 = 2\alpha kT/(|\gamma|M_s\text{Vol.})$. However, simulating the s-LLG equation in SPICE requires a significant amount of time. Therefore, we utilized a compact Verilog-A model for simulation speed where a resistor was modeled that stochastically switched from $0.5M\Omega$ to $1.5M\Omega$ with a retention time of $\mathcal{N}(0.5ns, 0.5ns)$, a transition time of $\mathcal{N}(0.2ns, 0.05ns)$, and a minimum retention and transition time of $0.01ns$. This provided behavior that was qualitatively similar to the results provided by the s-LLG and described in [3]. The embedded p-bit with the compact stochastic MTJ model was connected to a D-Flip-Flop to estimate the probability of spiking at the clock edge and was simulated for 100 Monte-Carlo runs with a clock period of 10ns for input voltages ranging from 0mV to 200mV with steps of 1mV for 1000ns each, and the resulting probability of spiking for each run is shown in Figure 1a. Based on this result, we modeled the spiking probability, $\rho(t)$, of each neuron in Brian2 with equation 1

$$\rho(t) = \frac{1}{1+e^{-\alpha v(t)+\beta}} \tag{2}$$

Where $\alpha = 500$, $\beta = \mathcal{N}(75, 9.75)$, and $v(t)$ is the input voltage at time t. Figure 1b shows 50 samples of equation 1 used for neurons in Brian2, which is very close to the behavior obtained from SPICE simulations.

B. Hybrid Spin-CMOS Synapse

Modeling the hybrid spin-CMOS synapse and homeostatic synapse circuits in Brian2 is challenging due to the complexity of CMOS behavior, especially in the presence of process variations at subthreshold voltages. Our approach is to use 10000 monte carlo SPICE simulations for each possible synapse strength, W0-W5 for the input-output synapses and W0-W3 for the homeostatic synapses, to fit the voltage increase seen at SUM in Figure 5 of the main paper, V_{weight}, to a gamma distribution with shape parameter a and scale parameter b, and then model the synaptic strength in Brian2 with such a distribution. We used $R_{SUM} = 200k\Omega$, $R_P = \mathcal{N}(20M\Omega, 4M\Omega)$, which is the resistance of the parallel state of the SHE-MTJs, and $R_{AP} = \mathcal{N}(50M\Omega, 10M\Omega)$, which is the resistance of the anti-parallel state of the SHE-MTJs. The

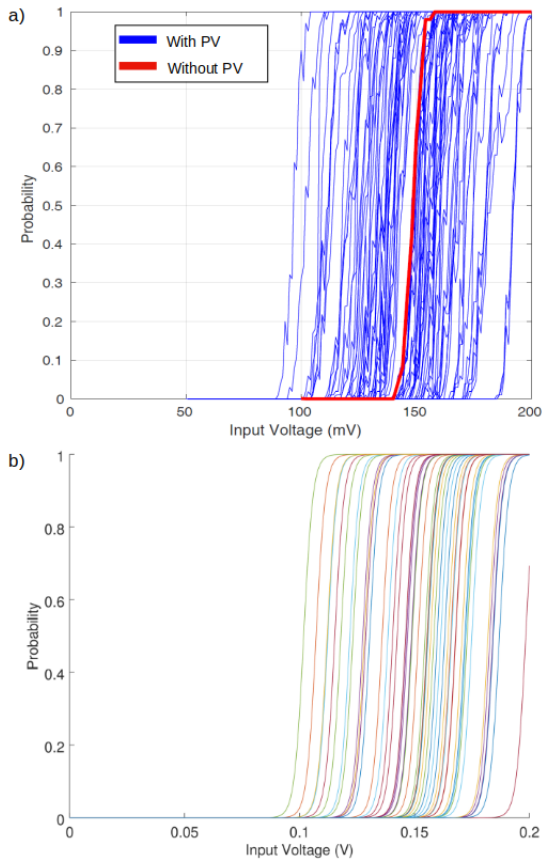

Fig. 1. SPICE and sigmoidal models of the probability of spiking for the stochastic spiking neuron. a) the probability of spiking for 100 monte carlo SPICE simulations of 100 clock periods at 1mV increments, b) the modeled sigmoidal probability of spiking used in Brian2.

