Learning excitatory-inhibitory neuronal assemblies in recurrent networks

Owen Mackwood[1,2], Laura B. Naumann[1,2], and Henning Sprekeler*[1,2]

[1] *Berlin Institute of Technology, Marchstr. 23, 10587 Berlin, Germany*

[2] *Bernstein Center for Computational Neuroscience Berlin, Philippstr. 13, 10115 Berlin, Germany*

May 2020

Abstract

In sensory circuits with poor feature topography, stimulus-specific feedback inhibition necessitates carefully tuned synaptic circuitry. Recent experimental data from mouse primary visual cortex (V1) show that synapses between pyramidal neurons and parvalbumin-expressing (PV) inhibitory interneurons tend to be stronger for neurons that respond to similar stimulus features. The mechanism that underlies the formation of such excitatory-inhibitory (E/I) assemblies is unresolved. Here, we show that activity-dependent synaptic plasticity on input and output synapses of PV interneurons generates a circuit structure that is consistent with mouse V1. Using a computational model, we show that both forms of plasticity must act synergistically to form the observed E/I assemblies. Once established, these assemblies produce a stimulus-specific competition between pyramidal neurons. Our model suggests that activity-dependent plasticity can enable inhibitory circuits to actively shape cortical computations.

Introduction

With the advent of modern optogenetics, the functional role of inhibitory interneurons has developed into one of the central topics of systems neuroscience [Fishell and Kepecs, 2019]. Aside from the classical perspective that inhibition serves to stabilize recurrent excitatory feedback loops in neuronal circuits [van Vreeswijk and Sompolinsky, 1996, Brunel, 2000, Murphy and Miller, 2009, Sprekeler,

2017], it is increasingly recognised as an active player in cortical computation [Isaacson and Scanziani, 2011, Priebe and Ferster, 2008, Rubin et al., 2015, Pouille and Scanziani, 2001, Letzkus et al., 2011, Adesnik et al., 2012, Hennequin et al., 2014, Phillips et al., 2017, Barron et al., 2016, 2017, Tovote et al., 2015].

Within cortical neurons, excitatory and inhibitory currents are often highly correlated in their response to stimuli [Wehr and Zador, 2003, Froemke et al., 2007, Tan et al., 2011, Bhatia et al., 2019], in time [Okun and Lampl, 2008, Dipoppa et al., 2018] and across neurons [Xue et al., 2014]. This co-tuning of excitatory and inhibitory currents has been attributed to different origins. In topographically organised sensory areas such as cat primary visual cortex, the co-tuning with respect to sensory stimuli could be a natural consequence of local feedback inhibition and does not impose strong constraints on inhibitory circuitry [Harris and Mrsic-Flogel, 2013]. In the case of feedforward inhibition, co-tuning of excitatory and inhibitory currents was suggested to arise from homeostatic synaptic plasticity in GABAergic synapses [Vogels et al., 2011, Clopath et al., 2016, Weber and Sprekeler, 2018, Hennequin et al., 2017].

In sensory areas with poor feature topography, such as primary visual cortex of rodents [Ohki et al., 2005], feedback inhibition has been hypothesised to be largely unspecific for stimulus features, a property inferred from the dense connectivity [Fino and Yuste, 2011, Packer and Yuste, 2011] and reliable presence of synapses connecting pyramidal (Pyr) neurons to inhibitory interneurons with dissimilar stimulus tuning [Harris and Mrsic-Flogel, 2013, Bock et al., 2011, Hofer et al., 2011]. However, recent results cast doubt on this idea of a "blanket of inhibition" [Fino and Yuste, 2011, Packer and Yuste, 2011].

In mouse primary visual cortex (V1), Znamenskiy et al. [2018] report that although the presence of synaptic connections between Pyr cells and parvalbumin-positive (PV) interneurons is independent of their respective stimulus responses, the efficacy of those synapses is correlated with their response similarity, both in PV $\rightarrow$ Pyr and in Pyr $\rightarrow$ PV connections. These mutual preferences in synaptic organization suggest that feedback inhibition may be more stimulus-specific than previously thought and that Pyr and PV neurons form specialized—albeit potentially overlapping—excitatory-inhibitory (E/I) assemblies [Chenkov et al., 2017, Yoshimura et al., 2005, Litwin-Kumar and Doiron, 2012, 2014]. While the presence of such E/I assemblies [Znamenskiy et al., 2018, Rupprecht and Friedrich, 2018] suggests the need for an activity-dependent mechanism for their formation and/or refinement [Khan et al., 2018, Najafi et al., 2020], the requirements such a mechanism must fulfil remain unresolved.

Here, we use a computational model to identify requirements for the development of stimulus-specific feedback inhibition. We find that the formation of E/I assemblies requires a synergistic action

of plasticity on two synapse types: the excitatory synapses from Pyr neurons onto PV interneurons, and the inhibitory synapses from those interneurons onto the Pyr cells. Using "knock-out experiments", in which we block plasticity in either synapse type, we show that both must be plastic to account for the observed functional microcircuits in mouse V1. In addition, after the formation of E/I assemblies, perturbations of individual Pyr neurons lead to a feature-specific suppression of other Pyr neurons as recently found in mouse V1 [Chettih and Harvey, 2019]. Thus, synergistic plasticity of the in- and outgoing synapses of PV interneurons can drive the development of stimulus-specific feedback inhibition, resulting in a competition between Pyr neurons with similar stimulus preference.

Results

To understand which activity-dependent mechanisms can generate specific feedback inhibition in circuits without feature topography—such as mouse V1 (Fig. 1a), we studied a rate-based network model consisting of $N^{\mathrm{E}} = 512$ excitatory Pyr neurons and $N^{\mathrm{I}} = 64$ inhibitory PV neurons. To endow the excitatory neurons with a stimulus-tuning similar to pyramidal cells in layer 2/3 of mouse V1 [Znamenskiy et al., 2018], each excitatory neuron receives external excitatory input that is tuned to orientation, temporal frequency and spatial frequency (Fig. 1b). The preferred stimuli of the Pyr neurons cover the stimulus space evenly. Because we are interested under which conditions feedback inhibition can acquire a stimulus-selectivity, inhibitory neurons receive external inputs without stimulus tuning, but are recurrently connected to Pyr neurons. While the network has no stimulus topography, Pyr neurons are preferentially connected to other Pyr neurons with similar stimulus tuning [Hofer et al., 2011, Cossell et al., 2015], and connection strength is proportional to the signal correlation of their external inputs. Note that the Pyr $\rightarrow$ Pyr connections only play a decisive role for the results in Fig. 4, but are present in all simulations for consistency. Connection probability across the network is $p = 0.6$, with the remaining network connectivity (Pyr $\rightarrow$ PV, PV $\rightarrow$ PV, PV $\rightarrow$ Pyr) initialised randomly according to a log-normal distribution [Song et al., 2005, Loewenstein et al., 2011], with a variability that is similar to that measured in the respective synapses [Znamenskiy et al., 2018].

E/I assemblies are formed by homeostatic plasticity rules in input and output connections of PV interneurons

In feedforward networks, a stimulus-specific balance of excitation and inhibition can arise from homeostatic inhibitory synaptic plasticity that aims to minimise the deviation of a neuron's firing rate from a target for all stimuli of a given set [Vogels et al., 2011, Clopath et al., 2016, Weber and Sprekeler,

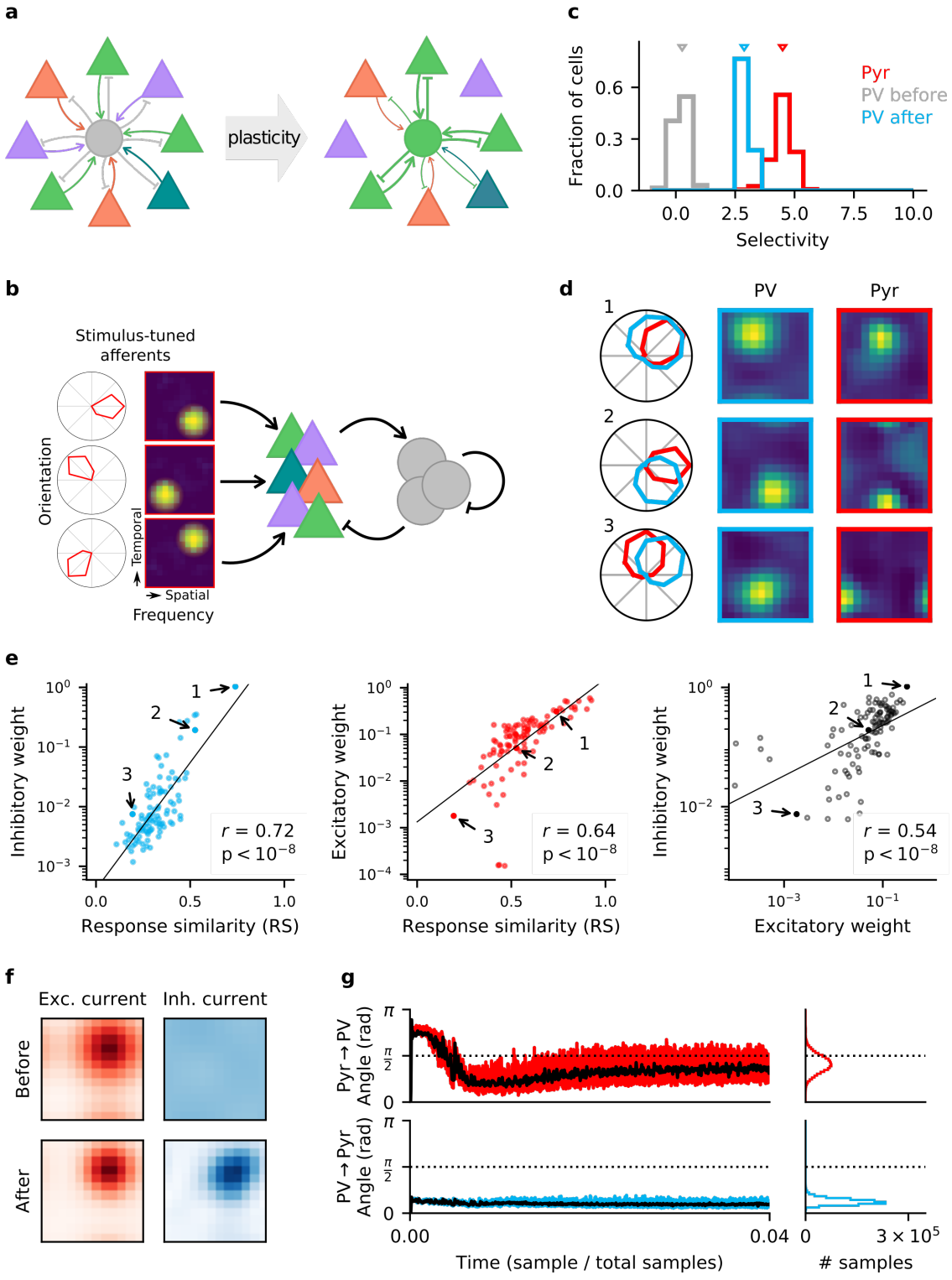

Figure 1: **Homeostatic plasticity in input and output synapses of interneurons drives the formation of E/I assemblies. a.** Emergence of E/I assemblies comprised of pyramidal neurons (triangles) and parvalbumin-expressing interneurons (circles) in circuits without feature topography. **b.** Network architecture and stimulus tuning of external inputs to pyramidal (Pyr) cells. **Continued on following page.**

Figure 1: **c.** Stimulus selectivity of Pyr neurons and PV interneurons (before and after learning). Arrows indicate the median. **d.** Example responses of reciprocally connected Pyr cells and PV interneurons. Examples chosen for large, intermediate and low response similarity (RS). Numbers correspond to points marked in (e). **e.** Relationship of synaptic efficacies of output (left) and input connections (centre) of PV interneurons with response similarity. Relationship of input and output efficacies (right). Black lines are obtained via linear regression. Reported r and associated p-value are Pearson's correlation. **f.** Stimulus tuning of excitatory and inhibitory currents onto an example Pyr cell, before and after learning. For simplicity, currents are shown for spatial and temporal frequency only, averaged across all orientations. **g.** Angle between the weight update and the gradient rule while following the local approximation for input (top) and output (bottom) connections of PV interneurons. Time course for first 4% of simulation (left) and final distribution (right) shown. Black lines are low-pass filtered time courses.

2018]. We wondered whether a stimulus-specific form of homeostasis can also generate stimulus-specific *feedback* inhibition by forming E/I assemblies. To that end, we derive synaptic plasticity rules for excitatory input and inhibitory output connections of PV interneurons that are homeostatic for the excitatory population (see Materials & Methods). A stimulus-specific homeostatic control can be seen as a "trivial" supervised learning task, in which the objective is that all pyramidal neurons should learn to fire at a given target rate ρ_0 for all stimuli. Hence, a gradient-based optimisation would effectively require a backpropagation of error [Rumelhart et al., 1985] through time [BPTT; Werbos, 1990].

