Cross-homeostatic plasticity enables analog neuromorphic circuits to exhibit robust computational primitives

Maryada[1], Saray Soldado-Magraner[2], Martino Sorbaro[1,3], Rodrigo Laje[4], Dean V. Buonomano[2], and Giacomo Indiveri[1]

[1]Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland
[2]Department of Neurobiology, University of California, Los Angeles, USA
[3]ETH AI Center, ETH Zurich, Switzerland
[4]Department of Science and Technology, Universidad Nacional de Quilmes, Bernal, Argentina

Abstract

Local neural microcircuits in the brain implement numerous powerful computations including working memory, motor control, and input amplification. The computational primitives underlying such computations emerge from self-sustained patterns of neural activity and rely on the brain's ability to balance excitation and inhibition in recurrent neural circuits. Neuromorphic electronic circuits that use the physics of silicon to faithfully emulate neuronal dynamics represent a promising technology for reproducing the brain's computational primitives. However, achieving the same robustness of biological networks in artificial recurrent neural networks implemented on neuromorphic analog circuits remains a challenge, due to their high degree of heterogeneity and variability. Inspired by the self-calibration features of real neural networks, we apply a biologically-plausible cross-homeostatic learning rule to balance excitation and inhibition in neuromorphic implementations of recurrent spiking neural networks. We show how this learning rule allows the neuromorphic neural networks to overcome device mismatch and to autonomously produce stable self-sustained attractor dynamics in an inhibition-stabilized regime. We present experimental results that demonstrate how these attractors are formed reliably and robustly, for different populations of silicon neurons, both on the same chip and across different chips. Finally, we use these networks to implement reliable working memory and to reproduce biologically relevant emergent neural dynamics, including the so-called "paradoxical effect", a well-studied phenomenon that is observed in cortical circuits. In addition to validating neuroscience models with a computational substrate that shares the many similar properties and limitations of real neural processing systems, this work enables the construction of ultra-low power mixed-signal neuromorphic technologies that can be automatically configured to compute reliably, despite the large on-chip and chip-to-chip variations of their analog circuits.

1 Introduction

Animal brains can perform complex computations including sensory processing, motor control, and working memory, as well as higher cognitive functions, such as decision-making and reasoning, in an efficient and reliable manner. At the neural network level, these processes are implemented using a variety of computational primitives that rely on neural dynamics within recurrent neocortical microcircuits [1–6]. Translating the computational primitives observed in the brain into novel technologies can potentially lead to radical innovations in artificial intelligence and edge-computing applications. A promising technology that can implement these primitives with compact and low-power devices is that of mixed-signal neuromorphic systems, which employ analog electronic circuits to emulate the biophysics of real neurons [7–9]. As opposed to software simulations, the emulation performed by these electronic circuits relies on the use of the physics of the silicon substrate to faithfully reproduce the dynamics of neurons and synapses asynchronously in real time. While systems designed following this approach have the advantage of ultra-low power consumption, they have limitations and constraints similar to those found in biology. These include a high degree of variability, heterogeneity, and sensitivity to noise [10].

Due to these constraints, implementing recurrent networks in silicon with the same stability and robustness observed in biological circuits remains an open challenge.

Previous studies have already demonstrated neuromorphic implementations of recurrent computational primitives such as winner-take-all or state dependent networks [11–14]. However, these systems required manual tuning and exhibited dynamics that would not always produce stable self-sustained regimes. Manually tuning on-chip recurrent networks with large numbers of parameters is challenging, tedious, and not scalable. Furthermore, because of cross-chip variability, this process requires re-tuning for each individual chip. Even when a stable regime is achieved with manual tuning, changes in the network conditions (e.g., due to temperature variations, increase or decrease of input signals, etc.) would require re-tuning. Therefore, mixed-signal neuromorphic technology can greatly benefit from automatic stabilizing and tuning mechanisms that can overcome the challenges imposed by the analog substrate.

Neocortical computations rely heavily on positive feedback imposed by recurrent connections between excitatory neurons, which allows networks to perform complex time-dependent computations and actively maintain information about past events. Recurrent excitation, however, also makes neocortical circuits vulnerable to "runaway excitation" and epileptic activity [15, 16]. In order to harness the computational power of recurrent excitation and avoid pathological regimes, the neocortical circuits operate in an inhibition-stabilized regime, in which positive feedback is held in check by recurrent inhibition [17–20]. And there is evidence that inhibition-stabilized regimes may comprise the default awake cortical dynamic regime [19, 21]. It is generally accepted that neocortical microcircuits have synaptic learning rules in place to homeostatically balance excitation and inhibition in order to generate dynamic regimes capable of self-sustained activity [22–25]. Indeed, recently, a family of synaptic learning rules that differentially operate at excitatory and inhibitory synapses has been proposed, which can drive simulated neural networks to self-sustained and inhibition-stabilized regimes in a self-organizing manner [26]. This family of learning rules is referred to as being "cross-homeostatic" because synaptic plasticity at excitatory and inhibitory synapses is dependent on both the inhibitory and excitatory setpoints. Taking inspiration from self-calibration principles in biological networks we use these cross-homeostatic plasticity rules to guide neuromorphic circuits to balanced excitatory-inhibitory regimes in a self-organizing manner.

We show that these rules can be successfully employed to autonomously calibrate analog spiking recurrent networks in silicon (Section 3.1). Specifically, by automatically tuning all synaptic weight classes in parallel, the networks of silicon neurons converge to fully self-sustained attractor dynamics, and configure themselves into the inhibition-stabilized regime. In addition, the emergent neural dynamics obtained in this way express the "paradoxical effect", a characteristic signature of inhibition-stabilized networks, which is widely observed in cortical circuits (Section 3.3) [17, 19]. The plasticity rules proposed prove resilient to hardware variability and noise, also across different chips, as well as to different parameter initializations. We demonstrate that inhibitory plasticity (often neglected in neuromorphic electronic systems) is necessary for successful convergence (Section 3.2). We additionally show that using such plasticity rules long-term memories can be implemented without disrupting the learned dynamics (Section 3.4), and that multiple such networks can be implemented on a single chip robustly (Section 3.5).

These results validate the robustness of cross-homeostatic plasticity in a physical substrate that presents similar challenges to those of biological networks, unlike idealized digital simulations. Additionally, this approach provides the neuromorphic community with a reliable method to autonomously and robustly calibrating recurrent neural networks in future mixed-signal analog/digital systems that are affected by device variability (e.g., comprising also memristive devices [27, 28]). This will open the door to designing autonomous systems that can interact in real time with the environment, and compute reliably with low-latency at extremely low-power using attractor dynamics, as is observed in biological circuits.