TABLE I
SYNAPSE FITTING PARAMETERS

Weight	Gamma Parameters	
	a	b
W0	1.8496	1.50E-4
W1	1.8018	2.60E-4
W2	1.7275	4.03E-4
W3	1.8340	4.17E-4
W4	1.8008	9.19E-4
W5	1.7715	1.772E-3

resulting fitted gamma distribution parameters for the synapse and homeostatic synapses are listed in Tables I and II. The synaptic weights from the SPICE simulations as well as 10000 samples from a gamma distribution of the fitted parameters for each weight are shown in Figure 2, demonstrating conformity between the SPICE simulations and the modeled weights. The homeostatic synapse weights had a similar conformity, and we do not show them for brevity.

C. Unsupervised Learning

Additional figures for the unsupervised learning results are provided for the 20x20, 30x30, and 30x30 with noise case as described in the main text in Figures 3-5, respectively. The distribution of average spike rates are provided for each test case in Figure 6. Please note that all output neuron did spike

TABLE II
HOMEOSTATIC SYNAPSE FITTING PARAMETERS

Weight	S1	S2	Gamma Parameters	
			a	b
W0	P	AP	1.8311	1.95E-4
W1	P	P	1.8213	3.83E-4
W2	AP	AP	1.8320	3.84E-4
W3	AP	P	1.8232	1.181E-3

to some degree, and those with a count of 0 spikes simply had an average spike rate between 0 and 1.

III. DISCUSSION

A. Connections to Neural Sampling

Legenstein et al. [4] built upon Neural Sampling to theoretically analyze cortically-inspired Spiking Neural Network motiffs with a form of Hebbian plasticity, and demonstrated that such networks approximate inference in a noisy-OR-like generative model while learning through an approximation of online Expectation Maximization. They modeled the cortically-inspired network motiff, $\mathcal{M}$, with binary input responses, $\vec{y}(t)$, a population of excitatory neurons binary responses $\vec{z}(t)$ that have rectangular PSPs of duration τ whenever a spike occurs and a refractory period of duration τ, fixed inhibitory connections between neurons of weight β, and synaptic strengths $\mathbf{W}$. The resulting membrane potential $u_m(t)$ is given by

$$u_m(t) = \sum_{i}^{N} w_{im} y_i(t) - \sum_{j \neq m} \beta z_j(t) + c_m \tag{3}$$

where c_m is the homeostatic regulation of each individual neuron. The instantaneous Poisson spike rate is modeled by

$$\rho(t) = \frac{1}{\tau} exp(u(t)) \tag{4}$$

Buesing et al. [5] showed that for such a network, the states, $\vec{z}$, can be represented by a Boltzmann distribution:

$$p(\vec{z}|\vec{y}, \mathbf{W}) = \frac{1}{Z}(\gamma(\sum_{i,m} w_{im} y_i z_m + \frac{1}{2} \sum_{m \neq l} \beta z_m z_l + \sum_{m} c_m z_m)) \tag{5}$$

where Z is a normalizing constant and γ is a scaling parameter in the neuron's response function. Legenstein et al. [4] goes on to show that such a distribution can be interpreted as probabilistic inference of the current input, $y(\vec{t})$, and therefore, $\mathcal{M}$ has a generative model of the input distribution, $p(\vec{y}|\vec{z}, \mathbf{W})$, with a prior, $p(\vec{z})$, that corresponds to the network constraints, such as the inhibition, since

$$p(\vec{z}|\vec{y}, \mathbf{W}) = \frac{p(\vec{z}) p(\vec{y}|\vec{z}, \mathbf{W})}{\sum_{z'} p(\vec{y}|\vec{z'}, \mathbf{W})} \tag{6}$$

and that local learning rules can be used to find a local minimum of the Kullback-Leibler divergence between $p(\vec{y}|\mathbf{W})$ and the actual input distribution $p^*(\vec{y})$.