Because backpropagation rules rely on non-local information that might not be available to the respective synapses, their biological plausibility is currently debated [Lillicrap et al., 2020, Sacramento et al., 2018, Guerguiev et al., 2017, Whittington and Bogacz, 2019, Bellec et al., 2020]. However, a local approximation of the full BPTT update can be obtained under the following assumptions: First, we assume that the sensory input to the network changes on a time scale that is slower than the intrinsic time scales in the network. This eliminates the necessity of backpropagating information through time, albeit still through the synapses in the network. This assumption results in what we call the "gradient-based" rules (Eq. 15 in the Supplementary Materials), which are spatially non-local. Second, we assume that synaptic interactions in the network are sufficiently weak that higher-order synaptic interactions can be neglected. Third and finally, we assume that over the course of learning, the Pyr $\to$ PV connections and the PV $\to$ Pyr connections become positively correlated [Znamenskiy et al., 2018], such that we can replace PV $\to$ Pyr synapses by the reciprocal Pyr $\to$ PV synapse in the Pyr $\to$ PV learning rule, without rotating the update too far from the true gradient (see Supplementary Materials).

The resulting learning rule for the output connections of the interneurons is similar to a previously suggested form of homeostatic inhibitory plasticity (Supp. Fig. S1a, left) [Vogels et al., 2011]. Specifically, PV output synapses $W^{\mathrm{E}\leftarrow\mathrm{I}}$ undergo Hebbian changes in proportion to presynaptic interneuron

activity r^{I} and the signed deviation of total postsynaptic pyramidal cell input h^{E} from the homeostatic target:

$$\Delta W_{ij}^{\mathrm{E}\leftarrow\mathrm{I}} \propto r_j^{\mathrm{I}}(h_i^{\mathrm{E}} - \rho_0) + \text{weight decay}\,.$$

In contrast, the PV input synapses $W^{\mathrm{I}\leftarrow\mathrm{E}}$ are changed such that the total excitatory drive $I_i^{\mathrm{E,\,rec}}$ from the Pyr population to each interneuron is close to some target value I_0 (Supp. Fig. S1a, right):

$$\Delta W_{ij}^{\mathrm{I}\leftarrow\mathrm{E}} \propto r_j^{\mathrm{E}}(I_i^{\mathrm{E,\,rec}} - I_0) + \text{weight decay}\,.$$

Both synapse types are subject to a weak weight decay, to avoid the redundancy that a multiplicative rescaling of input synapses can be compensated by a rescaling of the output synapses.

While our main results are obtained using the local approximations, we also simulated the gradient-based rules to verify that the approximation does not qualitatively change the results (Supp. Fig. S4).

When we endow the synapses of an initially randomly connected network of Pyr neurons and PV interneurons with plasticity in both the input and the output synapses of the interneurons, the network develops a synaptic weight structure and stimulus response that closely resemble that of mouse V1 [Znamenskiy et al., 2018]. Before learning, interneurons show poor stimulus selectivity (Fig. 1c), in line with the notion that in a random network, interneurons pool over many Pyr neurons with different stimulus tuning [Harris and Mrsic-Flogel, 2013]. The network is then exposed to randomly interleaved stimuli. By the end of learning, interneurons have developed a pronounced stimulus tuning, albeit weaker than that of Pyr neurons (Fig. 1c, d). Interneurons form strong bidirectional connections preferentially with Pyr neurons with a similar stimulus tuning, whereas connections between Pyr-PV pairs with dissimilar stimulus tuning are weaker (Fig. 1d, e). To make our results comparable to Znamenskiy et al. [2018], we randomly sample an experimentally feasible number of synaptic connections from the network ($n = 100$). Both the efficacy of PV input and output connections are highly correlated with the response similarity (see Materials & Methods) of the associated Pyr neurons and interneurons (Fig. 1e, left and center). For bidirectionally connected cell pairs, the efficacies of the respective input and output connections are highly correlated (Fig. 1e, right). The stimulus tuning of the inhibitory inputs onto the Pyr cells—initially flat—closely resembles that of the excitatory inputs after learning (Fig. 1f, Supp. Fig. S2) [Tan et al., 2011], i.e. the network develops a precise E/I balance [Hennequin et al., 2017].

Finally, the optimal gradient rules produce very similar results to the local approximations (Supp. Fig. S4). Over the course of learning, the weight updates by the approximate rules align to the updates

that would result from the gradient rules over (Fig. 1g, Supp. Fig. S3), presumably by a mechanism akin to feedback alignment [Lillicrap et al., 2016, Akrout et al., 2019].

In summary, these results show that combined homeostatic plasticity in input and output synapses of interneurons can generate a similar synaptic structure as observed in mouse V1, including the formation of E/I assemblies.

PV $\rightarrow$ Pyr plasticity is required for the formation of E/I assemblies

Having shown that homeostatic plasticity acting on both input and output synapses of interneurons are *sufficient* to learn E/I assemblies, we now turn to the question of whether both are *necessary*. To this end, we perform "knock-out" experiments, in which we selectively block synaptic plasticity in either of the synapses. The motivation for these experiments is the observation that the incoming PV synapses follow a long-tailed distribution [Znamenskiy et al., 2018]. This could provide a sufficient stimulus selectivity in the PV population for PV $\rightarrow$ Pyr plasticity alone to achieve a satisfactory E/I balance. A similar reasoning holds for static, but long-tailed outgoing PV synapses. This intuition is supported by result of Litwin-Kumar et al. [2017] that in a population of neurons analogous to our interneurons, the dimensionality of responses in that population can be high for static input synapses, when those are log-normally distributed.

When we knock-out output plasticity but keep input plasticity intact, the network fails to develop E/I assemblies and a stimulus-specific E/I balance. While there is highly significant change in the distribution of PV interneuron stimulus selectivity (Mann–Whitney U test, $U = 1207$, $p < 10^{-4}$), the effect is much stronger when output plasticity is also present (Fig. 2a,b). Importantly, excitatory and inhibitory currents in Pyr neurons are poorly co-tuned (Fig. 2c, Supp. Fig. S2b). In particular, feedback inhibition remains largely untuned because output connections are still random, so that Pyr neurons pool inhibition from many interneurons with different stimulus tuning.

To investigate whether the model without output plasticity is consistent with the synaptic structure of mouse V1, we repeatedly sample an experimentally feasible number of synapses ($n = 100$, Fig. 2d) and plot the distribution of the three pairwise Pearson correlation coefficients between the two classes of synaptic weights and response similarity (Fig. 2e). When both forms of plasticity are present in the network, a highly significant positive correlation ($p < 0.01$) is detected in all samples for all three correlation types (Fig. 2f). When output plasticity is knocked out, we still find a highly significant positive correlation between input weights and response similarity in 99% of the samples (Fig. 2d-f). In contrast, correlations between input and output synapses are weaker and cannot reliably be detected

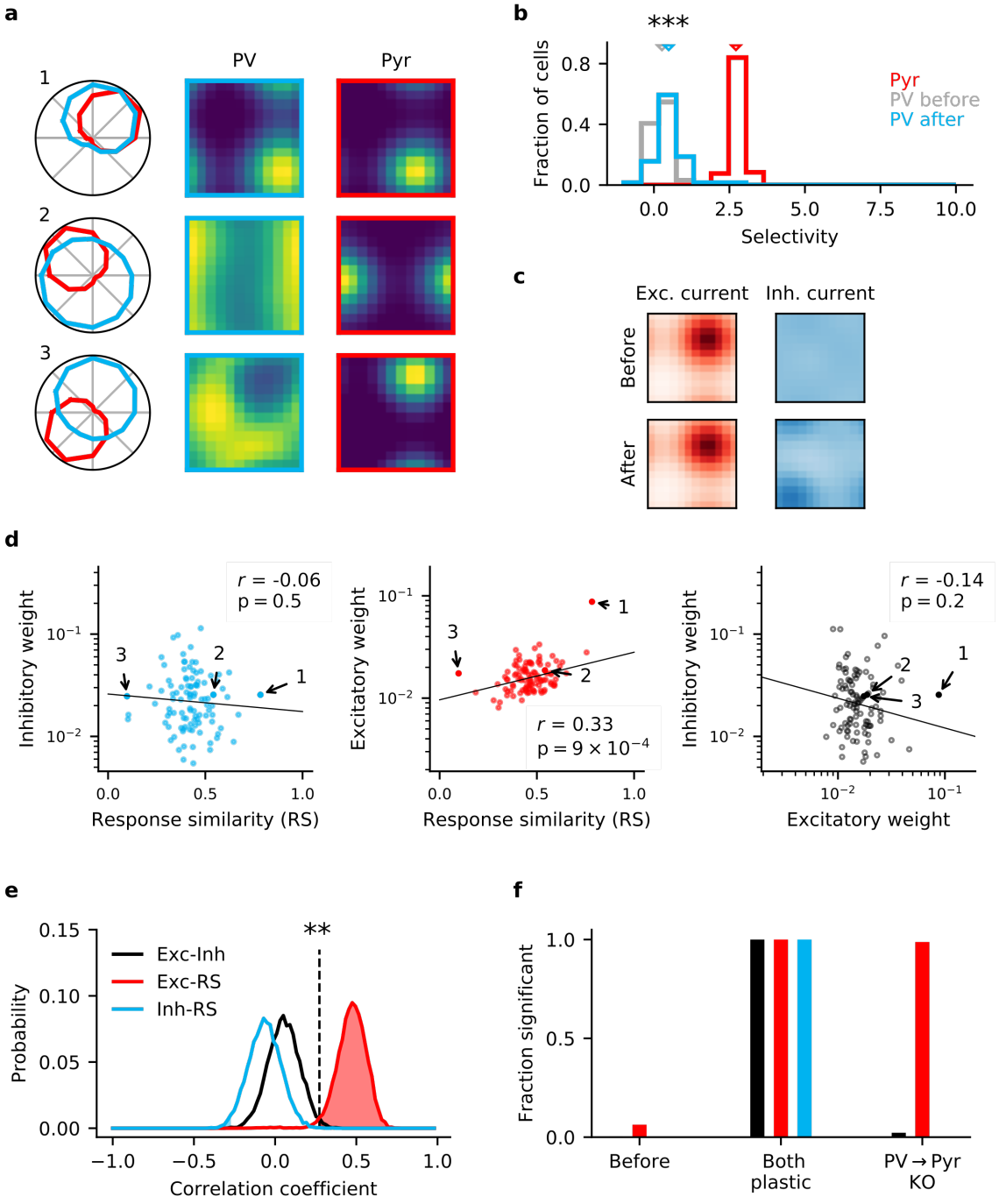

Figure 2: **Knock-out of plasticity in PV output connections prevents inhibitory co-tuning.** **a.** Example responses of reciprocally connected pyramidal (Pyr) cells and PV interneurons. Numbers correspond to points marked in (d). **b.** Stimulus selectivity of Pyr cells and PV interneurons (before and after learning; Mann–Whitney U test, $p < 10^{-4}$). Arrows indicate median. **c.** Stimulus tuning of excitatory and inhibitory input currents for two example Pyr cells. For simplicity, currents are shown for spatial and temporal frequency only, averaged across all orientations. **d.** Relationship of output (left) and input (centre) synaptic efficacies of PV interneurons with response similarity. Relationship of input and output efficacies (right). Plotted lines are obtained via linear regression. Reported r and associated p-value are the Pearson correlation. **Continued on following page.**

Figure 2: **e.** Distribution of Pearson correlation coefficients for multiple samples as shown in (d). Dashed line marks threshold of high significance ($p < 0.01$). **f.** Fraction of samples with highly significant positive correlation before plasticity, after plasticity in both input and output connections and for knock-out (KO) of plasticity in PV output connections (based on 10,000 random samples of 100 synaptic connections).

(2% of samples). Notably, we find a correlation between output weights and response similarity in 0.0% of samples (Fig. 2f). Finally, for an experimentally realistic sample size of $n = 100$, the probability of a correlation coefficient equal or higher than that observed by Znamenskiy et al. [2018] is 0.0% for the correlation between output weights and response similarity ($r = 0.55$), and 0.0% for the correlation between input and output synapses ($r = 0.52$).

The non-local gradient rule for the PV input synapses alone also does not permit the formation of E/I assemblies (Supp. Fig. S4a). While the selectivity of interneurons increases more than for the local approximation (Supp. Fig. S4b), feedback inhibition still remains untuned in the absence of output plasticity (Supp. Fig. S4c,d).

We therefore conclude that input plasticity alone is insufficient to generate the synaptic microstructure observed in mouse V1.

Pyr $\rightarrow$ PV plasticity is required for assembly formation

When we knock out input plasticity but keep output plasticity intact, we again observe no formation of E/I assemblies. This remains true even when using the gradient-based rule (Supp. Fig. S4). The underlying reason is that input weights remain random. Interneurons collect excitation from many Pyr neurons with different preferences, and absent plasticity on their input synapses, they maintain their initial poor stimulus selectivity (Fig. 3a-c). Because of the poor stimulus tuning of the interneurons, output plasticity cannot generate stimulus-specific inhibitory inputs to the Pyr neurons (Fig. 3d). Across the whole population, the similarity of excitatory and inhibitory currents onto Pyr neurons remains low (Supp. Fig. S2b,c).