2 Methods

2.1 Neuromorphic hardware: the spiking neural network chip

In this study we use a recently-developed mixed-signal spiking neural network chip denoted as Dynamic Neuromorphic Asynchronous Processor (DYNAP-SE2) [29]. This chip comprises 1024 adaptive-exponential integrate-and-fire (AdEx I&F) neurons [8, 30], distributed across four cores. The neuron parameters consist of

C_w	0	1	2	3	4	5
I_C (nA)	0.07	0.55	4.45	35.0	280	2250

Table 1: Nominal estimated coarse current values of the on-chip bias generator circuit for synaptic weights.

its spiking threshold, refractory period, membrane time constant, and gain. On hardware, the inputs to each neuron are limited to 64 incoming connections, implemented using differential pair integrator (DPI) synapse circuits [31]. Each synaptic input can be configured to express one of four possible temporal dynamics (via four different DPI circuits): AMPA, NMDA, GABA-A, and GABA-B. Besides expressing different time constants, the excitatory synapse circuits (AMPA and NMDA) differ in that the NMDA circuit also incorporates a voltage gating mechanism. Similarly, the inhibitory synapses (GABA-A, and GABA-B) differ in their subtractive (GABA-B) versus shunting (GABA-A) effect.

The bias parameters that determine the weights, neuronal and synaptic properties are set by an on-chip digital to analog (DAC) bias-generator circuit [32] and are shared globally within each core. These parameters are expressed as currents that drive the neuromorphic circuits and are represented internally through a "coarse" and a "fine" value, according to the following equation:

$$I = I_C(C_w)\frac{F_w}{256}, \tag{1}$$

where F_w represents the fine value and C_w the coarse one. The variable I_C represents a current estimated from the bias-generator circuit simulations, which depends on C_w as specified in Table 1.

Before training, the network is initialized with C_w randomly taking a value of 3, 4, or 5, and F_w uniformly distributed between 20 and 200. The corresponding currents will therefore range between 2.7 nA and 1757 nA. However, due to device mismatch and variability in the analog circuits, even when many circuits share the same nominal biases, the effective values will vary. On the DYNAP-SE2 and other mixed-signal neuromorphic processors implemented using similar technology nodes and circuit design styles, the coefficient of variation of these parameters is typically around 20% [10]. However, the learning rules proposed in this work make these networks robust to such variability and lead to a stable regime.

2.2 Network emulation

In the cortex, excitatory Pyramidal (Pyr) and inhibitory Parvalbumin-positive interneurons (PV) constitute the main neuronal subtypes and are primarily responsible for excitatory/inhibitory (E/I) balance [34, 35]. Pyr and PV neurons have different intrinsic biophysical properties: in terms of excitability, PV cells have a higher threshold and gain compared to Pyr neurons. These properties, along with other neuronal characteristics, such as refractory period and membrane time constants, contribute to the characteristic of the regular-spiking patterns of pyramidal cells and fast-spiking patterns of PV interneurons.

In our hardware experiments, we reproduce the properties of both Pyr and PV neuron types on the DYNAP-SE2 chip by tuning the refractory period, time constants, spiking threshold, and neuron gain parameters to match biologically measured values. Since parameters are shared by all neurons in a chip core, we assign each neuron population type to a different core (Fig. 1a). The neural dynamics and Input-Frequency (FI) curve of Fig. 1b-c show that the choice of parameters used for the silicon neurons is compatible with the behavior of biological PV and Pyr cells [35].

With these neuron parameters, we create an on-chip E/I network composed of 200 Pyr and 50 PV neurons. The two populations are sparsely connected with a 10% connection probability for each pair of neurons (both, within population and between populations). Synaptic weights take four sets of values shared globally among all Pyr-to-Pyr (w_{ee}), Pyr-to-PV (w_{ie}), PV-to-PV (w_{ii}), and PV-to-Pyr (w_{ei}) connections respectively. All excitatory weights represent synapses modeled as AMPA receptors, and inhibitory weights as GABA-A (somatic). Here, w_{ii} and w_{ei} are somatic inhibitory connections. Even though the weight parameters are shared, the synaptic efficacy of each synapse is subject to 20% variability due to device mismatch in the analog circuits [10].

During training synaptic weights are adjusted such that the average activity for both Pyr and PV population converge to a target firing rate (or *set-point*), and can thus maintain self-sustained activity in the absence of any

external input or bias current. To kickstart activity in the network, we provide a brief 4 ms external input of four spikes at 10 ms intervals to 80% of the Pyr cells, with random delays for each postsynaptic target neuron to introduce temporal sparsity to the initial response of the population. The network is emulated in real-time on the chip (i.e., the dynamics of the analog circuits evolve through physical time reproducing the dynamics of the biological neural circuits). Each experiment is run for 1 s and the spiking activity of the whole network is streamed to a computer, for data logging and analysis purposes.

In the analysis, the network response is measure by calculating the average firing rate of both neural populations, which is then used to calculate the error (*i.e.*, the difference between set-point and actual activity) that drives the weight update (see Section 2.3). As we are interested in the network's self-sustained behavior, we discard the response induced by the external input while calculating the firing rates (i.e., we ignore the first 60 ms of activity, which comprises the first 40 ms of external input and an additional stabilization window of 20 ms). During the initial stages of the training procedure, before E/I balance is reached, the network could generate non-sustained bursts of activity. Therefore, we calculate the firing rate by considering only the time window when the neural populations are active, i.e. the in-burst firing rate, as opposed to the average over the whole trial. We empirically find that this choice leads to the desired sustained activity.

2.3 Learning rule

To bring the recurrent dynamics to a fully self-sustained inhibition-stabilized regime we employ the recently proposed cross-homeostatic plasticity rule [26], which tunes all four synaptic weight classes in parallel.

In classic forms of homeostatic plasticity, both excitatory and inhibitory populations self-regulate calcium levels based on their own activity [23, 36]. In contrast to these classic forms, where a neuron tunes its weights to maintain its own average output of activity (a firing rate set-point), the cross-homeostatic weight update rule aims at bringing the neuron's presynaptic partners of opposite polarity to an average population firing rate setpoint, according to the following equations:

$$\begin{aligned} \Delta w_{ee} &= +\alpha \bar{E}(t)(I_{target} - \bar{I}(t)), & \Delta w_{ie} &= -\alpha \bar{E}(t)(E_{target} - \bar{E}(t)), \\ \Delta w_{ei} &= -\alpha \bar{I}(t)(I_{target} - \bar{I}(t)), & \Delta w_{ii} &= +\alpha \bar{I}(t)(E_{target} - \bar{E}(t)), \end{aligned} \tag{2}$$

where α is the learning rate ($\alpha = 0.05$), $\bar{E}(t)$ and $\bar{I}(t)$ are the average measured firing rates during a trial t and $E_{target} = 20$ Hz, $I_{target} = 40$ Hz are the desired firing rate set-points, for the Pyr and PV populations respectively. This type of plasticity can be interpreted biologically as having a target on the input current that a given neuron receives. For example, an excitatory neuron could have indirect access to the average firing rate of their presynaptic inhibitory neurons based on a running average of the total GABA it receives.

While biologically plausible target activities are typically lower (*e.g.*, 5 Hz for the E population and 14 Hz for the I one as proposed in [26]), in our experiments we had to use higher set-point rates to attain reliable behavior, because of the relatively small network size and the low probability of connections used. In Section 3.5 we quantify the effects of network size and connectivity with additional experiments.

As the weights are shared globally within each population, the learning rules are implemented at the whole population level. Therefore the network maintains a given firing rate at the population level, rather than at the individual neuron level (see [26] for a local implementation of the rules). However, the heterogeneity in the analog circuits and their inherent noise produce diverse behaviors among neurons, for example leading some of the neurons to fire intermittently at slightly increased rates (while maintaining the desired set-point firing rate at the full population level).