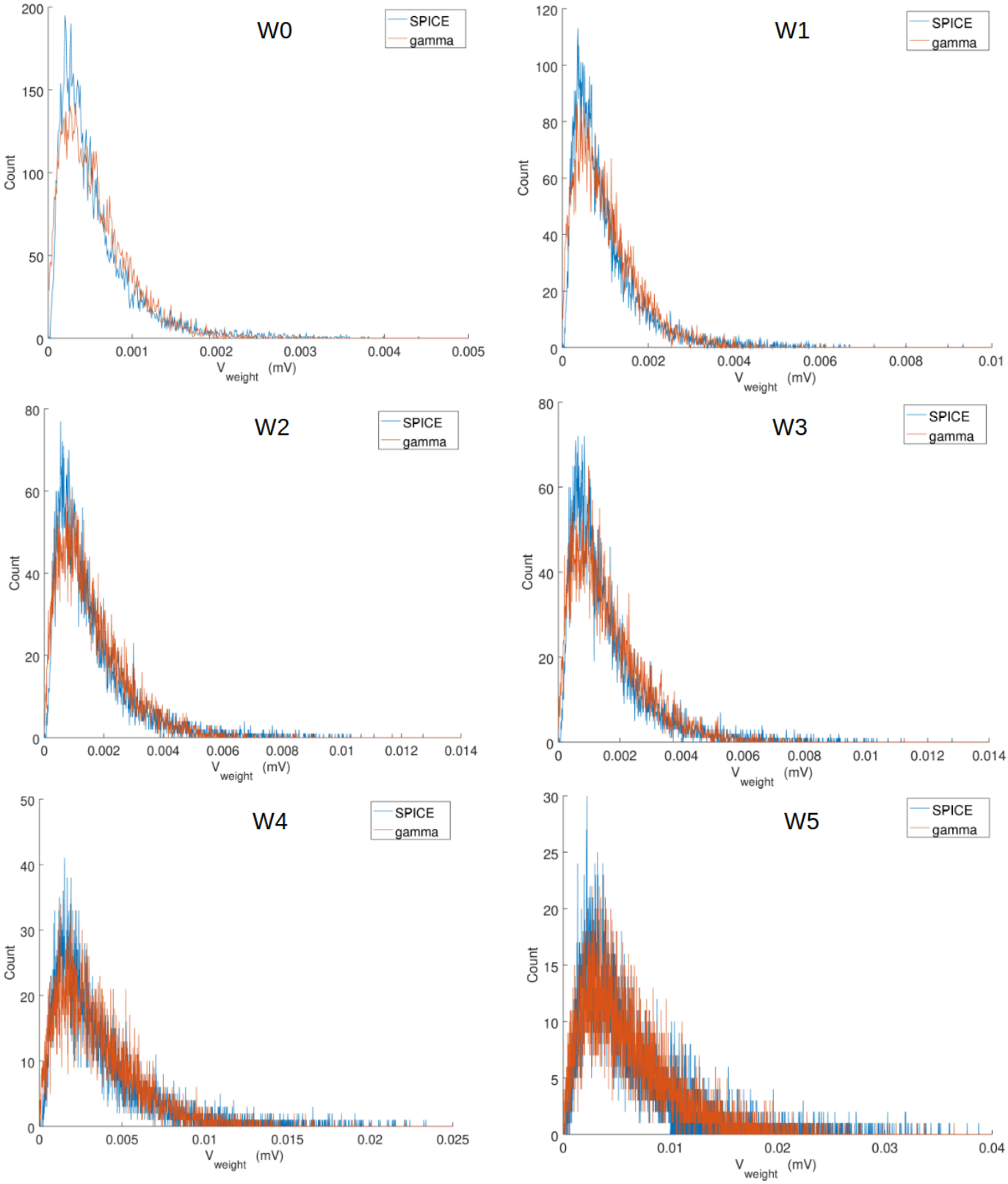

Fig. 2. The distribution of synaptic weights for each synapse circuit configuration obtained from 10000 monte-carlo SPICE simulations and 10000 samples from the fitted gamma distribution used for modeling the weights in Brian2.