Note that interneurons still possess a weak, but consistent stimulus tuning that arises from random variations in their input weights. A particularly strong input connection will cause the postsynaptic interneuron to prefer similar stimuli to the presynaptic Pyr. Because of the resulting correlated activity, the Hebbian nature of the output plasticity potentiates inhibitory weights for such cell pairs that are reciprocally connected. This tendency of strong input synapses to generate a strong corresponding output synapse is reflected in a positive correlation between them (Fig. 3e, Supp. Fig. S5a), despite the fact that input synapses remain random.

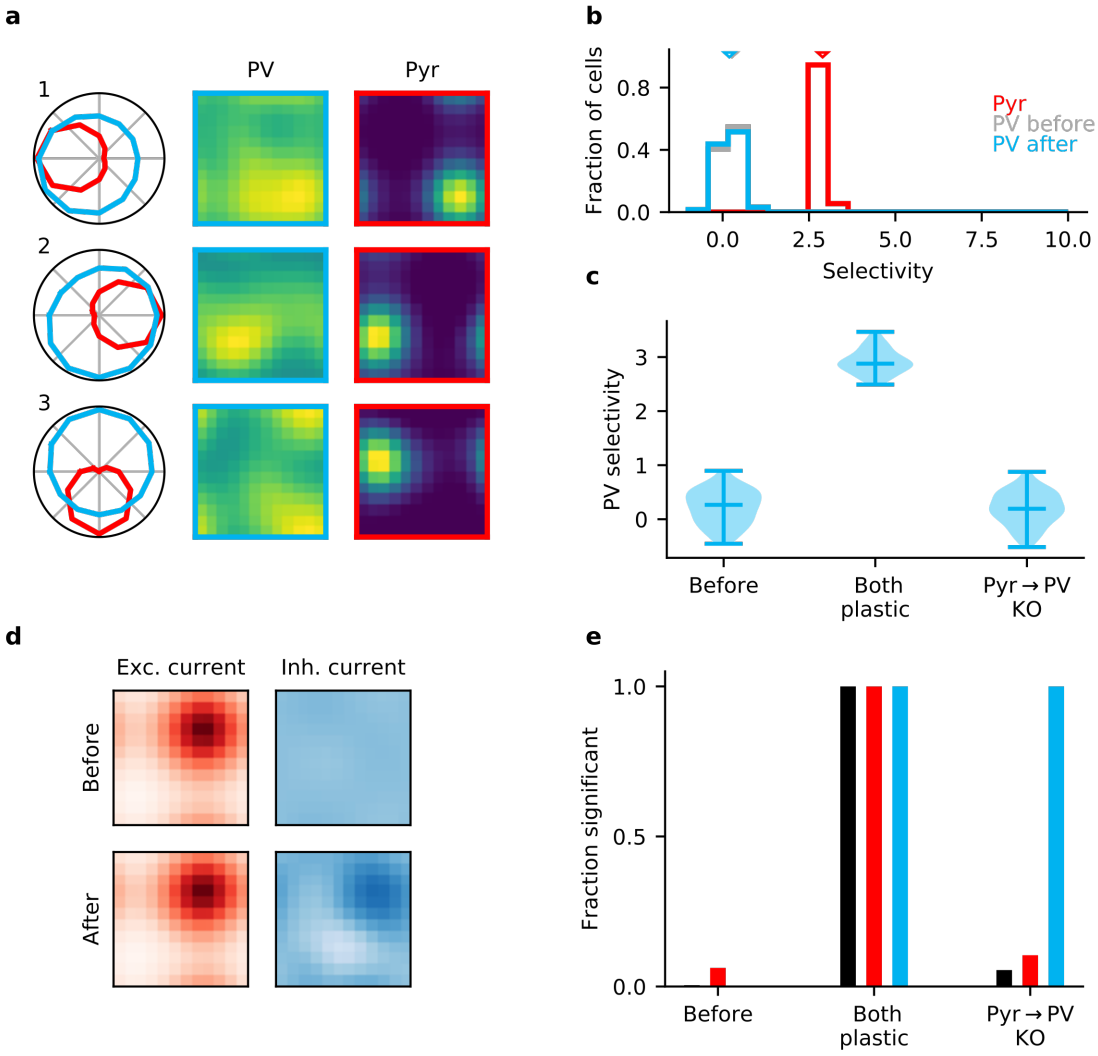

Figure 3: **Plasticity of PV input connections is required for inhibitory stimulus selectivity and current co-tuning. a.** Example responses of reciprocally connected pyramidal (Pyr) cells and PV interneurons. **b.** Stimulus selectivity of Pyr cells and PV interneurons (before and after learning). Arrows indicate median. **c.** Violin plots of inhibitory stimulus selectivity before plasticity, after learning with plasticity in both input and output connections of PV interneurons and for knock-out of plasticity in PV input connections. **d.** Stimulus tuning of excitatory and inhibitory currents in a Pyr cell before and after learning. Dimensions correspond to spatial and temporal frequency of the stimuli averaged across all orientations. **e.** Fraction of samples with highly significant ($p < 0.01$) positive correlation before plasticity, after plasticity in both input and output connections, and for knock-out (KO) of plasticity in PV input connections (based on 10,000 random samples of 100 synaptic connections).

Collectively, these results indicate that plasticity of both the inhibitory output and the excitatory input synapses of PV interneurons is required for the formation of E/I assemblies in cortical areas without feature topography, such as mouse V1.

Single Neuron Perturbations

Our findings demonstrate that in networks without feature topography, only a synergy of excitatory and inhibitory plasticity can account for the emergence of E/I assemblies. But how does stimulus-specific feedback inhibition affect interactions between excitatory neurons? In layer 2/3 of V1 similarly tuned excitatory neurons tend to have stronger and more frequent excitatory connections [Ko et al., 2011]. It has been hypothesised that this tuned excitatory connectivity supports reliable stimulus responses by amplifying the activity of similarly tuned neurons [Cossell et al., 2015]. However, the presence of co-tuned feedback inhibition could also induce the opposite effect, such that similarly tuned excitatory neurons are in competition with each other [Chettih and Harvey, 2019, Moreno-Bote and Drugowitsch, 2015].

To investigate the effect of stimulus-specific inhibition in our network, we simulate the perturbation experiment of Chettih and Harvey [2019]: First, we again expose the network to the stimulus set, with PV input and output plasticity in place to learn E/I assemblies. Second, both before and after learning, we probe the network with randomly selected stimuli from the same stimulus set, while perturbing a single Pyr cell with additional excitatory input, and measure the resulting change in activity of other Pyr neurons in the network (Fig. 4a).

While the activity of the perturbed neuron increases, many of the other Pyr neurons are inhibited in response to the perturbation (Fig. 4b). Although comparing the pairwise influence of Pyr neurons on each other does not reveal any apparent trend (Fig. 4c), recent experiments report that the influence a single-cell perturbation has on other neurons depends on the similarity of their stimulus feature tuning [Chettih and Harvey, 2019]. To test whether we observe the same feature-specific suppression, we compute the influence of perturbing a Pyr on the rest of the network as a function of the receptive field correlation of the perturbed cell and each measured cell. In line with recent perturbation studies [Chettih and Harvey, 2019, Sadeh and Clopath, 2020], we observe that—on average—neurons are more strongly inhibited if they have a similar tuning to the perturbed neuron (Fig. 4d). The opposite holds before learning: the effect of single-neuron perturbations on the network is increasingly excitatory as receptive field correlation increases. Notably, the networks in which input or output plasticity was knocked out during learning (and therefore did not develop E/I assemblies) show the same excitatory

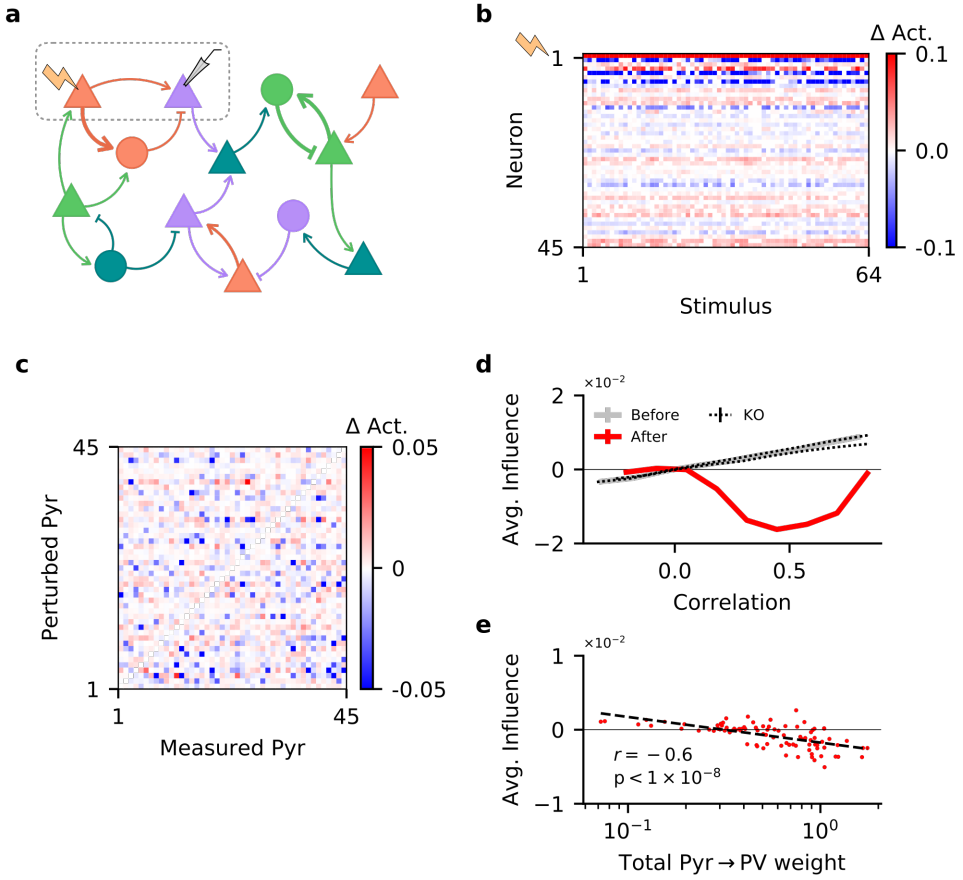

Figure 4: **Single neuron perturbations suppress responses of similarly tuned neurons. a.** Perturbation of a single pyramidal (Pyr) neuron. Responses of other Pyr neurons are recorded for different stimuli, both with and without perturbation. **b.** Perturbation-induced change in activity (Δ Act.) of a subset of Pyr cells, for a random subset of stimuli (with neuron 1 being perturbed). **c.** Influence of perturbing a Pyr neuron on the other Pyr neurons, averaged across all stimuli, for a subset of Pyr neurons. **d.** Dependence of influence among Pyr neurons on their receptive field correlation (Pearson's r), across all neurons in the network (see Materials & Methods). Dotted lines indicate plasticity knock-out (KO) experiments, see Supp. Fig. S6b for details. Error bars correspond to the standard error of the sample mean, but are not visible due to their small values. **e.** Total strength of output synapses from a Pyr neuron predicts the average effect perturbing it has on other neurons. Dashed line is the result of a linear regression, while r and its associated p-value correspond to the Pearson correlation.

effect (Fig. 4d, Supp. Fig. S6b). This confirms that a "blanket of inhibition" does not account for feature-specific suppression between excitatory neurons [Sadeh and Clopath, 2020].

To better understand this behaviour, we use the Pyr-Pyr receptive field correlations to compute the coefficient of determination for all pairs (R^2, which quantifies how well the receptive field of one Pyr neuron predicts that of another). Learning changes the correlative structure in the network (Supp. Fig. S6a), and thereby decreases the coefficient of determination on average, indicating a reduction in Pyr-Pyr correlations within the network ($\mathrm{E}[R^2] = 0.06$ before learning, 0.02 after). Thus, plasticity suppresses some of the strongest correlations, resulting in "feature competition" which is believed to aid sensory processing [Lochmann et al., 2012, Moreno-Bote and Drugowitsch, 2015].

While on average the network exhibits feature competition, the influence of individual Pyr neurons on the rest of the network is highly variable. According to recent modeling work [Sadeh and Clopath, 2020], the strength of Pyr $\rightarrow$ PV synapses strongly influences whether a network will exhibit feature competition. In our network, the total out-going weight of a Pyr cell onto the PV neurons indeed predicts the average influence that neuron will have on the rest of the network when perturbed (Fig. 4e; $r = -0.6$).

In summary, the stimulus-specific feedback inhibition that emerges in the model also captures the paradoxical suppression of similarly tuned excitatory neurons observed in single-cell perturbation experiments.

Discussion

The idea that feedback inhibition serves as a "blanket of inhibition" [Packer and Yuste, 2011, Fino and Yuste, 2011] that can be selectively broken [Karnani et al., 2016] has been gradually relaxed over recent years and replaced by the notion that feedback inhibition can be rather selective [Rupprecht and Friedrich, 2018] and could thereby support specific neuronal computations [Vogels and Abbott, 2009, Hennequin et al., 2014, Denève and Machens, 2016, Najafi et al., 2020], even in networks without topographic organisation [Znamenskiy et al., 2018, Rupprecht and Friedrich, 2018]. Here, we used a computational model to show that the development of E/I assemblies similar to those observed in mouse V1 [Znamenskiy et al., 2018] or zebrafish olfactory areas [Rupprecht and Friedrich, 2018] can be driven by a homeostatic form of plasticity of the in- and outgoing synapses of inhibitory interneurons. Based on the results of virtual knock-out experiments we suggest that, on their own, input or output plasticity of interneurons are insufficient to explain the Pyr-PV microcircuitry in mouse V1 and that input and output plasticity in interneurons must act in synergy for stimulus-

specific feedback inhibition to develop. To investigate how the presence of E/I assemblies affects interactions between excitatory neurons, we mimicked a perturbation experiment and found that—as in mouse visual cortex—stimulating single excitatory cells paradoxically suppresses similarly tuned neurons [Chettih and Harvey, 2019]. Our findings suggest that, by driving the development of tuned feedback inhibition, plasticity of interneurons can fundamentally shape cortical processing.