2.4 Chip-in-the-loop training

Although the DYNAP-SE2 chip can be tuned to exhibit biologically plausible dynamics, it does not incorporate on-chip synaptic plasticity learning circuits. To overcome this limitation we used a "chip-in-the-loop" learning paradigm: the activity produced by the on chip is transmitted to a computer, which analyzes the data at the software level and calculates the parameter updates, which are then sent back to the chip.

In particular, the network is emulated for 1 s on the chip. Each emulation is repeated five times to reduce the influence of noise and obtain more reliable results. The output spikes are sent back to the computer, which

averages the results, calculates the appropriate weight update, applies stochastic rounding (see below), and sends the updated weights to the chip for the next iteration (i.e., after 5 repetitions). After each iteration, we drain the accumulated residual current from both the neuron and synapse circuits. We stop the procedure after 400 trials, as we found it to be sufficient for the network to converge to the target firing rates regardless of the initial conditions. Note that most simulations converge to target firing rate before trial 200, which shows the long-term stability of the cross-homeostasis rules in a converged network.

For all experiments, we applied the same set of parameters to neurons and synapses at the beginning of the training, with the exception of the random initialization of weights. The training produced stable and reliable activity around the programmed set-points 100% of the time, without having to re-tune any of the neuron or synapse parameters (except the synaptic weights), across multiple days, and for multiple chips.

Stochastic rounding As the values of the synaptic weight parameters are set by the bias-generator DAC, the weight updates can only take discrete values $\Delta F_w = 0, \pm 1, \pm 2, \ldots$. For very large learning rates, this could lead to unstable learning, and for very low ones to no changes at all. To overcome the constraints of limited resolution weight updates, a common strategy used is to use stochastic rounding (SR) [37, 38]. We adopted the SR technique by interpreting the Δw calculated by the learning rule as a probability to increase or decrease the weight: The stochastic rounding operation is as follows:

$$\mathrm{SR}(x) = \begin{cases} \lfloor x \rfloor & \text{with prob. } x - \lfloor x \rfloor \\ \lceil x \rceil & \text{otherwise.} \end{cases} \tag{3}$$

where $\lfloor \cdot \rfloor$ indicates the floor operation and $\lceil \cdot \rceil$ the ceiling operation.

Values for the desired weight updates Δw are given by the learning rules (2). We also define upper and lower bounds for F_w, respectively $F_- = 20$ and $F_+ = 250$. We then obtain the new coarse and fine values C'_w, F'_w of the weight as follows:

$$\begin{cases} C'_w = C_w,\ F'_w = F_w + \mathrm{SR}(\Delta w) & \text{if } F_- \leq F_w + \mathrm{SR}(\Delta w) \leq F_+ \\ C'_w = C_w - 1,\ F'_w = F_+ & \text{if } F_w + \mathrm{SR}(\Delta w) < F_- \\ C'_w = C_w + 1,\ F'_w = F_- & \text{if } F_w + \mathrm{SR}(\Delta w) > F_+. \end{cases} \tag{4}$$

In other words, the fine value is updated with the stochastically-rounded version of Δw; when it oversteps the bound, the fine value is reset to the bound, and the coarse value is changed instead.

Note that this SR strategy is compatible with neuromorphic hardware implementations, as demonstrated in [39]. Our setup using a chip-in-the-loop enables the evaluation of different learning mechanisms and test the robustness of different design choices. The design of future on-chip learning circuit can therefore be informed by our results, following a hardware-software co-design approach.

3 Results

3.1 Convergence to stable self-sustained dynamics

As an overview of the experimental paradigm, the training procedure begins with random initialization of the network weights. Then, at each iteration, we record the resulting firing rates, calculate the weight update and apply the new weight values to the chip.

Fig. 2a illustrates an example of the evolution of firing activity during the course of training for Pyr and PV cells. Starting from a random value determined by the initialization of weights, the average rates converge to their target values. Fig. 2b plots the evolution of weights during training. On most iterations, only the fine value F_w is updated, but updates to the coarse value C_w occur occasionally (eq. 4). Even though the coarse updates cause abrupt jumps in the weight value (due to the nature of the bias-generator circuit implementation), the cross-homeostatic rule is robust enough to guide the network to a stable regime.

When repeating the training process from different initial conditions of randomly sampled weights, the weights converge to different values, which however produce the same desired network average firing rate [26].

Fig. 2c illustrates the rate space with multiple initialization and convergence to respective target set-points (dashed lines). The root mean square error from the target is 1.02 Hz for $\bar{E}$ and 6.32 Hz for $\bar{I}$.

The converged weight values are approximately aligned to a linear manifold (Fig. 2d), with sets of weights w_{ii}, w_{ie} and w_{ee}, w_{ei} highly correlated (w_{ee} and w_{ei} : ρ_{green} = 0.81 and p-value = $1.55e^{-06}$; ρ_{orange} = 0.88, p-value = $6.6e^{-12}$; w_{ii} and w_{ie} : ρ_{green} = 0.63 and p-value = 0.0009; ρ_{orange} = 0.75, p-value = $3.9e^{-07}$, where *green* and *orange* represents two different chips). This is well aligned with the theoretical solution derived for rate E-I networks at their set-points, when the neuron transfer function is linear or threshold-linear [26, eqn. 4-5]. In our case, the linear transfer function assumption does not hold, because the AdEx silicon neuron models saturate at a rate that is inversely proportional to the refractory period parameter. However, the linear approximation holds well in the region of operation forced by the learning rule, set at relatively low firing rates (see Fig. 1b).

3.2 Inhibitory plasticity is necessary to achieve reliable convergence

The equations for the weights in absence of noise [26, eqn. 4-5] and the experimental results in Fig. 2d indicate that two of the four weight parameters, for example the inhibitory weights w_{ii} and w_{ei}, can be chosen arbitrarily, since one can always find a solution by learning the other two. This appears to suggest that homeostatic plasticity of the inhibitory weights might not be required in an ideal case. However, we experimentally show that excitatory plasticity alone is not sufficient to reliably converge in the presence of the noise and non-idealities of the analog substrate (see Fig. 3). In absence of inhibitory plasticity, both the excitatory and inhibitory populations approach their respective set-points, but they fail to converge to a stable activity levels. (Fig. 3a). In contrast, the full cross-homeostatic plasticity rule robustly drives the network activity towards the set-points after just a few trials (Fig. 2a). These results hold for multiple initializations (Fig. 3b), indicating that inhibitory plasticity is necessary for robust convergence.

Homeostatic plasticity in inhibitory neurons is less studied than in excitatory neurons [23, 36], although there is evidence that inhibitory neurons also regulate their activity levels based on activity set-points [40, 41]. Considering our on-chip results, we therefore speculate that homeostatic plasticity in all excitatory and inhibitory connections plays an important stabilization role also in biological systems.

3.3 Network is in an inhibition-stabilized asynchronous-irregular firing regime

Here we investigate the properties of the final network dynamics after learning. Fig. 4a shows a sample of activity of the converged network, as recorded from the chip, in which the excitatory and inhibitory populations exhibit asynchronous-irregular activity patterns. The low levels of correlations between neurons (Fig. 4b) are proof of asynchrony. Irregularity is shown by the coefficient of variation (CV^2) of most neurons being close to 1 (Fig. 4b). The CV^2 is computed as the variance of the inter-spike interval divided by its squared mean, and equals 0 for perfectly regular firing, and 1 for Poisson firing [42]. The asynchronous-irregular firing regime is what typical in realistic cortical networks [43].