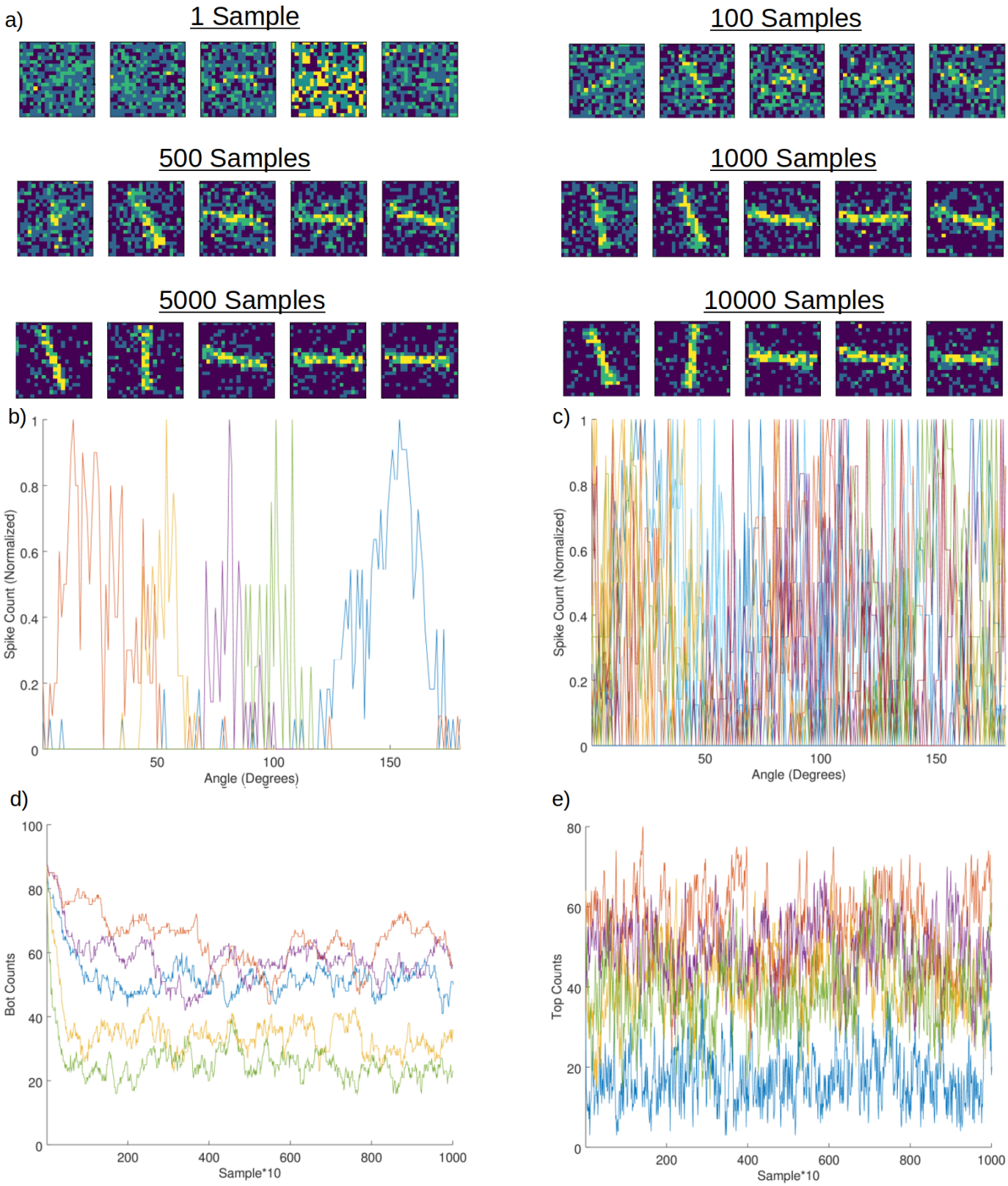

Fig. 3. Relevant figures for the 20x20 test case. a) the evolution of receptive fields for a random selection of five output neurons. b) the tuning curves for a random selection five output neurons. c) the tuning curves for all 50 output neurons. d) the temporal evolution of the number of homeostatic synapses with potentiated long-term SHE-MTJs (bot) for a random selection of five output neurons. e) the temporal evolution of the number of homeostatic synapses with potentiated short-term SHE-MTJs (top) for a random selection of five output neurons.