The learning rules for the input and output synapses of PV interneurons are based on a single homeostatic objective that aims to keep the net synaptic current onto Pyr neurons close to a given target for all stimuli. The two forms of plasticity fulfil different purposes, however. Plasticity of input synapses is required for interneurons to acquire a stimulus selectivity, whereas plasticity of output synapses can exploit interneuron selectivity to shape inhibitory currents onto excitatory cells. The output plasticity we derived for our recurrent network is very similar to a previously suggested form of inhibitory plasticity [Vogels et al., 2011, Sprekeler, 2017]. Homeostatic plasticity rules for inhibitory synapses are now used regularly in computational studies to stabilise model circuits [Vogels et al., 2011, Hennequin et al., 2017, Landau et al., 2016]. In contrast, a theoretically grounded approach for the plasticity of excitatory input synapses onto inhibitory neurons is missing.

Homeostatic changes in excitatory synapses onto interneurons in response to lesions or sensory deprivation have been reported [Keck et al., 2011, Takesian et al., 2013, Kuhlman et al., 2013], but the specific mechanisms and functions of this form of interneuron plasticity are not resolved. The plasticity rule we derived for the input synapses of interneurons effectively changes the selectivity of those neurons according to the demands of the Pyr cells, i.e. such that the interneurons can best counteract deviations of Pyr activity from the target. By which mechanisms such a (nearly teleological) form of plasticity can be achieved is at its core a problem of credit assignment, whose biological implementation remains open [Lillicrap et al., 2016, Guerguiev et al., 2017, Sacramento et al., 2018].

Here, we used a local approximation of the gradient, backpropagation rules, which produces qualitatively similar results, and which we interpret as a recurrent variant of feedback alignment, applied to the specific task of a stimulus-specific E/I balance [Lillicrap et al., 2016, Akrout et al., 2019]. The excitatory input connections onto the interneurons serve as a proxy for the transpose of the output connections. The intuition why this replacement is reasonable is the following: The task of balancing excitation by feedback inhibition favours symmetric connections, because excitatory cells that strongly drive a particular PV interneuron should receive a strong feedback connection in return. Therefore, E/I balance favours a positive correlation between the incoming and outgoing synapses of PV neurons and thus the two weight matrices will be aligned in a final balanced state [Lillicrap et al., 2016, Akrout et al., 2019]. This weight replacement effectively replaces the "true" feedback errors by a deviation

of the total excitatory input to the PV neurons from a target [Hertäg and Sprekeler, 2020]. The rule therefore has the structure of a homeostatic rule for the recurrent excitatory drive received by PV neurons.

A cellular implementation of such a plasticity rule would require the following ingredients: i) a signal that reflects the cell-wide excitatory current ii) a mechanism that changes Pyr $\rightarrow$ PV synapses in response to variations in this signal. On PV interneurons, NMDA receptors are enriched in excitatory feedback relative to feedforward connections [Le Roux et al., 2013]. Intracellular sodium and calcium could hence be a proxy of recurrent excitatory input. In addition, the activation of NMDA receptors has been shown to track intracellular sodium concentration [Yu and Salter, 1998] which at least partially reflects glutamatergic synaptic currents. Due to a lack of spines in PV dendrites, both postsynaptic sodium and calcium are expected to diffuse more broadly in the dendritic arbor [Hu et al., 2014, Kullmann and Lamsa, 2007], and thus might provide a signal for overall dendritic excitatory currents. Depending on how the excitatory inputs are distributed on PV interneuron dendrites [Larkum and Nevian, 2008, Jia et al., 2010, Grienberger et al., 2015], this integration does not need to be cell-wide, but could be local, e.g. to a dendrite, if the local excitatory input is a proxy for the global input. NMDA receptors at IN excitatory input synapses can mediate Hebbian long-term plasticity [Kullmann and Lamsa, 2007], and blocking excitatory currents can abolish plasticity in those synapses [Le Roux et al., 2013]. Furthermore, NMDAR-dependent plasticity is expressed postsynaptically, and seems to require presynaptic activation [Kullmann and Lamsa, 2007]. Other molecular signals that reflect excitatory activity have been implicated in the homeostatic regulation of synapses onto INs, including Narp and BDNF [Chang et al., 2010, Rutherford et al., 1998, Lamsa et al., 2007]. In summary, we conjecture that PV interneurons and their excitatory inputs have the necessary prerequisites to implement the suggested local Pyr $\rightarrow$ PV plasticity rule.

If excitatory inputs to Pyr neurons are much larger than required to reach the target, the homeostatic objective of bringing net currents to that target effectively requires a balance of excitation and inhibition on a stimulus-by-stimulus basis, with a small overshoot of excitation (or, in spiking networks, membrane potential fluctuations) that allows Pyr neurons to fire at the target rate. We speculate that E/I assemblies could be learned not only from the homeostatic objective used here, but by any other objective that enforces a positive correlation of the stimulus tuning of excitatory and inhibitory inputs to neurons in the circuit.

We expect that the rules we suggest here are only one set of many that can establish E/I assemblies. Given that the role of the input plasticity in the interneurons is the formation of a stimulus specificity, it is tempting to assume that this could equally well be achieved by classical forms of plasticity like the

Bienenstock-Cooper-Munro (BCM) rule [Bienenstock et al., 1982], which is commonly used in models of receptive field formation. However, in our hands, the combination of BCM plasticity in Pyr → PV synapses with homeostatic inhibitory plasticity in the PV → Pyr synapses showed complex dynamics, an analysis of which is beyond the scope of this article. In particular, this combination of rules often did not converge to a steady state, probably for the following reason. BCM rules tend to make the postsynaptic neuron as stimulus-selective as possible. Given the limited number of interneurons in our circuit, this can lead to a situation in which parts of stimulus space are not represented by any interneurons. As a result, Pyr neurons that respond to those stimuli cannot recruit inhibition and maintain a high firing rate far above the target. Other Pyr cells, which have access to interneurons with a similar stimulus tuning, can recruit inhibition to gradually reduce their firing rates towards the target rate. Because the BCM rule is Hebbian, it tends to strengthen input synapses from Pyr neurons with high activity. This shifts the stimulus tuning of the interneurons to those stimuli that were previously underrepresented. However, this in turn renders a different set of stimuli uncovered by inhibition and withdraws feedback inhibition from the corresponding set of Pyr cells, which can now fire at high rates.

We suspect that this instability can also arise for other Hebbian forms of plasticity in interneuron input synapses when they are combined with homeostatic inhibitory plasticity [Vogels et al., 2011] in their output synapses. The underlying reason is that for convergence, the two forms of plasticity need to work synergistically towards the same goal, i.e., the same steady state. For two arbitrary synaptic plasticity rules acting in different sets of synapses, it is likely that they aim for two different overall network configurations. Such competition can easily result in latching dynamics with a continuing turn-over of transiently stable states, in which the form of plasticity that acts more quickly gets to reach its goal transiently, only to be undermined by the other one later.

Both Pyr → PV and PV → Pyr plasticity have been studied in slice [for reviews, see, e.g., Kullmann and Lamsa, 2007, Vogels et al., 2013], but mostly in isolation. The idea that the two forms of plasticity should act in synergy suggests that it may be interesting to study both forms in the same system, e.g., in reciprocally connected Pyr-PV pairs.

Like all computational models, the present one contains simplifying design choices. First, we did not include stimulus-specific *feedforward* inhibition, because the focus lay on the formation of stimulus-specific *feedback* inhibition. The model could be enriched by feedforward inhibition in different ways. In particular, we expect that the two forms of plasticity will establish E/I assemblies even in the presence of stimulus-selective external inputs to the interneurons, because stimulus-specific external excitation should always be more supportive of the homeostatic objective than unspecific inputs. It may be

worth exploring whether adding feedforward inhibition leaves more room for replacing the PV input plasticity that we used by classical Hebbian rules, because the activity of the external inputs remains unaltered by the plasticity in the network (such that the complex instability described above may be mitigated). Given that the focus of this work was on feedback inhibition, an extensive evaluation of the different variants of feedforward inhibition are beyond the scope of the present article.

Second, we neglected much of the complexity of cortical interneuron circuits by including only one class of interneurons. We interpret these interneurons as PV-expressing interneurons, given that PV interneurons provide local feedback inhibition [Hu et al., 2014] and show a stimulus-selective circuitry akin to E/I assemblies [Znamenskiy et al., 2018]. With their peri-somatic targets on Pyr cells, PV-expressing (basket) cells are also a prime candidate for the classical feedback model of E/I balance [van Vreeswijk and Sompolinsky, 1996]. Note that our results do not hinge on any assumptions that are specific to PV neurons, and may thus also hold for other interneuron classes that provide feedback inhibition [Tremblay et al., 2016]. Given that the division of labour of the various cortical interneuron classes is far from understood, an extension to complex interneuron circuits [Litwin-Kumar et al., 2016, Hertäg and Sprekeler, 2019, 2020] is clearly beyond the present study.

Similarly tuned pyramidal cells tend to be recurrently connected [Cossell et al., 2015, Harris and Mrsic-Flogel, 2013], in line with the notion that excitatory cells with similar tuning mutually excite each other. This notion is questioned by a recent perturbation experiment demonstrating feature-specific suppression between pyramidal cells with similar tuning [Chettih and Harvey, 2019]. It has been suggested that this apparently paradoxical effect requires strong and tuned connections between excitatory and inhibitory neurons [Sadeh and Clopath, 2020]. The E/I assemblies that develop in our model provide sufficiently strong and specific inhibitory feedback to cause a suppression between similarly tuned Pyr neurons in response to perturbations. Hence, despite the presence of stimulus-specific excitatory recurrence, Pyr neurons with similar stimulus preference effectively compete. Computational arguments suggest that this feature competition may be beneficial for stimulus processing, e.g. by generating a sparser and more efficient representation of the stimuli [Olshausen and Field, 2004, Denève and Machens, 2016].

In addition to predicting that knocking out plasticity of inhibitory input or output synapses should prevent the development of E/I assemblies, our model also predicts different outcomes for single neuron perturbation experiments in juvenile and adult mice. Given that in rodents, stimulus-tuning of inhibitory currents occurs later in development than that of excitation [Dorrn et al., 2010], we expect that in juvenile mice single-cell perturbations would not cause feature-specific suppression but amplification due to excitatory recurrence and unspecific feedback inhibition.

Materials & Methods

Network & stimuli

We use custom software to simulate a rate-based recurrent network model containing $N^{\mathrm{E}} = 512$ excitatory and $N^{\mathrm{I}} = 64$ inhibitory neurons. The activation of the neurons follows Wilson-Cowan dynamics:

$$\tau_{\mathrm{E}} \frac{d}{dt} \boldsymbol{h}^{\mathrm{E}} = -\boldsymbol{h}^{\mathrm{E}} + W^{\mathrm{E}\leftarrow\mathrm{E}} \mathbf{r}^{E} - W^{\mathrm{E}\leftarrow\mathrm{I}} \boldsymbol{r}^{\mathrm{I}} + I^{\mathrm{bg}} + \mathbf{I}(\mathbf{s}) \tag{1a}$$

$$\tau_{\mathrm{I}} \frac{d}{dt} \boldsymbol{h}^{\mathrm{I}} = -\boldsymbol{h}^{\mathrm{I}} + W^{\mathrm{I}\leftarrow\mathrm{E}} \boldsymbol{r}^{\mathrm{E}} - W^{\mathrm{I}\leftarrow\mathrm{I}} \boldsymbol{r}^{\mathrm{I}} + I^{\mathrm{bg}} \,. \tag{1b}$$

Here, $\boldsymbol{r}^{\mathrm{E}} = [\boldsymbol{h}^{\mathrm{E}}]_+$, $\boldsymbol{r}^{\mathrm{I}} = [\boldsymbol{h}^{\mathrm{I}}]_+$ denote the firing rates of the excitatory and inhibitory neurons, which are given by their rectified activation. $W^{Y\leftarrow X}$ denotes the matrix of synaptic efficacies from population X to population Y ($X, Y \in \{\mathrm{E}, \mathrm{I}\}$). The external inputs $\mathbf{I}(\mathbf{s})$ to the excitatory neurons have a bell-shaped tuning in the three-dimensional stimulus space consisting of spatial frequency, temporal frequency and orientation [Znamenskiy et al., 2018]. To avoid edge effects, the stimulus space is periodic in all three dimensions, with stimuli ranging from -π to π. The stimulus tuning of the external inputs is modeled by a von Mises function with a maximum of 50 Hz and a tuning width $\kappa = 1$. The preferred stimuli of the $N^{\mathrm{E}} = 512$ excitatory cells cover the stimulus space evenly on a $12 \times 12 \times 12$ grid. All neurons receive a constant background input of $I^{\mathrm{bg}} = 5$ Hz.