As further evidence of the inhibition-stabilized regime of the network, we demonstrate the occurrence of the "paradoxical effect" in our deployed network [17, 19]: we inject an external depolarizing pulse to all PV neurons for 0.2 s and verify that the firing rate of both populations decreases for the duration of the stimulus, returning back to the target FR when the stimulus ceases (Fig. 4c).

3.4 Neuronal ensembles remain intact under cross-homeostasis

The effect of homeostasis is to maintain a set level of activity in neurons by scaling synaptic efficacy, as done in this work, or by changing the intrinsic properties of the neuron. This raises concerns about the stability – plasticity dilemma, e.g., the competition between the homeostatic stability mechanism and the associative forms of synaptic plasticity that contribute to the formation of functional *neural ensembles*, as they may counteract each other. A neural ensemble corresponds to a group of interconnected neurons (e.g., that increased their synaptic weights via learning) that contributes to the representation of a particular feature or concept. Typically, neurons forming an ensemble exhibit shared selectivity.

The ensemble could potentially be affected by the homeostatic scaling process, which attempts to equalize the activity levels across the network. To verify that the cross-homeostatic plasticity used in this work does not

interfere or disrupt the role of neural ensembles we artificially formed a "crafted memory" into the network during the homeostatic training process (Fig. 5). This memory is implemented as a neural ensemble within the excitatory population of the network. Specifically, the ensemble consists of 32 neurons, in which each neuron has a connection probability with other neurons of the same ensemble of 0.5, much higher than the rest of the network (Fig. 5a).

After 250 cross-homeostatic plasticity training iterations, we manually form the ensemble by introducing the additional connections in it. We observe an immediate increase in the overall activity of both excitatory and inhibitory neurons (Fig. 5b) after ensemble implantation, which is however reduced as the homeostatic training continues. The change in weights (Fig. 5c) illustrates how inhibitory plasticity (w_{ei}) compensates for increased activity, bringing the firing rates back to the target point. As the the values of weights within and outside of the ensemble remain differentiated (Fig. 5c), even after homeostatic plasticity converges, the neural ensemble remains intact and retains its selectivity for the specific feature it represents. We conclude that the cross-homeostatic learning rule brings the activity back to the target while preserving the implanted memory.

3.5 Multiple clusters with attractor dynamics

In the network discussed so far, we modeled a single stable point attractor using an E/I-balanced network. We used 200 out of the 256 neurons available on a single core of our multi-core neuromorphic chip (DYNAP-SE2).

Even though it is a small-scale network with sparse connectivity, for computational purposes, having a single attractor formed from 200 neurons is not an efficient and optimal utilization of the resources on the chip. Employing smaller networks would allow us to implement multiple attractors on a single chip, whereby each subnetwork could represent, for example, a distinct choice in a decision making task. The model could then be extended to introduce competition among the clusters leading to winner-take-all dynamics to facilitate decision-making processes.

Therefore, we also run scaled down variations of the previous network, and assess whether cross-homeostatic plasticity can still bring activity to desired levels. Table 2 summarizes the population count, connectivity scaling, and adjustment in the target firing rate that we adopt for each experiment.

Number of neurons (E/I)	Connection Probability	Target (E/I)
200/50	0.1	20/40 Hz
100/25	0.1	20/40 Hz
64/16	0.35	25/50 Hz
48/12	0.35	40/60 Hz

Table 2: Variation in network size, connection scaling, and stable target firing rate (set-points) for reliable self-sustained activity.

Furthermore, we experiment with implementing multiple copies of these scaled-down networks on the same chip. We create 5 instances of different E/I networks, each with 50 excitatory and 15 inhibitory neurons. The connectivities within each sub-network are drawn from the same distribution. Each subnetwork does not cross-talk with others. The configuration is illustrated in Fig. 6a. To implement this configuration on the chip, we group the excitatory neuron populations into a single core, and used another core for the inhibitory populations. All parameter values, including weights, are shared between sub-networks. The challenge is to check if the homeostatic plasticity converges for all subnetworks, since the effect of the weights is different for each E/I network due to the in-homogeneity of the chip.

During the training process, we sequentially stimulate one sub-network at a time during each iteration. We then compute the Δw based on the activity of the neuron population within that E/I network. The weight is configured on the chip after each iteration based on the average Δw across all sub-networks. Fig. 6b shows that all five subnetworks, with shared weights, can indeed converge to mean firing rate regimes close to the target firing rate.

4 Conclusions

In this paper, we demonstrated that cross-homeostatic plasticity can autonomously bring analog neuromorphic hardware to a fully self-sustained inhibition-stabilized regime, the regime that is thought to be the default state of the awake neocortex [18–21]. These are the first results to show how biologically inspired learning rules can be used to address a long-standing challenge in neuromorphic engineering: how to self-tune neuromorphic hardware into computationally powerful states while simultaneously solving the problem of heterogeneity and variability in the hardware.

Our results also validate the robustness of cross-homeostatic plasticity in a physical substrate that presents similar challenges to those of faced by biological networks. We also highlight the crucial contribution of inhibitory plasticity, specifically, that homeostatic plasticity of inhibitory connections is necessary to maintain stability in both biological and silicon networks. Notably, we constructed an on-chip stable attractor neural network that effectively maintains a long-term memory represented by a neural ensemble, and show that multiple attractors can stably coexist within the hardware-[17, 43]. In the future, this stable multi-attractor model can be extended to simulate decision-making processes, along with self-generated dynamics that are thought to underlie a number of computations including motor control, timing, and sensory integration [44–46].

By taking inspiration from the self-calibration features of biological networks, we achieve a remarkable level of resilience to hardware variability, enabling recurrent networks to maintain stability and self-sustained activity across different analog neuromorphic chips. Cross-homeostatic plasticity endows the instantiation of neural networks on different chips without the need for chip-specific tuning, thereby addressing the inherent mismatch in analog devices. Such bio-plausible algorithm provides a promising avenue for the development of robust and scalable hardware implementations of recurrent neural networks, while mitigating the challenges associated with the intrinsic mismatch in analog devices.

In short, these results not only provide insights into the underlying mechanisms configuring biological circuits, but also hold great promise for the development of neuromorphic applications that unlock the power of recurrent computations. This research presents a substantial step forward in our understanding of neural dynamics and paves new possibilities for practical deployment of neuromorphic devices that compute, like biological circuits, with extremely low-power and low-latency using attractor dynamics.

Author Contributions

Conceptualization: M, SSM and RL. Investigation: M, MS and SSM. Methodology, Software, Data curation and Visualization: M and MS. Supervision SSM, RL, DB and GI. Writing Original Draft: M and MS. Writing and editing: SSM, RL, DB and GI. Funding acquisition: SSM, MS, DB and GI.