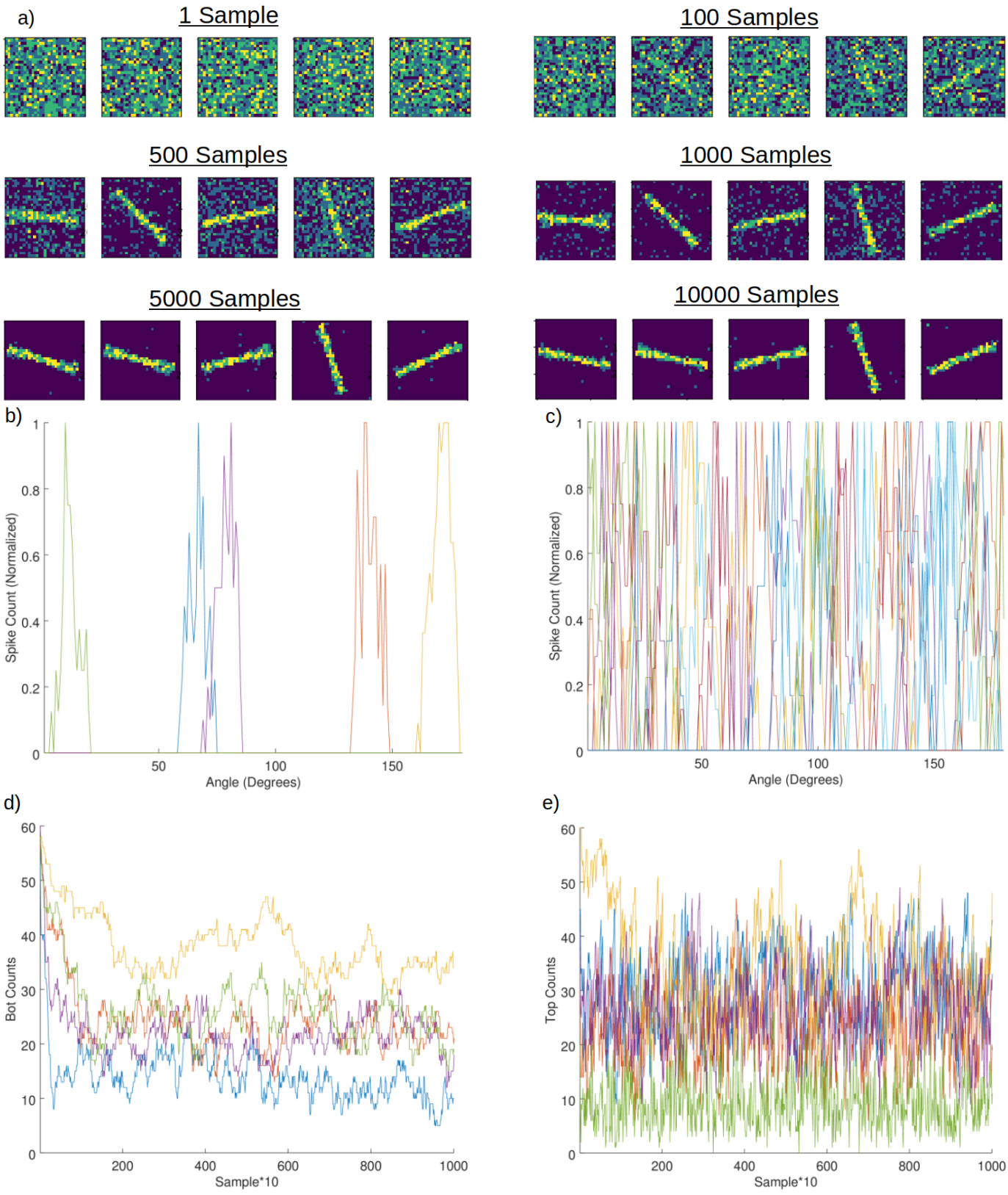

Fig. 4. Relevant figures for the 30x30 test case. a) the evolution of receptive fields for a random selection of five output neurons. b) the tuning curves for a random selection five output neurons. c) the tuning curves for all 50 output neurons. d) the temporal evolution of the number of homeostatic synapses with potentiated long-term SHE-MTJs (bot) for a random selection of five output neurons. e) the temporal evolution of the number of homeostatic synapses with potentiated short-term SHE-MTJs (top) for a random selection of five output neurons.