Recurrent connections $W^{\mathrm{E}\leftarrow\mathrm{E}}$ among excitatory neurons have synaptic weight between neurons i and j that grows linearly with the signal correlation of their external inputs:

$$W_{ij}^{\mathrm{E}\leftarrow\mathrm{E}} = [corr(I_i(\mathbf{s}), I_j(\mathbf{s})) - C]_+ \,. \tag{2}$$

The cropping threshold C is chosen such that the overall connection among the excitatory neurons probability is 0.6. The remaining synaptic connections (E→I, I→E, I→I) are initially random, with a connection probability $p = 0.6$ and log-normal weights. For parameters please refer to Table 1.

During learning, we repeatedly draw all $12 \times 12 \times 12$ preferred stimuli of the Pyr neurons, in random order. This procedure is repeated 500 times to ensure convergence of synaptic weights. To reduce simulation time, we present each stimulus long enough for all firing rates to reach steady state and only then update the synaptic weights.

Synaptic plasticity

The PV → Pyr and Pyr → PV synapses follow plasticity rules that aim to minimize the deviation of the excitatory activations from a target rate ρ_0 ($\rho_0 = 1$ Hz):

$$\mathcal{E}\left(\boldsymbol{h}^{\mathrm{E}}\right) = \left\langle \frac{1}{2} \sum_{j=1}^{N^{\mathrm{E}}} \left(h_j^{\mathrm{E}} - \rho_0\right)^2 \right\rangle_{\mathbf{s}}, \tag{3}$$

where $\langle\cdot\rangle_{\mathbf{s}}$ denotes the average over all stimuli. When plastic, synaptic weights change according to

$$\Delta W_{ji}^{\mathrm{E}\leftarrow\mathrm{I}} \propto \left(h_j^{\mathrm{E}} - \rho_0\right) r_i^{\mathrm{I}}, \tag{4a}$$

$$\Delta W_{ij}^{\mathrm{I}\leftarrow\mathrm{E}} \propto \left[\sum_{k=1}^{N^{\mathrm{E}}} W_{ik}^{\mathrm{I}\leftarrow\mathrm{E}} \left(h_k^{\mathrm{E}} - \rho_0\right)\right] r_j^{\mathrm{E}} \tag{4b}$$

$$\approx \left[\sum_{k=1}^{N^{\mathrm{E}}} W_{ik}^{\mathrm{I}\leftarrow\mathrm{E}} \left(r_k^{\mathrm{E}} - \rho_0\right)\right] r_j^{\mathrm{E}}$$

$$= \left(I_i^{\mathrm{E,rec}} - I_0\right) r_j^{\mathrm{E}}. \tag{4c}$$

After every update of the Pyr → PV matrix, the incoming weights for each PV interneuron are multiplicatively scaled such that their sum is $J^{\mathrm{I}\leftarrow\mathrm{E}}$ [Akrout et al., 2019]. In that case, the rule in Eq. (4b) is approximately local in that it compares the excitatory input current $I_i^{E,\mathrm{rec}}$ received by the postsynaptic PV neuron to a target value $I_0 = J^{\mathrm{I}\leftarrow\mathrm{E}}\rho_0$, and adjusts the incoming synapses in proportion to this error and to presynaptic activity [see Eq. (4c)].

Both plasticity rules are approximations of the gradient of the objective function Eq. (3). Interested readers are referred to the supplementary methods for their mathematical derivation. For the results in Supp. Fig. S4, we use the Adaptive Moment Estimation (Adam) algorithm [Kingma and Ba, 2014] to improve optimisation performance.

We used a standard reparameterization method to ensure the sign constraints of an E/I network. Moreover, all weights are subject to a small weight-dependent decay term, which aids to keep the firing rates of the interneurons in a reasonable range. For details, please refer to the Supplementary Methods. The learning rule Eq. (4a) for the output synapses of the inhibitory neurons is similar to the rule proposed by Vogels et al. [2011], wherein each inhibitory synapse increases in strength if the deviation of the postsynaptic excitatory cell from the homeostatic target ρ_0 is positive (and decreases it when negative). In contrast, the learning rule Eq. (4b) increases activated input synapses for an interneuron if the weighted sum of deviations in its presynaptic excitatory population is positive (and

decreases them if it is negative). Though it is local, when operating in conjunction with the plasticity of Eq. (4a), this leads to feedback alignment in our simulations, and effectively performs backpropagation without the need for weight transport [Akrout et al., 2019].

Note that the objective function Eq. (3) can also be interpreted differently. The activation h^{E} of a neuron is essentially the difference between its excitatory and inhibitory inputs. Therefore, the objective function Eq. (3) is effectively the mean squared error between excitation and inhibition, aside from a small constant offset ρ_0. The derived learning rules can therefore be seen as supervised learning of the inhibitory inputs, with excitation as the label. They hence aim to establish the best co-tuning of excitation and inhibition that is possible given the circuitry.

Perturbation experiments

The perturbation experiments in Fig. 4 are performed in a network in which both forms of plasticity have converged. The network is then exposed to different stimuli, while the afferent drive to a single excitatory cell i is transiently increased by $\Delta I = 10$ Hz. For each stimulus, we compute the steady state firing rates r_j of all excitatory cells both with and without the perturbation. The influence of the perturbation of neuron i on neuron j is defined as the difference between these two firing rates, normalized by the pertubation magnitude [Sadeh and Clopath, 2020]. This stimulation protocol is repeated for 90 randomly selected excitatory neurons. The dependence of the influence on the tuning similarity (Fig. 4d) is obtained by binning the influence of the perturbed neuron i and the influenced neuron j according to their stimulus response correlation, and then averaging across all influences in the bin. During the perturbation experiments, synaptic plasticity was disabled.

Quantitative measures

The response similarity (RS) of the stimulus tuning of two neurons i and j is measured by the dot product of their steady state firing rates in response to all stimuli, normalized by the product of their norms [Znamenskiy et al., 2018]:

$$\mathrm{RS}(r_i, r_j) = \frac{\sum_{\mathbf{s}} r_i(\mathbf{s}) r_j(\mathbf{s})}{\left(\sum_{\mathbf{s}} \left(r_i(\mathbf{s})\right)^2 \sum_{\mathbf{s}} \left(r_j(\mathbf{s})\right)^2\right)^{1/2}}. \tag{5}$$

The same measure is used for the similarity of synaptic currents onto excitatory neurons in Supp. Fig. S2c & S4d.

There is no structural plasticity, i.e. synapses are never added or pruned. However, when calculating

N^{E}	512	N^{I}	64	Number of exc. & inh. neurons.
τ_{E}	50 ms	τ_{I}	25 ms	Rate dynamics time constants.
dt	1 ms			Numerical integration time step.
$p^{\mathrm{E}\leftarrow X}$	0.6	$p^{\mathrm{I}\leftarrow X}$	0.6	Connection probability to exc. & inh. neurons.
$J_i^{\mathrm{E}\leftarrow\mathrm{E}}$	2	$J_i^{\mathrm{I}\leftarrow\mathrm{E}}$	5	Total of exc. weights onto neuron i: $\sum_j W_{ij}^{X\leftarrow\mathrm{E}}$
$J_i^{\mathrm{E}\leftarrow\mathrm{I}}$	1	$J_i^{\mathrm{I}\leftarrow\mathrm{I}}$	1	Total of inh. weights onto neuron i: $\sum_j W_{ij}^{X\leftarrow\mathrm{I}}$
$\sigma^{\mathrm{E}\leftarrow X}$	0.65	$\sigma^{\mathrm{I}\leftarrow X}$	0.65	Std. deviation of the logarithm of the weights.
$\theta^{\mathrm{E}\leftarrow\mathrm{I}}$	10^{-4}	$\theta^{\mathrm{I}\leftarrow\mathrm{E}}$	10^{-4}	Experimental detection threshold for synapses.
I^{bg}	5 Hz	$\max(\mathbf{I}(\mathbf{s}))$	50 Hz	Background & maximum stimulus-specific input.
N^S	$12\times12\times12$	N^{trials}	500	Number of stimuli & trials.
R^S	$2\pi\times2\pi\times2\pi$	κ	1	Range of stimuli & Pyr RF von Mises width.
ΔI	10 Hz			Change of input for perturbation experiments.
$\eta^{\mathrm{Approx.}}$	10^{-5}	$\eta^{\mathrm{Grad.}}$	10^{-3}	Learning rates (approx. & gradient rules).
$\delta^{\mathrm{E}\leftarrow\mathrm{I}}$	0.1	$\delta^{\mathrm{I}\leftarrow\mathrm{E}}$	0.1	Weight decay rates.
ρ_0	1 Hz			Homeostatic plasticity target.
β_1	0.9	β_2	0.999	Adam parameters for gradient rules.
ϵ	10^{-9}			

Table 1: **Model parameters.**

Pearson's correlation between synaptic weights and RS, we exclude synapses that are too weak to be detected using the experimental protocol employed by Znamenskiy et al. [2018]. The threshold values $\theta^{\mathrm{E}\leftarrow\mathrm{I}}$ & $\theta^{\mathrm{I}\leftarrow\mathrm{E}}$ were chosen to be approximately four orders of magnitude weaker than the strongest synapses in the network. The rules that we investigate here tend to produce bimodal distributions of weights, with the lower mode well below this threshold (Supp. Fig. S7).

The stimulus selectivity of the neurons is measured by the skewness of their response distribution across all stimuli:

$$\gamma_i = \frac{\left\langle \left(r_i(\mathbf{s}) - \bar{r}_i\right)^3 \right\rangle_{\mathbf{s}}}{\left\langle \left(r_i(\mathbf{s}) - \bar{r}_i\right)^2 \right\rangle_{\mathbf{s}}^{3/2}} \tag{6}$$

where $\bar{r}_i = \langle r_i(\mathbf{s}) \rangle_{\mathbf{s}}$. Both the response similarity Eq. (5) and the stimulus selectivity Eq. (6) are adapted from Znamenskiy et al. [2018].

Finally, the angle θ between the gradient G from Eq. (15) and its approximation A from Eq. (4) is given by:

$$\theta = \arccos\left(\frac{\sum_{ij} G_{ij} A_{ij}}{\left(\sum_{ij} G_{ij}^2 \sum_{ij} A_{ij}^2\right)^{1/2}}\right) \tag{7}$$

References

H. Adesnik, W. Bruns, H. Taniguchi, Z. J. Huang, and M. Scanziani. A neural circuit for spatial summation in visual cortex. *Nature*, 490(7419):226–231, 2012.

M. Akrout, C. Wilson, P. Humphreys, T. Lillicrap, and D. B. Tweed. Deep learning without weight transport. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 976–984. Curran Associates, Inc., 2019.

H. Barron, T. Vogels, U. Emir, T. Makin, J. O'shea, S. Clare, S. Jbabdi, R. J. Dolan, and T. Behrens. Unmasking latent inhibitory connections in human cortex to reveal dormant cortical memories. *Neuron*, 90(1):191–203, 2016.

H. C. Barron, T. P. Vogels, T. E. Behrens, and M. Ramaswami. Inhibitory engrams in perception and memory. *Proceedings of the National Academy of Sciences*, 114(26):6666–6674, 2017.

G. Bellec, F. Scherr, A. Subramoney, E. Hajek, D. Salaj, R. Legenstein, and W. Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *bioRxiv*, page 738385, 2020.

A. Bhatia, S. Moza, and U. S. Bhalla. Precise excitation-inhibition balance controls gain and timing in the hippocampus. *eLife*, 8:e43415, 2019.

E. Bienenstock, L. Cooper, and P. Munroe. Theory of the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2:32–48, 1982.

D. D. Bock, W.-C. A. Lee, A. M. Kerlin, M. L. Andermann, G. Hood, A. W. Wetzel, S. Yurgenson, E. R. Soucy, H. S. Kim, and R. C. Reid. Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, 471(7337):177–182, 2011.

N. Brunel. Dynamics of sparsely connected networks of excitatory and inhibitory neurons. *Journal of Computational Neuroscience*, 8(3):183–208, 2000.

M. C. Chang, J. M. Park, K. A. Pelkey, H. L. Grabenstatter, D. Xu, D. J. Linden, T. P. Sutula, C. J. McBain, and P. F. Worley. Narp regulates homeostatic scaling of excitatory synapses on parvalbumin-expressing interneurons. *Nature neuroscience*, 13(9):1090–1097, 2010.

N. Chenkov, H. Sprekeler, and R. Kempter. Memory replay in balanced recurrent networks. *PLoS Computational Biology*, 13(1):e1005359, 2017.

S. N. Chettih and C. D. Harvey. Single-neuron perturbations reveal feature-specific competition in V1. *Nature*, 567(7748):334–340, 2019.

C. Clopath, T. P. Vogels, R. C. Froemke, and H. Sprekeler. Receptive field formation by interacting excitatory and inhibitory synaptic plasticity. *bioRxiv*, page 066589, 2016.