Acknowledgments

We thank the organizers of the CapoCaccia Workshop Towards Neuromorphic Intelligence 2022, where this work was conceived. We thank Chenxi Wu, Ole Richter and German Koestinger for DYNAP-SE2 support, and Emre Neftci and Matthew Cook for useful discussions. SSM was supported by the Swiss National Science Foundation (SNSF) grant nos. P2ZHP3-187943 and P500PB-203133. M was supported by the SNSF Sinergia Project (CRSII5-180316). The work of MS was made possible by a postdoctoral fellowship from the ETH AI Center. GI was supported by the EU ERC Grant "NeuroAgents" (No. 724295), DVB and RL were supported by the National Institutes of Health grant no. NS116589. RL was supported by Universidad Nacional de Quilmes, CONICET, and NeurotechEU (Faculty Scholarship).

References

[1] Guillaume Hennequin, Tim P Vogels, and Wulfram Gerstner. “Optimal control of transient dynamics in balanced networks supports generation of complex movements”. In: *Neuron* 82.6 (2014), pp. 1394–1406.

[2] Dean V Buonomano and Wolfgang Maass. "State-dependent computations: spatiotemporal processing in cortical networks". In: *Nature Reviews Neuroscience* 10.2 (2009), pp. 113–125.

[3] Saurabh Vyas, Matthew D Golub, David Sussillo, and Krishna V Shenoy. "Computation through neural population dynamics". In: *Annual review of neuroscience* 43 (2020), pp. 249–275.

[4] Rodney J. Douglas and Kevan A.C. Martin. "Recurrent neuronal circuits in the neocortex". In: *Current Biology* 17.13 (2007), R496–R500. DOI: 10.1016/j.cub.2007.04.024.

[5] Rodney J. Douglas, Christof Koch, Misha Mahowald, Kevan A. C. Martin, and Humbert H. Suarez. "Recurrent Excitation in Neocortical Circuits". In: *Science* 269.5226 (1995), pp. 981–985. DOI: 10.1126/science.7638624.

[6] Xiao-Jing Wang. "Synaptic reverberation underlying mnemonic persistent activity". In: *Trends in neurosciences* 24.8 (2001), pp. 455–463.

[7] Carver Mead. "Neuromorphic Engineering: In Memory of Misha Mahowald". In: *Neural Computation* (2022), pp. 1–41. DOI: 10.1162/neco_a_01553.

[8] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri. "Neuromorphic electronic circuits for building autonomous cognitive systems". In: *Proceedings of the IEEE* 102.9 (Sept. 2014), pp. 1367–1388. ISSN: 0018-9219. DOI: 10.1109/JPROC.2014.2313954.

[9] C. Mead. "Neuromorphic Electronic Systems". In: *Proceedings of the IEEE* 78.10 (1990), pp. 1629–36. DOI: 10.1109/5.58356.

[10] Dmitrii Zendrikov, Sergio MG Solinas, and Giacomo Indiveri. "Brain-inspired methods for achieving robust computation in heterogeneous mixed-signal neuromorphic processing systems". In: *Neuromorphic Computing and Engineering* (2023).

[11] U. Rutishauser and R. Douglas. "State-Dependent Computation Using Coupled Recurrent Networks". In: *Neural Computation* 21 (2009), pp. 478–509.

[12] E. Neftci et al. "Synthesizing cognition in neuromorphic electronic systems". In: *Proceedings of the National Academy of Sciences* 110.37 (2013), E3468–E3476.

[13] M. Giulioni et al. "Robust working memory in an asynchronously spiking neural network realized in neuromorphic VLSI". In: *Frontiers in Neuroscience* 5.149 (2012). ISSN: 1662-453X. DOI: 10.3389/fnins.2011.00149.

[14] T. Pfeil et al. "Six networks on a universal neuromorphic computing substrate". In: *Frontiers in Neuroscience* 7 (2013). ISSN: 1662-453X. DOI: 10.3389/fnins.2013.00011.

[15] Mircea Steriade, David A. McCormick, and Terrence J. Sejnowski. "Thalamocortical oscillations in the sleeping and aroused brain". In: *Science* 262.5134 (1993), pp. 679–685.

[16] DAVID A McCormick. "GABA as an inhibitory neurotransmitter in human cerebral cortex". In: *Journal of neurophysiology* 62.5 (1989), pp. 1018–1027.

[17] Misha V Tsodyks, William E Skaggs, Terrence J Sejnowski, and Bruce L McNaughton. "Paradoxical effects of external modulation of inhibitory interneurons". In: *Journal of neuroscience* 17.11 (1997), pp. 4382–4388.

[18] Hirofumi Ozeki, Ian M Finn, Evan S Schaffer, Kenneth D Miller, and David Ferster. "Inhibitory stabilization of the cortical network underlies visual surround suppression". In: *Neuron* 62.4 (2009), pp. 578–592.

[19] Alessandro Sanzeni et al. "Inhibition stabilization is a widespread property of cortical networks". In: *Elife* 9 (2020), e54875.

[20] Sadra Sadeh and Claudia Clopath. "Inhibitory stabilization and cortical computation". In: *Nature Reviews Neuroscience* 22.1 (2021), pp. 21–37.

[21] Alain Destexhe, Stuart W Hughes, Michelle Rudolph, and Vincenzo Crunelli. "Are corticothalamic 'up'states fragments of wakefulness?" In: *Trends in neurosciences* 30.7 (2007), pp. 334–342.

[22] G.G. Turrigiano and S.N. Nelson. "Homeostatic plasticity in the developing nervous system". In: *Nature Reviews Neuroscience* 5 (Feb. 2004), pp. 97–107.

[23] Gina Turrigiano. "Homeostatic synaptic plasticity: local and global mechanisms for stabilizing neuronal function". In: *Cold Spring Harbor perspectives in biology* 4.1 (2012), a005736.

[24] Guillaume Hennequin, Everton J. Agnes, and Tim P. Vogels. "Inhibitory Plasticity: Balance, Control, and Codependence". In: *Annual Review of Neuroscience* 40 (2017), pp. 557–579. ISSN: 15454126. DOI: 10.1146/annurev-neuro-072116-031005.

[25] Robert C Froemke. "Plasticity of cortical excitatory-inhibitory balance". In: *Annual review of neuroscience* 38 (2015), pp. 195–219.

[26] Saray Soldado-Magraner, Michael J Seay, Rodrigo Laje, and Dean V Buonomano. "Paradoxical self-sustained dynamics emerge from orchestrated excitatory and inhibitory homeostatic plasticity rules". In: *Proceedings of the National Academy of Sciences* 119.43 (2022), e2200621119.

[27] Bipin Rajendran, Damien Querlioz, Sabina Spiga, and Abu Sebastian. "Memristive devices for spiking neural networks". In: *Memristive Devices for Brain-Inspired Computing*. Elsevier, 2020, pp. 399–405.

[28] Dennis Valbjørn Christensen et al. "2022 roadmap on neuromorphic computing and engineering". In: *Neuromorphic Computing and Engineering* (2022). DOI: 10.1088/2634-4386/ac4a83.

[29] O. Richter et al. "DYNAP-SE2: a scalable multi-core dynamic neuromorphic asynchronous spiking neural network processor". In preparation.

[30] G. Indiveri et al. "Neuromorphic silicon neuron circuits". In: *Frontiers in Neuroscience* 5 (2011), pp. 1–23. ISSN: 1662-453X. DOI: 10.3389/fnins.2011.00073.