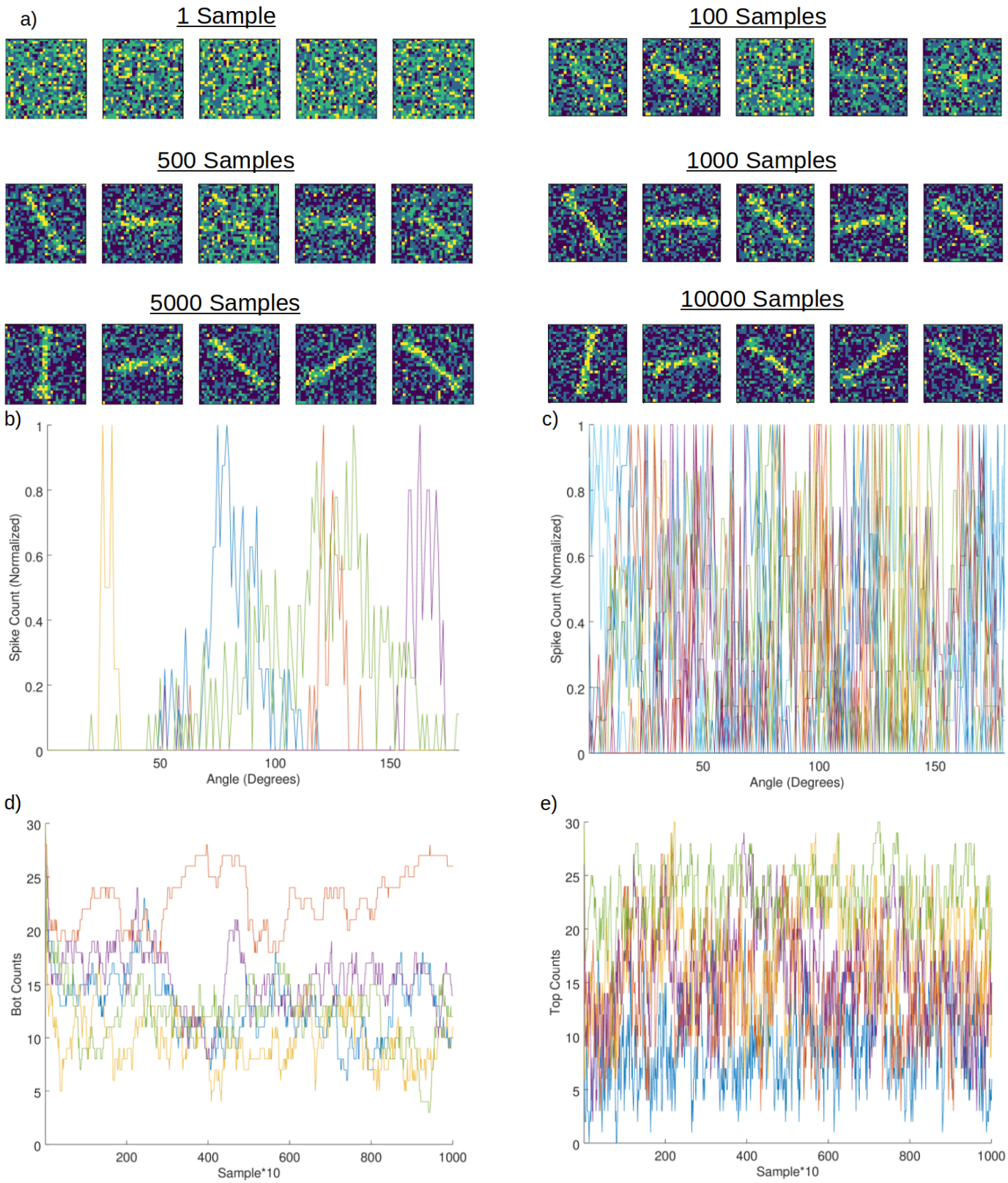

Fig. 5. Relevant figures for the 30x30 test case with noise. a) the evolution of receptive fields for a random selection of five output neurons. b) the tuning curves for a random selection five output neurons. c) the tuning curves for all 50 output neurons. d) the temporal evolution of the number of homeostatic synapses with potentiated long-term SHE-MTJs (bot) for a random selection of five output neurons. e) the temporal evolution of the number of homeostatic synapses with potentiated short-term SHE-MTJs (top) for a random selection of five output neurons.