L. Cossell, M. F. Iacaruso, D. R. Muir, R. Houlton, E. N. Sader, H. Ko, S. B. Hofer, and T. D. Mrsic-Flogel. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature*, 518(7539):399–403, 2015.

S. Denève and C. K. Machens. Efficient codes and balanced networks. *Nature Neuroscience*, 19(3): 375, 2016.

M. Dipoppa, A. Ranson, M. Krumin, M. Pachitariu, M. Carandini, and K. D. Harris. Vision and locomotion shape the interactions between neuron types in mouse visual cortex. *Neuron*, 98(3): 602–615, 2018.

A. L. Dorrn, K. Yuan, A. J. Barker, C. E. Schreiner, and R. C. Froemke. Developmental sensory experience balances cortical excitation and inhibition. *Nature*, 465(7300):932–936, 2010.

E. Fino and R. Yuste. Dense inhibitory connectivity in neocortex. *Neuron*, 69(6):1188–1203, 2011.

G. Fishell and A. Kepecs. Interneuron types as attractors and controllers. *Annual Review of Neuroscience*, 43, 2019.

R. C. Froemke, M. M. Merzenich, and C. E. Schreiner. A synaptic memory trace for cortical receptive field plasticity. *Nature*, 450:425–429, 2007.

C. Grienberger, X. Chen, and A. Konnerth. Dendritic function in vivo. *Trends in neurosciences*, 38 (1):45–54, 2015.

J. Guerguiev, T. P. Lillicrap, and B. A. Richards. Towards deep learning with segregated dendrites. *eLife*, 6:e22901, 2017.

K. D. Harris and T. D. Mrsic-Flogel. Cortical connectivity and sensory coding. *Nature*, 503(7474): 51–58, 2013.

G. Hennequin, T. P. Vogels, and W. Gerstner. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, 82(6):1394–1406, 2014.

G. Hennequin, E. J. Agnes, and T. P. Vogels. Inhibitory plasticity: Balance, control, and codependence. *Annual Review of Neuroscience*, 40:557–579, 2017.

L. Hertäg and H. Sprekeler. Amplifying the redistribution of somato-dendritic inhibition by the interplay of three interneuron types. *PLoS Computational Biology*, 15(5):e1006999, 2019.

L. Hertäg and H. Sprekeler. Learning prediction error neurons in a canonical interneuron circuit. *bioRxiv*, 2020.

S. B. Hofer, H. Ko, B. Pichler, J. Vogelstein, H. Ros, H. Zeng, E. Lein, N. A. Lesica, and T. D. Mrsic-Flogel. Differential connectivity and response dynamics of excitatory and inhibitory neurons in visual cortex. *Nature Neuroscience*, 14(8):1045, 2011.

H. Hu, J. Gan, and P. Jonas. Fast-spiking, parvalbumin+ GABAergic interneurons: From cellular design to microcircuit function. *Science*, 345(6196):1255263, 2014.

J. S. Isaacson and M. Scanziani. How inhibition shapes cortical activity. *Neuron*, 72(2):231–243, 2011.

H. Jia, N. L. Rochefort, X. Chen, and A. Konnerth. Dendritic organization of sensory input to cortical neurons in vivo. *Nature*, 464(7293):1307–1312, 2010.

M. M. Karnani, J. Jackson, I. Ayzenshtat, A. H. Sichani, K. Manoocheri, S. Kim, and R. Yuste. Opening holes in the blanket of inhibition: Localized lateral disinhibition by VIP interneurons. *Journal of Neuroscience*, 36(12):3471–3480, 2016.

T. Keck, V. Scheuss, R. I. Jacobsen, C. J. Wierenga, U. T. Eysel, T. Bonhoeffer, and M. Hübener. Loss of sensory input causes rapid structural changes of inhibitory neurons in adult mouse visual cortex. *Neuron*, 71(5):869–882, 2011.

A. G. Khan, J. Poort, A. Chadwick, A. Blot, M. Sahani, T. D. Mrsic-Flogel, and S. B. Hofer. Distinct learning-induced changes in stimulus selectivity and interactions of GABAergic interneuron classes in visual cortex. *Nature Neuroscience*, 21(6):851–859, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, and T. D. Mrsic-Flogel. Functional specificity of local synaptic connections in neocortical networks. *Nature*, 473(7345):87–91, 2011.

S. J. Kuhlman, N. D. Olivas, E. Tring, T. Ikrar, X. Xu, and J. T. Trachtenberg. A disinhibitory microcircuit initiates critical-period plasticity in the visual cortex. *Nature*, 501(7468):543–546, 2013.

D. M. Kullmann and K. P. Lamsa. Long-term synaptic plasticity in hippocampal interneurons. *Nature Reviews Neuroscience*, 8(9):687–699, 2007.

K. Lamsa, E. E. Irvine, K. P. Giese, and D. M. Kullmann. Nmda receptor-dependent long-term potentiation in mouse hippocampal interneurons shows a unique dependence on ca2+/calmodulin-dependent kinases. *The Journal of physiology*, 584(3):885–894, 2007.

I. D. Landau, R. Egger, V. J. Dercksen, M. Oberlaender, and H. Sompolinsky. The impact of structural heterogeneity on excitation-inhibition balance in cortical networks. *Neuron*, 92(5):1106–1121, 2016.

M. E. Larkum and T. Nevian. Synaptic clustering by dendritic signalling mechanisms. *Current opinion in neurobiology*, 18(3):321–331, 2008.

N. Le Roux, C. Cabezas, U. L. Böhm, and J. C. Poncer. Input-specific learning rules at excitatory synapses onto hippocampal parvalbumin-expressing interneurons. *The Journal of physiology*, 591 (7):1809–1822, 2013.

J. Letzkus, S. Wolff, E. Meyer, P. Tovote, J. Courtin, C. Herry, and A. Lüthi. A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature*, 480:331–335, December 2011.

T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7(1):1–10, 2016.

T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, pages 1–12, 2020.

A. Litwin-Kumar and B. Doiron. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nature Neuroscience*, 2012.

A. Litwin-Kumar and B. Doiron. Formation and maintenance of neuronal assemblies through synaptic plasticity. *Nature Communications*, 5(1):1–12, 2014.

A. Litwin-Kumar, R. Rosenbaum, and B. Doiron. Inhibitory stabilization and visual coding in cortical circuits with multiple interneuron subtypes. *Journal of Neurophysiology*, 115(3):1399–1409, 2016.

A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, and L. Abbott. Optimal degrees of synaptic connectivity. *Neuron*, 93(5):1153–1164, 2017.

T. Lochmann, U. A. Ernst, and S. Deneve. Perceptual inference predicts contextual modulations of sensory responses. *Journal of Neuroscience*, 32(12):4179–4195, 2012.

Y. Loewenstein, A. Kuras, and S. Rumpel. Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex in vivo. *Journal of Neuroscience*, 31(26): 9481–9488, 2011.

R. Moreno-Bote and J. Drugowitsch. Causal inference and explaining away in a spiking network. *Scientific Reports*, 5:17531, 2015.

B. Murphy and K. Miller. Balanced amplification: A new mechanism of selective amplification of neural activity patterns. *Neuron*, 61(4):635–648, 2009. ISSN 0896-6273.

F. Najafi, G. F. Elsayed, R. Cao, E. Pnevmatikakis, P. E. Latham, J. P. Cunningham, and A. K. Churchland. Excitatory and inhibitory subnetworks are equally selective during decision-making and emerge simultaneously during learning. *Neuron*, 105(1):165–179, 2020.

E. O. Neftci, H. Mostafa, and F. Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

K. Ohki, S. Chung, Y. H. Ch'ng, P. Kara, and R. C. Reid. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature*, 433(7026):597–603, 2005.

M. Okun and I. Lampl. Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nature Neuroscience*, 11(5):535–537, 2008.

B. A. Olshausen and D. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14 (4):481–487, 2004.

A. M. Packer and R. Yuste. Dense, unspecific connectivity of neocortical parvalbumin-positive interneurons: A canonical microcircuit for inhibition? *Journal of Neuroscience*, 31(37):13260–13271, 2011.

E. A. Phillips, C. E. Schreiner, and A. R. Hasenstaub. Cortical interneurons differentially regulate the effects of acoustic context. *Cell Reports*, 20(4):771–778, 2017.

F. Pouille and M. Scanziani. Enforcement of temporal fidelity in pyramidal cells by somatic feed-forward inhibition. *Science*, 293(5532):1159–1163, 2001.

N. J. Priebe and D. Ferster. Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron*, 57(4):482–497, 2008.

D. B. Rubin, S. D. Van Hooser, and K. D. Miller. The stabilized supralinear network: A unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85(2):402–417, 2015.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

P. Rupprecht and R. W. Friedrich. Precise synaptic balance in the zebrafish homolog of olfactory cortex. *Neuron*, 100(3):669–683, 2018.

L. C. Rutherford, S. B. Nelson, and G. G. Turrigiano. Bdnf has opposite effects on the quantal amplitude of pyramidal neuron and interneuron excitatory synapses. *Neuron*, 21(3):521–530, 1998.

J. Sacramento, R. P. Costa, Y. Bengio, and W. Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. In *Advances in Neural Information Processing Systems*, pages 8721–8732, 2018.

S. Sadeh and C. Clopath. Theory of neuronal perturbome: Linking connectivity to coding via perturbations. *bioRxiv*, 2020.

S. Song, P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biology*, 3(3), 2005.

H. Sprekeler. Functional consequences of inhibitory plasticity: Homeostasis, the excitation-inhibition balance and beyond. *Current Opinion in Neurobiology*, 43:198–203, 2017.

A. E. Takesian, V. C. Kotak, N. Sharma, and D. H. Sanes. Hearing loss differentially affects thalamic drive to two cortical interneuron subtypes. *Journal of Neurophysiology*, 110(4):999–1008, 2013.

A. Y. Tan, B. D. Brown, B. Scholl, D. Mohanty, and N. J. Priebe. Orientation selectivity of synaptic input to neurons in mouse and cat primary visual cortex. *Journal of Neuroscience*, 31(34):12339–12350, 2011.

P. Tovote, J. P. Fadok, and A. Lüthi. Neuronal circuits for fear and anxiety. *Nature Reviews Neuroscience*, 16(6):317–331, 2015.

R. Tremblay, S. Lee, and B. Rudy. GABAergic interneurons in the neocortex: From cellular properties to circuits. *Neuron*, 91(2):260–292, 2016.

C. van Vreeswijk and H. Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274:1724–1726, 1996.

T. Vogels and L. Abbott. Gating multiple signals through detailed balance of excitation and inhibition in spiking networks. *Nature Neuroscience*, 12(4):483–491, 2009. ISSN 1097-6256.

T. Vogels, H. Sprekeler, F. Zenke, C. Clopath, and W. Gerstner. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science*, 334(6062):1569–1573, 2011.

T. P. Vogels, R. C. Froemke, N. Doyon, M. Gilson, J. S. Haas, R. Liu, A. Maffei, P. Miller, C. Wierenga, M. A. Woodin, et al. Inhibitory synaptic plasticity: spike timing-dependence and putative network function. *Frontiers in neural circuits*, 7:119, 2013.

S. N. Weber and H. Sprekeler. Learning place cells, grid cells and invariances with excitatory and inhibitory plasticity. *eLife*, 7:e34560, 2018.

M. Wehr and A. Zador. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature*, 426:442–446, 2003.

P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

J. C. Whittington and R. Bogacz. Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, 2019.

M. Xue, B. V. Atallah, and M. Scanziani. Equalizing excitation–inhibition ratios across visual cortical neurons. *Nature*, 511(7511):596–600, 2014.

Y. Yoshimura, J. L. Dantzker, and E. M. Callaway. Excitatory cortical neurons form fine-scale functional networks. *Nature*, 433(7028):868–873, 2005.

X.-M. Yu and M. W. Salter. Gain control of nmda-receptor currents by intracellular sodium. *Nature*, 396(6710):469–474, 1998.

P. Znamenskiy, M.-H. Kim, D. R. Muir, M. F. Iacaruso, S. B. Hofer, and T. D. Mrsic-Flogel. Functional selectivity and specific connectivity of inhibitory neurons in primary visual cortex. *bioRxiv*, page 294835, 2018.

Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to H.S.

Acknowledgements

We thank Joram Keijser for helpful discussions that inspired parts of this work. He, along with Denis Alevi, Loreen Hertäg and Robert T. Lange also provided careful proof-reading of the manuscript. This project was funded by the German Federal Ministry for Science and Education through a Bernstein Award (BMBF, FKZ 01GQ1201) and by the German Research Foundation (DFG, collaborative research center FOR 2143).

Contributions

O.M. & H.S. conceived the model. O.M. wrote the simulator, and performed all of the simulations. H.S. supervised the project, and acquired the funding. All authors contributed to the experimental design, interpretation of results, and writing of the manuscript.

Competing Interests

The authors declare no competing interests.