[31] C. Bartolozzi and G. Indiveri. "Synaptic dynamics in analog VLSI". In: *Neural Computation* 19.10 (Oct. 2007), pp. 2581–2603. DOI: 10.1162/neco.2007.19.10.2581.

[32] T. Delbruck and A. Van Schaik. "Bias Current Generators with Wide Dynamic Range". In: *Analog Integrated Circuits and Signal Processing* 43.3 (2005), pp. 247–268.

[33] S.-C. Liu, J. Kramer, G. Indiveri, T. Delbruck, and R.J. Douglas. *Analog VLSI:Circuits and Principles*. MIT Press, 2002.

[34] Brielle R. Ferguson and Wen-Jun Gao. "PV Interneurons: Critical Regulators of E/I Balance for Prefrontal Cortex-Dependent Behavior and Psychiatric Disorders". In: *Frontiers in Neural Circuits* 12 (2018), p. 37. DOI: 10.3389/fncir.2018.00037.

[35] Juan L Romero-Sosa, Helen Motanis, and Dean V Buonomano. "Differential excitability of PV and SST neurons results in distinct functional roles in inhibition stabilization of up states". In: *Journal of Neuroscience* 41.34 (2021), pp. 7182–7196.

[36] Gina G Turrigiano and Sacha B Nelson. "Homeostatic plasticity in the developing nervous system". In: *Nature reviews neuroscience* 5.2 (2004), pp. 97–107.

[37] Lorenz K Muller and Giacomo Indiveri. "Rounding methods for neural networks with low resolution synaptic weights". In: *arXiv preprint arXiv:1504.05767* (2015), pp. 1–11.

[38] Matteo Croci, Massimiliano Fasi, Nicholas J Higham, Theo Mary, and Mantas Mikaitis. "Stochastic rounding: implementation, error analysis and applications". In: *Royal Society Open Science* 9.3 (2022), p. 211631.

[39] Matteo Cartiglia et al. "Stochastic dendrites enable online learning in mixed-signal neuromorphic processing systems". In: *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2022, pp. 1–5.

[40] Sandra J Kuhlman et al. "A disinhibitory microcircuit initiates critical-period plasticity in the visual cortex". In: *Nature* 501.7468 (2013), pp. 543–546.

[41] Zhengyu Ma, Gina G Turrigiano, Ralf Wessel, and Keith B Hengen. "Cortical circuit dynamics are homeostatically tuned to criticality in vivo". In: *Neuron* 104.4 (2019), pp. 655–664.

[42] Arvind Kumar, Sven Schrader, Ad Aertsen, and Stefan Rotter. "The high-conductance state of cortical networks". In: *Neural computation* 20.1 (2008), pp. 1–43.

[43] Alfonso Renart et al. “The asynchronous state in cortical circuits”. In: *science* 327.5965 (2010), pp. 587–590.

[44] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. “Context-dependent computation by recurrent dynamics in prefrontal cortex”. In: *Nature* 503.7474 (2013), pp. 78–84. ISSN: 00092665. DOI: 10.1038/nmeth.2835.Multiplexed.

[45] Vishwa Goudar and Dean V Buonomano. “Encoding sensory and motor patterns as time-invariant trajectories in recurrent neural networks”. In: *Elife* 7 (2018), e31134.

[46] Brian DePasquale, Christopher J Cueva, Kanaka Rajan, G Sean Escola, and LF Abbott. “full-FORCE: A target-based method for training recurrent networks”. In: *PloS one* 13.2 (2018), e0191527.