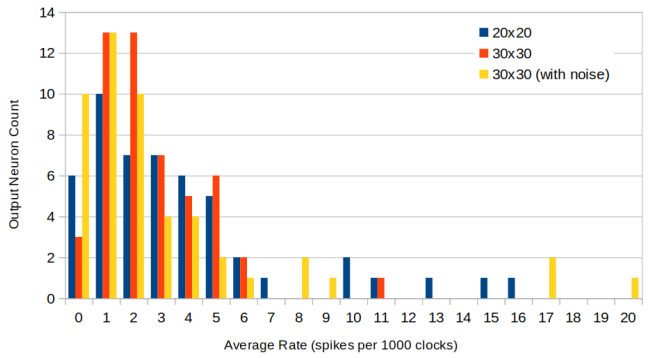

Fig. 6. The distribution of average spike rates for all 50 output neurons for each of the test cases.

The work herein take a circuit and architecture perspective towards implementing a form of this theoretical work by designing the stochastic spiking neuron to compute $\rho(t)$ and implement its associated PSP and refractory period, the low-precision hybrid spin-CMOS synapse to realize $\mathbf{W}$, the inhibitory feedback mechanism to compute $\beta z_j(t)$ in equation 2, the homeostasis mechanism to realize c_m, and Probabilistic Hebbian Plasticity to learn a generative model of the input distribution using only local information and low-precision components.

B. Update Phase

The primary motivations and contributions of this work is the investigation of using imprecise and stochastic components to realize robust neuromorphic hardware that has very low operational power. Therefore, since the stochastic switching behavior of spintronic devices is well established [6], and yet, highly dependent upon the device parameters and switching mechanism, for which there is no standard SHE-MTJ foundry process yet, determining the exact voltages and pulse durations needed for a given probability of switching will differ from any assumptions that could be made herein. Therefore, we model the update mechanisms outlined in Algorithm 1 of the main paper with a Gaussian-distributed probability of switching listed in Table III for each of the SHE-MTJs in the synapses, S1-S3, or the homeostatic synapses, S1-S2. Although the unsupervised learning results herein are obtained using the parameters listed in Table III, we also explored switching probabilites with standard deviations of up to 500% and found no qualitative differences in the results, illustrating that the exact switching probabilites are not important as long as the average behavior is similar to those in Table III, and therefore, such a scheme should be robust to process variations. It is also worth noting that the probability of switching for each SHE-MTJ is very small, and thus, would yield a very low-power update mechanism compared to approaches that would require larger switching probabilities. Also, if the update mechanism required too much power to update the entire NSC in parallel, time-multiplexing could be used to update smaller portions at a time, thanks to the non-volatility of the design.

TABLE III
SHE-MTJ SWITCHING PROBABILITY DURING EVENTS

Event	SHE-MTJ	P_{SW}
Synaptic potentiation	S1-S3	$\mathcal{N}(0.01, 0.0025)$
Synaptic depression	S1-S3	$\mathcal{N}(0.001, 0.00025)$
Homeostatic potentiation		
	S1	$\mathcal{N}(0.0001, 0.000025)$
	S2	$\mathcal{N}(0.00001, 0.0000025)$
Homeostatic depression		
	S1	$\mathcal{N}(0.01, 0.0025)$
	S2	$\mathcal{N}(0.001, 0.00025)$

C. Integration Complexity

The integration of multiple technologies on a single chip will always cause some increase in integration complexity. Additionally, since low-barrier spintronic devices are still an emerging research topic, it is not exactly understood what is the best method for fabricating and integrating these devices at scale. However, it is understood that the energy barrier can be manipulated by adjusting the volume of the device as well as the in-plane dimensions for in-plane devices [7]. Therefore, it could be possible to integrate both low-barrier and high-barrier devices on a single chip by adjusting the length and width of the MTJ structures, which can be done on the same mask, adding a minimal increase in integration complexity. It is true that the homeostatic synapse design in Figure 4a of the main paper utilizes SHE-MTJs with different energy barriers to realize different switching probabilities, which would certainly increase integration complexity to some degree, but it is a proposed trade off for reducing area and other complexities such as additional wires and signalling overheads. As shown in Figure 4b of the main paper, it is also possible to utilize SHE-MTJs with the same structures and dimensions to achieve the same results by using additional circuitry, wiring, and signaling overheads as discussed in the main paper.