Supplementary Materials

Plasticity rules

The general framework we follow to derive homeostatic rules is to minimise the mean squared deviation of individual excitatory (Pyr) neuron activations from a target for all stimuli. More specifically, we perform gradient descent on the following objective function:

$$\mathcal{E}\left(\boldsymbol{h}^{\mathrm{E}}\right) = \left\langle \frac{1}{2} \sum_{j=1}^{N^{\mathrm{E}}} \left(h_j^{\mathrm{E}} - \rho_0\right)^2 \right\rangle_{\mathbf{s}}.$$

Note that the activations $\boldsymbol{h}^{\mathrm{E}}$ are given by the difference between the excitatory and the inhibitory inputs to the excitatory neurons. Our approach can hence be interpreted as supervised learning of the inhibitory circuitry, with the goal of minimising the mean squared loss between the inhibitory and the excitatory inputs (plus the constant target ρ_0). In this sense, the derived gradient rules aim to generate the best possible E/I balance across stimuli that is possible with the circuitry at hand.

For reasons of readability, we will first simply state the derived rules. The details of their derivation can be found in the following section.

The sign constraints in excitatory-inhibitory networks require all synaptic weights to remain positive. To ensure this, we reparameterised all plastic weights of the network by a strictly positive soft-plus function $W = s^+(V) = \alpha^{-1} \ln\left(1 + \exp \alpha V\right)$ and optimised the weight parameter V by gradient descent.

In summary, the derived learning rules for the synaptic weight parameters between excitatory neuron j and inhibitory interneuron i are given by

$$\Delta V_{ji}^{\mathrm{E}\leftarrow\mathrm{I}} = \eta^{\mathrm{I}} \left(h_j^{\mathrm{E}} - \rho_0\right) \frac{\partial W_{ji}^{\mathrm{E}\leftarrow\mathrm{I}}}{\partial V_{ji}^{\mathrm{E}\leftarrow\mathrm{I}}} r_i^{\mathrm{I}} - \delta^{\mathrm{I}} W_{ji}^{\mathrm{E}\leftarrow\mathrm{I}}, \tag{8a}$$

$$\Delta V_{ij}^{\mathrm{I}\leftarrow\mathrm{E}} = \eta^{\mathrm{E}} \left[\sum_{k=1}^{N^{\mathrm{E}}} W_{ik}^{\mathrm{I}\leftarrow\mathrm{E}} \left(h_k^{\mathrm{E}} - \rho_0\right)\right] \frac{\partial r_i^{\mathrm{I}}}{\partial h_i^{\mathrm{I}}} \frac{\partial W_{ij}^{\mathrm{I}\leftarrow\mathrm{E}}}{\partial V_{ij}^{\mathrm{I}\leftarrow\mathrm{E}}} r_j^{\mathrm{E}} - \delta^{\mathrm{E}} W_{ij}^{\mathrm{I}\leftarrow\mathrm{E}}. \tag{8b}$$

Please note that we added a small weight decay to both learning rules. The purpose of this decay term is to avoid an ambiguity in the solution. When the firing rates of the interneurons are increased, but their output weights are decreased accordingly, the firing rates of the excitatory population remain unchanged. Pure gradient-based rules can therefore generate extreme values for the synaptic weights, in which the interneurons have biologically unrealistic firing rates. The additional decay terms in the learning rules solve this issue.

Finally, we replaced the derivative $\frac{\partial r}{\partial h}$ (which should be a Heaviside function, because rates are the rectified activations) by the derivative of a soft-plus function with finite sharpness ($\alpha = 1$). This allows interneurons to recover from a silent state, in which all gradients vanish. Note that this replacement is done only in the learning rules. The firing rates are still the rectified activations. This method is similar to recent surrogate gradient approaches in spiking networks [Neftci et al., 2019].

Derivation of the homeostatic plasticity rules in recurrent networks

The challenging aspect of the derivation of the learning rules lies in the recurrence of the network. The effects of changes in individual synapses can percolate through the network and thereby change the firing rates of all neurons. Moreover, the temporal dynamics of the network would in principle require a backpropagation of the gradient through time. We circumvent this complication by assuming that the external stimuli to the network change slowly compared to the dynamical time scales of the network, and that the network adiabatically follows the fixed point in its dynamics as the stimulus changes. This assumption significantly simplifies the derivation of the gradient.

The goal is to minimise the total deviation of the excitatory activations $\boldsymbol{h}^{\mathrm{E}}$ from the homeostatic target value ρ_0. To this end, we calculate the gradient of the objective function in Eq. (3) with respect to a given synaptic weight parameter $v \in \{V_{ij}^{\mathrm{I}\leftarrow\mathrm{E}}, V_{ji}^{\mathrm{E}\leftarrow\mathrm{I}}\}$:

$$\frac{\partial}{\partial v}\mathcal{E}(\boldsymbol{h}^{\mathrm{E}}) = \left\langle \left(\boldsymbol{h}^{\mathrm{E}} - \rho_0\right)^T \frac{\partial \boldsymbol{h}^{\mathrm{E}}}{\partial v} \right\rangle_{\mathbf{s}} . \tag{9}$$

We therefore need the gradient of the activations $\boldsymbol{h}^{\mathrm{E}}$ of excitatory cells with respect to a parameter v. In the steady state, the activations are given by

$$\boldsymbol{h}^{\mathrm{E}} = W^{\mathrm{E}\leftarrow\mathrm{E}}\boldsymbol{r}^{\mathrm{E}} - W^{\mathrm{E}\leftarrow\mathrm{I}}\boldsymbol{r}^{\mathrm{I}} + I_{\mathrm{bg}} + \mathbf{I}(\mathbf{s}) . \tag{10}$$

The gradient of the activations $\boldsymbol{h}^{\mathrm{E}}$ is therefore given by the following implicit condition:

$$\frac{\partial \boldsymbol{h}^{\mathrm{E}}}{\partial v} = W^{\mathrm{E}\leftarrow\mathrm{E}} D^{\mathrm{E}} \frac{\partial \boldsymbol{h}^{\mathrm{E}}}{\partial v} - \left[\frac{\partial W^{\mathrm{E}\leftarrow\mathrm{I}}}{\partial v}\boldsymbol{r}^{\mathrm{I}} + W^{\mathrm{E}\leftarrow\mathrm{I}} D^{\mathrm{I}} \frac{\partial \boldsymbol{h}^{\mathrm{I}}}{\partial v} \right] , \tag{11}$$

where we introduced the diagonal matrices $D_{ij}^{\mathrm{E/I}} := \delta_{ij}\partial r_i^{\mathrm{E/I}}/\partial h_i^{\mathrm{E/I}}$ for notational convenience, δ_{ij} being the Kronecker symbol. Derivatives of expressions that do not depend on any of the synaptic weights in question are excluded.

Eq. (11) requires the gradient $\frac{\partial \boldsymbol{h}^{\mathrm{I}}}{\partial v}$ of the inhibitory activations with respect to the parameter v,

which can be calculated by a similar approach

$$\begin{aligned}\frac{\partial \boldsymbol{h}^{\mathrm{I}}}{\partial v} &= \frac{\partial}{\partial v}\left(W^{\mathrm{I}\leftarrow\mathrm{E}}\boldsymbol{r}^{\mathrm{E}} - W^{\mathrm{I}\leftarrow\mathrm{I}}\boldsymbol{r}^{\mathrm{I}} + I_{\mathrm{bg}}\right)\\ &= \left(\frac{\partial W^{\mathrm{I}\leftarrow\mathrm{E}}}{\partial v}\boldsymbol{r}^{\mathrm{E}} + W^{\mathrm{I}\leftarrow\mathrm{E}}D^{\mathrm{E}}\frac{\partial \boldsymbol{h}^{\mathrm{E}}}{\partial v}\right) - W^{\mathrm{I}\leftarrow\mathrm{I}}D^{\mathrm{I}}\frac{\partial \boldsymbol{h}^{\mathrm{I}}}{\partial v}\,.\end{aligned}$$

Introducing the effective interaction matrix $\mathcal{M} := \mathbb{I} + W^{\mathrm{I}\leftarrow\mathrm{I}}D^{\mathrm{I}}$ among the interneurons ($\mathbb{I}$ being the identity matrix) allows to solve for the gradient of $\boldsymbol{h}^{\mathrm{I}}$:

$$\frac{\partial \boldsymbol{h}^{\mathrm{I}}}{\partial v} = \mathcal{M}^{-1}\left[W^{\mathrm{I}\leftarrow\mathrm{E}}D^{\mathrm{E}}\frac{\partial \boldsymbol{h}^{\mathrm{E}}}{\partial v} + \frac{\partial W^{\mathrm{I}\leftarrow\mathrm{E}}}{\partial v}\boldsymbol{r}^{\mathrm{E}}\right]$$

Inserting this expression into Eq. (11) yields

$$\frac{\partial \boldsymbol{h}^{\mathrm{E}}}{\partial v} = \left[W^{\mathrm{E}\leftarrow\mathrm{E}}D^{\mathrm{E}} - W^{\mathrm{E}\leftarrow\mathrm{I}}D^{\mathrm{I}}\mathcal{M}^{-1}W^{\mathrm{I}\leftarrow\mathrm{E}}D^{\mathrm{E}}\right]\frac{\partial \boldsymbol{h}^{\mathrm{E}}}{\partial v} - \frac{\partial W^{\mathrm{E}\leftarrow\mathrm{I}}}{\partial v}\boldsymbol{r}^{\mathrm{I}} - W^{\mathrm{E}\leftarrow\mathrm{I}}D^{\mathrm{I}}\mathcal{M}^{-1}\frac{\partial W^{\mathrm{I}\leftarrow\mathrm{E}}}{\partial v}\boldsymbol{r}^{\mathrm{E}}\,,$$

Introducing the effective interaction matrix $\mathcal{W} = \mathbb{I} - W^{\mathrm{E}\leftarrow\mathrm{E}}D^{\mathrm{E}} + W^{\mathrm{E}\leftarrow\mathrm{I}}D^{\mathrm{I}}\mathcal{M}^{-1}W^{\mathrm{I}\leftarrow\mathrm{E}}D^{\mathrm{E}}$ among the excitatory neurons yields an explicit expression for the gradient of $\boldsymbol{h}^{\mathrm{E}}$:

$$\frac{\partial \boldsymbol{h}^{\mathrm{E}}}{\partial v} = -\mathcal{W}^{-1}\frac{\partial W^{\mathrm{E}\leftarrow\mathrm{I}}}{\partial v}\boldsymbol{r}^{\mathrm{I}} - \mathcal{W}^{-1}W^{\mathrm{E}\leftarrow\mathrm{I}}D^{\mathrm{I}}\mathcal{M}^{-1}\frac{\partial W^{\mathrm{I}\leftarrow\mathrm{E}}}{\partial v}\boldsymbol{r}^{\mathrm{E}}\,, \tag{12}$$

To obtain gradients with respect to a particular network parameter, we simply substitute the chosen parameter into Eq. (12). For the parameters $V_{ij}^{\mathrm{I}\leftarrow\mathrm{E}}$ of the input synapses to the interneurons, the gradient reduces to

$$\frac{\partial \boldsymbol{h}^{\mathrm{E}}}{\partial V^{\mathrm{I}\leftarrow\mathrm{E}}} = -\mathcal{W}^{-1}W^{\mathrm{E}\leftarrow\mathrm{I}}D^{\mathrm{I}}\mathcal{M}^{-1}\frac{\partial W^{\mathrm{I}\leftarrow\mathrm{E}}}{\partial V^{\mathrm{I}\leftarrow\mathrm{E}}}\boldsymbol{r}^{\mathrm{E}}\,, \tag{13}$$

and for the parameters $V_{ij}^{\mathrm{E}\leftarrow\mathrm{I}}$ of the output synapses from the interneurons we get

$$\frac{\partial \boldsymbol{h}^{\mathrm{E}}}{\partial V^{\mathrm{I}\leftarrow\mathrm{E}}} = -\mathcal{W}^{-1}\frac{\partial W^{\mathrm{E}\leftarrow\mathrm{I}}}{\partial V^{\mathrm{E}\leftarrow\mathrm{I}}}\boldsymbol{r}^{\mathrm{I}}\,. \tag{14}$$

By inserting these expressions into Eq. (9) and dropping the average, we obtain online learning rules for the input and output synapses of the interneurons:

$$\Delta V^{\mathrm{I}\leftarrow\mathrm{E}} \propto \left[(\boldsymbol{h}^{\mathrm{E}} - \rho_0)^{\intercal}\mathcal{W}^{-1}W^{\mathrm{E}\leftarrow\mathrm{I}}D^{\mathrm{I}}\mathcal{M}^{-1}\right]\frac{\partial W^{\mathrm{I}\leftarrow\mathrm{E}}}{\partial V^{\mathrm{I}\leftarrow\mathrm{E}}}\boldsymbol{r}^{\mathrm{E}} \tag{15a}$$

$$\Delta V^{\mathrm{E}\leftarrow\mathrm{I}} \propto \left[(\boldsymbol{h}^{\mathrm{E}} - \rho_0)^{\intercal}\mathcal{W}^{-1}\right]\frac{\partial W^{\mathrm{E}\leftarrow\mathrm{I}}}{\partial V^{\mathrm{E}\leftarrow\mathrm{I}}}\boldsymbol{r}^{\mathrm{I}}\,. \tag{15b}$$

Note that the same approach also yields learning rules for the threshold and the gain of the transfer

function of the inhibitory interneurons, if those are parameters of the system. Although we did not use such intrinsic plasticity rules, we include them here for the interested reader. We assumed a threshold linear transfer function of the interneurons: $r_i^{\mathrm{I}} = g_i \left[h_i^{\mathrm{I}} - \theta_i\right]^+$, where g_i is the gain of the neuronal transfer function and θ_i a firing threshold. While the firing threshold can become negative, gain is reparameterised via the strictly positive soft-plus $g_i = s^+(v_i^g)$.