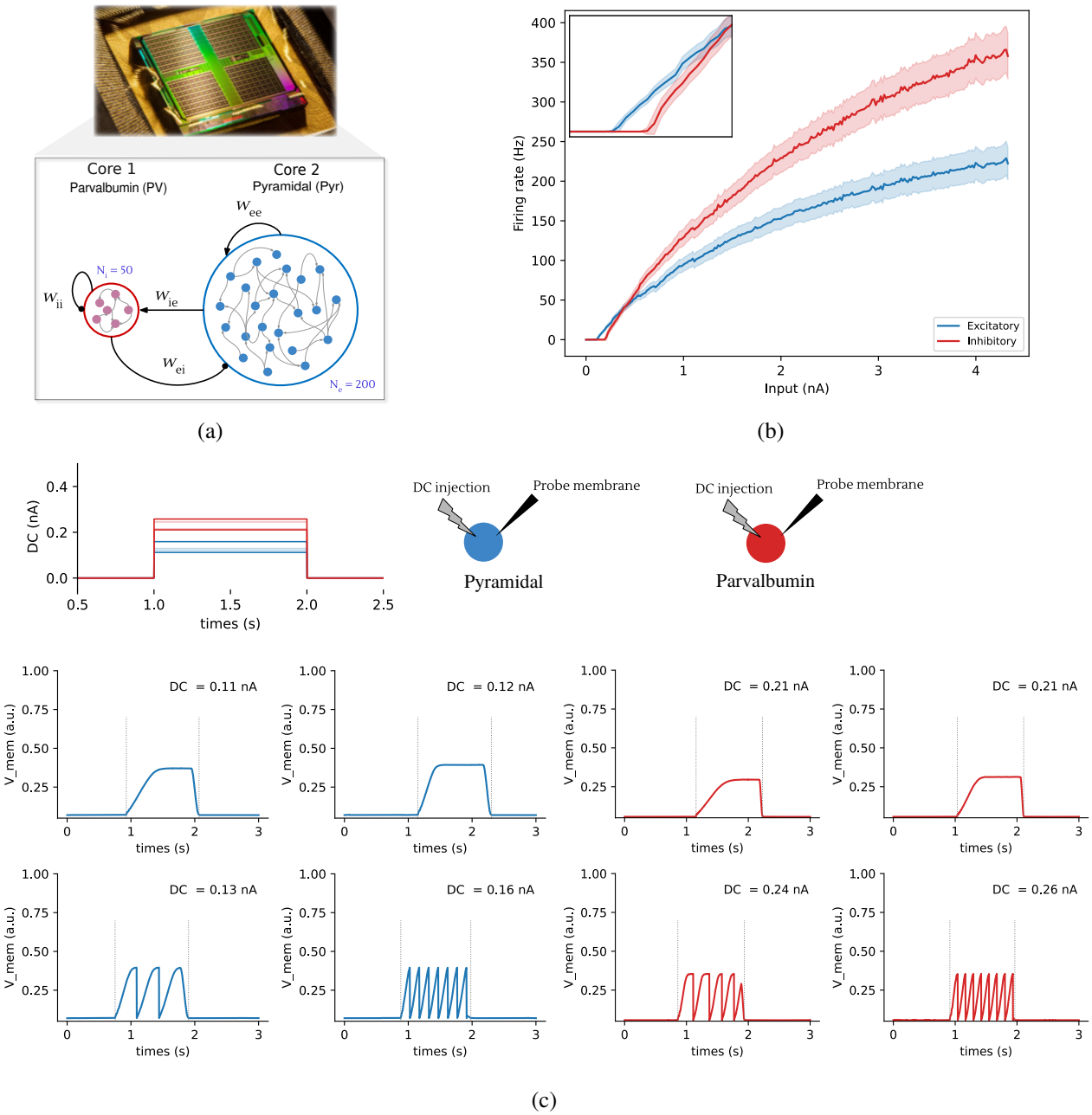


Figure 1: (a) Image of the DYNAP-SE2 chip, showing the four cores. An E-I network is implemented on chip. Excitatory (Pyr) and Inhibitory (PV) neurons occupy different cores, and are connected by four weight classes, Pyr-to-Pyr (w_{ee}), Pyr-to-PV (w_{ie}), PV-to-PV (w_{ii}), and PV-to-Pyr (w_{ei}). (b) Input-response curves for the resulting Pyr and PV neuron models on-chip. Inset: zoomed-in view at low firing rates. (c) Neuron membrane potential traces over time as a response to DC injection. Blue traces (left) are recorded from the chip core that contains the silicon neurons modeling the Pyramidal cell and red traces (right) are from the core used for the PV cells. The relationship between the measured voltage and the internal current-mode circuit output that represents the membrane potential is given by an exponential function that governs the subthreshold transistor operation mode [33]. (Top) Neuronal dynamics in the sub-threshold range. (Bottom) Neuronal dynamics as the membrane potential surpasses the spike threshold. On-set and off-set of DC injection are marked by grey lines.

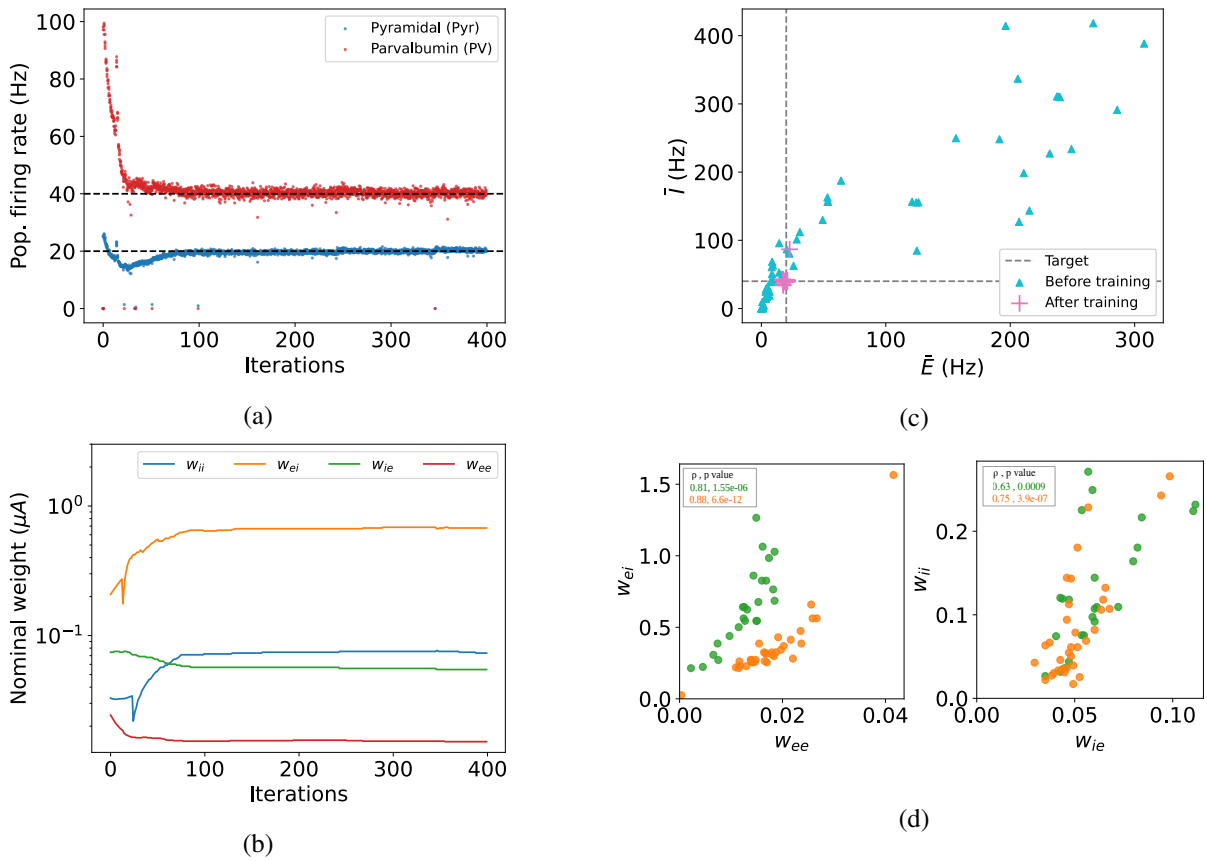


Figure 2: (a) A representative experiment of the convergence of excitatory (red) and inhibitory (blue) firing rates during learning. The gray dashed lines show the desired set-point targets. (b) A representative experiment (the same as in panel a) of the convergence of weight values during learning. The plotted value is nominal, inferred from eq. (1). (c) Initial and final firing rates across different chips (n=57, across two chip) and different initial conditions. For each initial condition, network connectivity was randomized. (d) Left: relationship between the final w_{ee} and w_{ei} values across different trials. Orange and green colors correspond to two different chips. Right: same, for w_{ii} and w_{ie} (axis in µA) .