D. Benefits of Homeostasis

Since homeostasis mechanisms are found in all biological neurons for ensuring balanced neural activity and are critical for realizing in circuit adaptivity to the effects of process variations, the NSC was designed from the beginning with a homeostatic mechanism in mind. Therefore, the input synapses of the NSC are designed as such that they typically won't generate enough voltage to drive the neuron to spike without any homeostatic synapses unless different design choices are made, like increasing the resistance of R_{SUM} or the width of $M1$ in the synapse circuit. Although this can mitigate the need for homeostatic synapses to elicit neuron spiking, the adaptivity that the homeostatic mechanism provides is critical for ensuring that all neurons participate in the network activity even with their intrinsic heterogeneity arising from process variation. This is demonstrated in Figure 7, that shows the temporal evolution of the 30x30 test case described previously, but with all homeostatic synapses fixed with S1 and S2 in the AP state. With no homeostatic adaptations, the neurons with intrinsically greater excitability dominate the network activity, and due to the inhibition mechanism, prevent intrinsically less

excitable neurons from ever spiking or learning, rendering them to be wasted area and power.

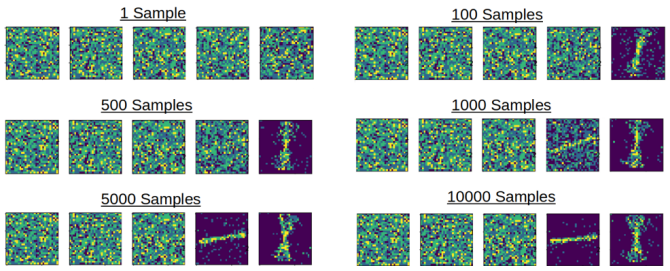

Fig. 7. Receptive fields for a random selection of 5 output neurons in the 30x30 test case with no homeostatic updates.

ACKNOWLEDGMENT

This work was supported in part by the Center for Probabilistic Spin Logic for Low-Energy Boolean and Non-Boolean Computing (CAPSL), one of the Nanoelectronic Computing Research (nCORE) Centers as task 2759.006, a Semiconductor Research Corporation (SRC) program sponsored by the NSF through CCF 1739635.

REFERENCES

[1] M. Stimberg, D. F. Goodman, V. Benichoux, and R. Brette, "Equation-oriented specification of neural models for simulations," *Frontiers in neuroinformatics*, vol. 8, p. 6, 2014.

[2] S. Sinha, G. Yeric, V. Chandra, B. Cline, and Y. Cao, "Exploring sub-20nm finfet design with predictive technology models," in *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*. IEEE, 2012, pp. 283–288.

[3] K. Y. Camsari, S. Salahuddin, and S. Datta, "Implementing p-bits with embedded mtj," *IEEE Electron Device Letters*, vol. 38, no. 12, pp. 1767–1770, 2017.

[4] R. Legenstein, Z. Jonke, S. Habenschuss, and W. Maass, "A probabilistic model for learning in cortical microcircuit motifs with data-based divisive inhibition," *arXiv preprint arXiv:1707.05182*, 2017.

[5] L. Buesing, J. Bill, B. Nessler, and W. Maass, "Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons," *PLoS computational biology*, vol. 7, no. 11, p. e1002211, 2011.

[6] T. Devolder, J. Hayakawa, K. Ito, H. Takahashi, S. Ikeda, P. Crozat, N. Zerounian, J.-V. Kim, C. Chappert, and H. Ohno, "Single-shot time-resolved measurements of nanosecond-scale spin-transfer induced switching: Stochastic versus deterministic aspects," *Physical review letters*, vol. 100, no. 5, p. 057206, 2008.

[7] C. M. Liyanagedera, A. Sengupta, A. Jaiswal, and K. Roy, "Stochastic spiking neural networks enabled by magnetic tunnel junctions: From non-telegraphic to telegraphic switching regimes," *Physical Review Applied*, vol. 8, no. 6, p. 064017, 2017.