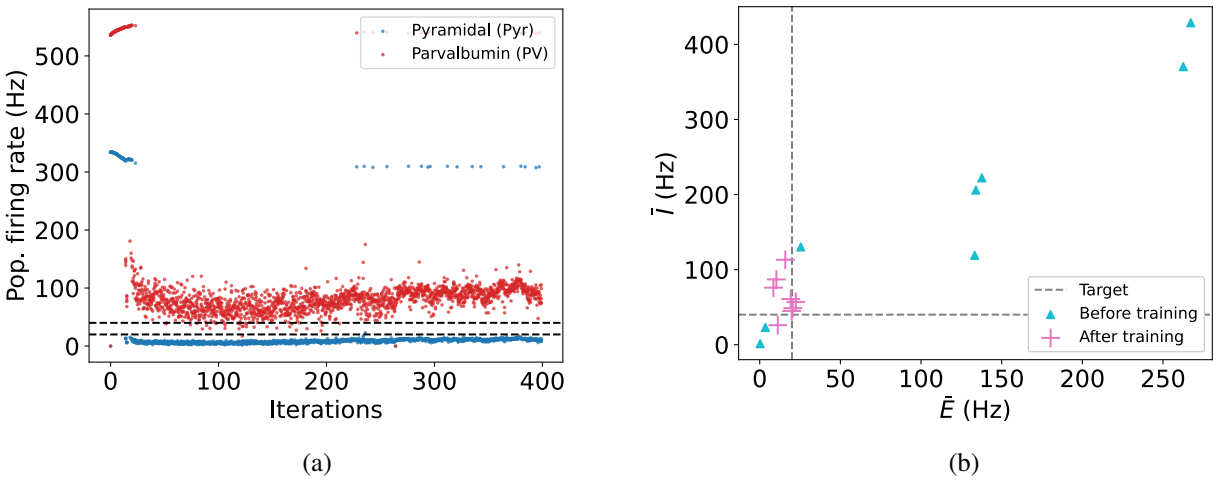
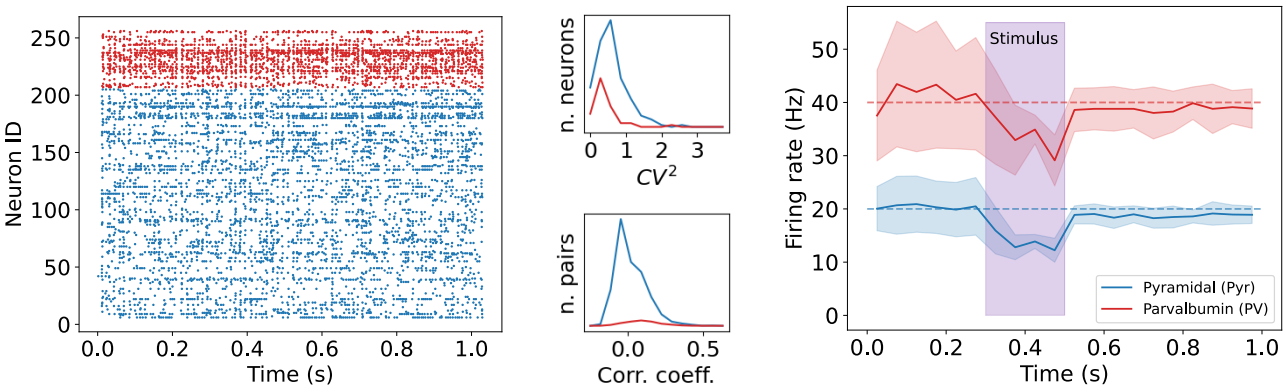
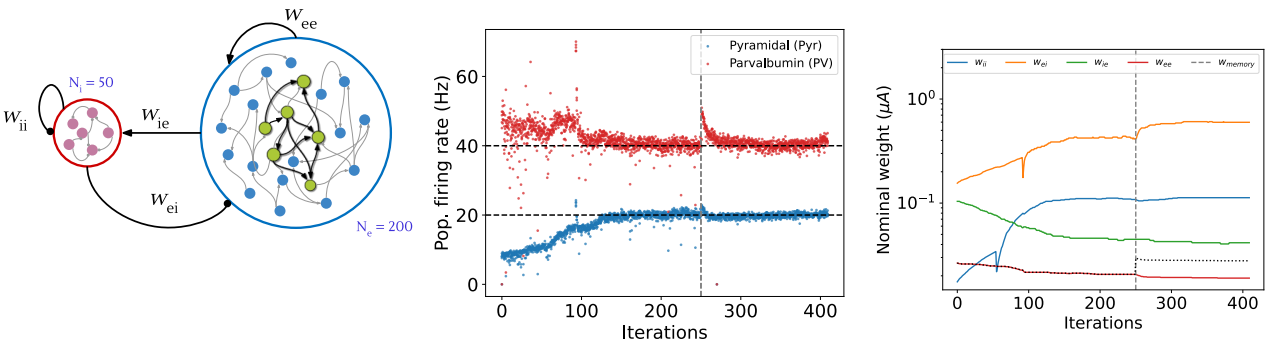


Figure 3: (a) Training results with "fixed" inhibitory synapse. Convergence is impaired if inhibitory plasticity is turned off. (b) Random networks ran with different initialization of all weights (n=8). The firing rates do not converge to the targets as precisely as in Fig. 2a, having a much higher root mean square error from the target (RMS on $\bar{E}$ 6.39, on $\bar{I}$ 35.11).



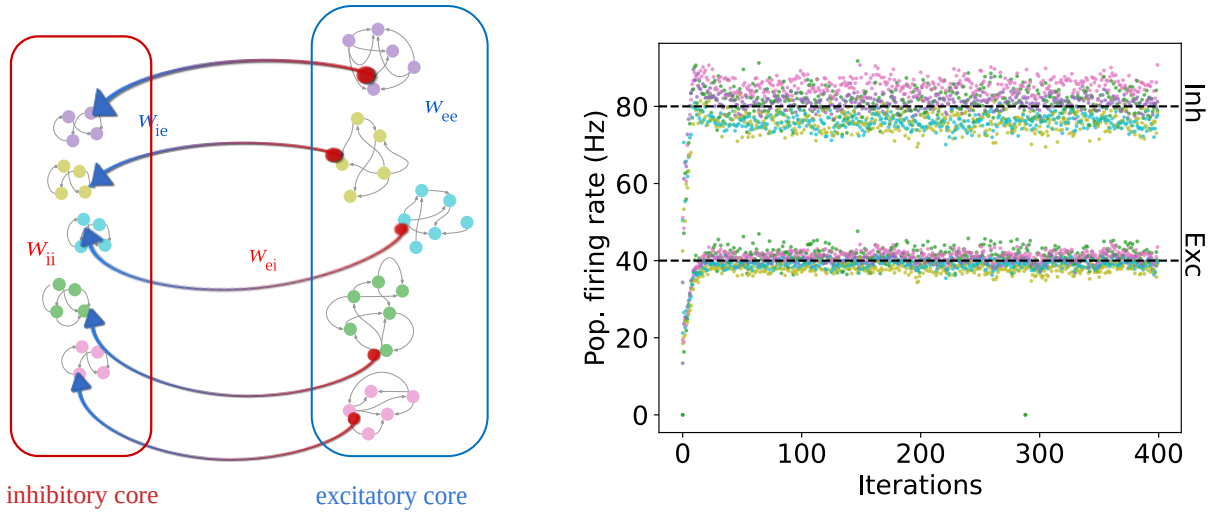
(a) (b) (c)

Figure 4: (a). Raster plot exemplifying on-chip firing activity during a single emulation run at the end of training. Both Pyr and PV neurons successfully converge to an asynchronous-irregular firing pattern at the desired FRs. (b). Distributions for the coefficient of variation (CV^2) and the 5 ms correlation coefficients between neurons, demonstrating low regularity and low synchrony. (c). The paradoxical effect can be demonstrated by applying a depolarizing current to the inhibitory units. The resulting decrease in the firing rate of the inhibitory units demonstrates that the on-chip network is in the inhibition-stabilized regime. No. of trials = 14, shading region represents standard deviation in firing rate across trials.



(a) (b) (c)

Figure 5: Training result with a memory implanted in the network. (a) Random recurrent E/I network with a neural ensemble (size = 32, green neurons). (b) Convergence of both excitatory (Pyr) and inhibitory (PV) populations to their respective set-point. A subset of the weights is manually changed to represent an ensemble at iteration 250, well after the network has converged to its set-point. A jump in activity is observed in both Pyr and PV cells due to implanted memory. (c) The weight (w_{ee}) value for 32 neurons is increased by a factor of 0.0088 µA, illustrated by a dotted line. All weights undergo alteration again to compensate for the ensemble, to bring the network back to the set point once memory is implanted. Although the memory remains intact, represented by higher recurrent weight compare to random excitatory connectivity (w_{ee}).



(a) (b)

Figure 6: Training results with multiple clusters. (a) Schematic of clusters on the chip. One core contains excitatory populations, and another contains inhibitory populations. The clusters are not connected to each other. (b) Rate convergence for 5 clusters sharing the same nominal weights. Each color represents one cluster.