The gradient-based learning rule for the firing thresholds θ_i of the interneurons is given by

$$\Delta\theta_i \propto -\left[\left(\boldsymbol{h}^{\mathrm{E}} - \rho_0\right)^\intercal \mathcal{W}^{-1} W^{\mathrm{E}\leftarrow\mathrm{I}} \mathcal{M}^{-1}\right]_i \frac{\partial r_i^{\mathrm{I}}}{\partial \theta_i}\,, \tag{16}$$

and the corresponding learning rule for the interneuron gain g_i is

$$\Delta v_i^g \propto \left[\left(\boldsymbol{h}^{\mathrm{E}} - \rho_0\right)^\intercal \mathcal{W}^{-1} W^{\mathrm{E}\leftarrow\mathrm{I}} \mathcal{M}^{-1}\right]_i \frac{\partial r_i^{\mathrm{I}}}{\partial g_i} \frac{\partial g_i}{\partial v_i^g}\,. \tag{17}$$

Approximating the gradient rules

In the gradient-based rules derived in the previous section, the $\mathcal{W}^{-1}$ and $\mathcal{M}^{-1}$ terms account for the fact that a change in a given synaptic connections percolates through the network. As a result, the learning rules are highly nonlocal and hard to implement in a biologically plausible way. To resolve this challenge, we begin by noting that

$$\mathcal{W}^{-1} = \left(\mathbb{I} - \hat{\mathcal{W}}\right)^{-1} = \sum_{k=0}^{\infty} \hat{\mathcal{W}}^k,$$

which holds if $||\hat{\mathcal{W}}|| < 1$. $\hat{\mathcal{W}}$ is a matrix that depends on the synaptic weights in the network. A similar relation holds for $\mathcal{M}^{-1}$. Since those matrices are contained in Eq. (15a), we substitute the equivalent sums into the relevant sub-expression and truncate the geometric series after the 0-th order, as in

$$\begin{aligned}
\mathcal{W}^{-1} W^{\mathrm{E}\leftarrow\mathrm{I}} D^{\mathrm{I}} \mathcal{M}^{-1} &= \left(\sum_{k=0}^{\infty} \hat{\mathcal{W}}^k\right) W^{\mathrm{E}\leftarrow\mathrm{I}} D^{\mathrm{I}} \left(\sum_{k=0}^{\infty} \hat{\mathcal{M}}^k\right) \\
&= W^{\mathrm{E}\leftarrow\mathrm{I}} D^{\mathrm{I}} + \hat{\mathcal{W}} W^{\mathrm{E}\leftarrow\mathrm{I}} D^{\mathrm{I}} + W^{\mathrm{E}\leftarrow\mathrm{I}} D^{\mathrm{I}} \hat{\mathcal{M}} + \left(\sum_{k=1}^{\infty} \hat{\mathcal{W}}^k\right) W^{\mathrm{E}\leftarrow\mathrm{I}} D^{\mathrm{I}} \left(\sum_{k=1}^{\infty} \hat{\mathcal{M}}^k\right) \\
&\approx W^{\mathrm{E}\leftarrow\mathrm{I}} D^{\mathrm{I}}.
\end{aligned}$$

The truncation to 0-th order in the last line should yield an acceptable approximation if synapses are sufficiently weak. The effect of higher-order interactions in the network can then be ignored. This approximation can be substituted into Eq. (15a) and yields an equation that resembles a backpropa-

gation rule in a feedforward network (E $\to$ I $\to$ E) with one hidden layer—the interneurons. The final, local approximation used for the simulations in the main text is then reached by replacing the output synapses of the interneurons by the transpose of their input synapses. While there is no mathematical argument why this replacement is valid, it turns out to be in the simulations, presumably because of a mechanism akin to feedback alignment [Lillicrap et al., 2016], see discussion in the main text. In feedback alignment, the matrix that backpropagates the errors is replaced by a random matrix B. Here, we instead use the feedforward weights in the layer below. Similar to the extension to feedback alignment of Akrout et al. [2019], those weights are themselves plastic. However, we believe that the underlying mechanism of feedback alignment still holds. The representation in the hidden layer (the interneurons) changes as if the weights to the output layer (the Pyr neurons) were equal to the weight matrix they are replaced with (here, the input weights to the PV neurons). To exploit this representation, the weights to the output layer then align to the replacement weights, justifying the replacement post-hoc (Fig. 1G).

Note that the condition for feedback alignment to provide an update in the appropriate direction ($e^T B^T W e > 0$, where e denotes the error, W the weights in the second layer, and B the random feedback matrix) reduces to the condition that $W^{\mathrm{E}\leftarrow\mathrm{I}} W^{\mathrm{I}\leftarrow\mathrm{E}}$ is positive definite (assuming the errors are full rank). One way of assuring this is a sufficiently positive diagonal of this matrix product, i.e. a sufficiently high correlation between the incoming and outgoing synapses of the interneurons. A positive correlation of these weights is one of the observations of Znamenskiy et al. [2018] and also a result of learning in our model.

While such a positive correlation is not necessarily present for all learning tasks or network models, we speculate that it will be for the task of learning an E/I balance in networks that obey Dale's law.

The same logic of using a 0-th order approximation of $\mathcal{W}^{-1}$ that neglects higher order interactions is employed to recover the inhibitory synaptic plasticity rule of Vogels et al. [2011] from Eq. (15b).

Overall, the local approximation of the learning rule relies on three assumptions: Slowly varying inputs, weak synaptic weights and alignment of input and output synapses of the interneurons. These assumptions clearly limit the applicability of the learning rules for other learning tasks. In particular, the learning rules will not allow the network to learn temporal sequences.

Supplementary Figures

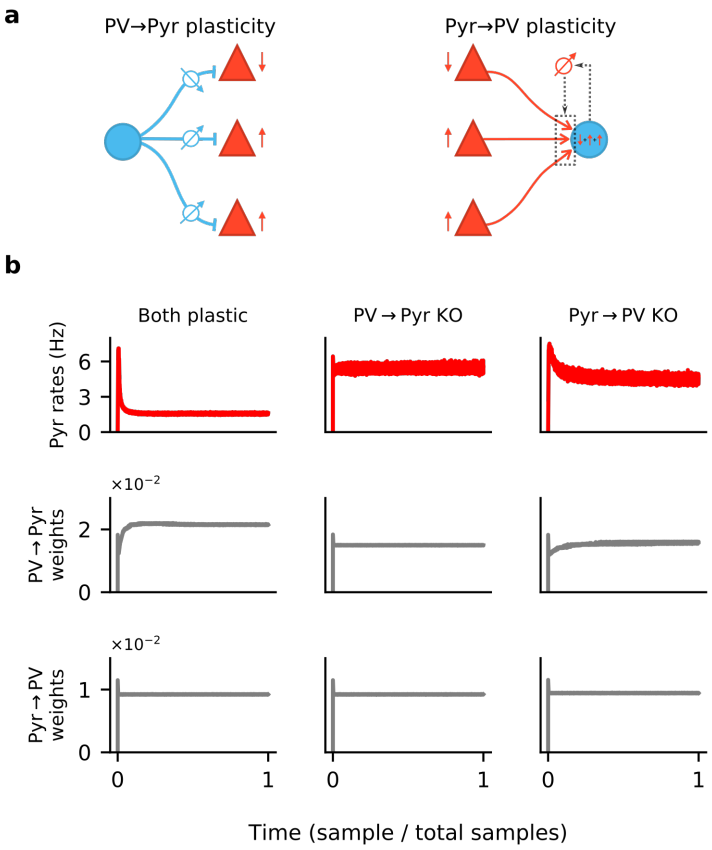

Figure S1: **Synaptic plasticity and convergence. a.** Schematics of PV → Pyr plasticity (left) and Pyr → PV plasticity (right). PV → Pyr plasticity follows a simple logic: A given inhibitory synapse is potentiated if the postsynaptic Pyr neuron fires above target, and is depressed if below. The Pyr → PV plasticity compares the excitatory input current received by the postsynaptic PV neuron to a target value, and adjusts the incoming synapses in proportion to this error and to presynaptic activity. **b.** Time plots of the Pyr population firing rate (top), mean of all PV → Pyr synaptic weights (middle) and Pyr → PV weights (bottom). Columns correspond to simulations in which both PV → Pyr and Pyr → PV plasticity are present (left), only Pyr → PV is present (middle), and only PV → Pyr is present (right).

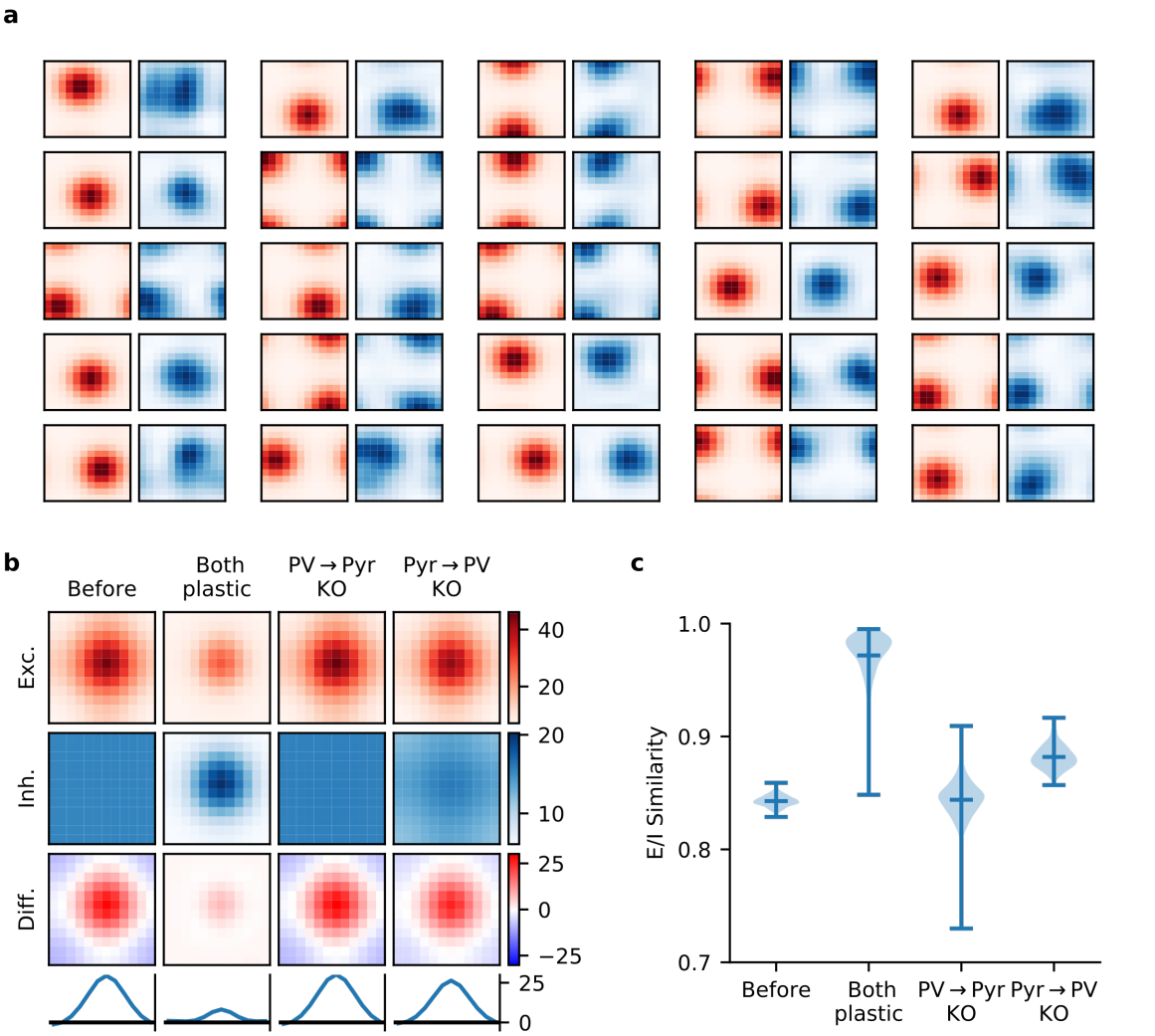

Figure S2: **Synaptic currents onto Pyr neurons.** **a.** Excitatory (red) and inhibitory (blue) synaptic currents onto a random selection of Pyr neurons, as a function of temporal and spatial stimulus frequency (averaged over all orientations), when both incoming and outgoing PV synapses are plastic. **b.** The network-averaged excitatory (first row) and inhibitory (second row) synaptic currents onto Pyr neurons, both centred according to the peak excitatory current before averaging. After averaging their difference is taken (third row), and a slice is plotted (bottom row). When both plasticities are present, currents are well-balanced across stimuli with a modest excitatory bias for preferred stimuli. **c.** Quality of E-I current co-tuning for every Pyr in the network quantified by the distribution of their cosine similarities. Only when both plasticities are present do most Pyr neurons receive well co-tuned E-I synaptic currents.

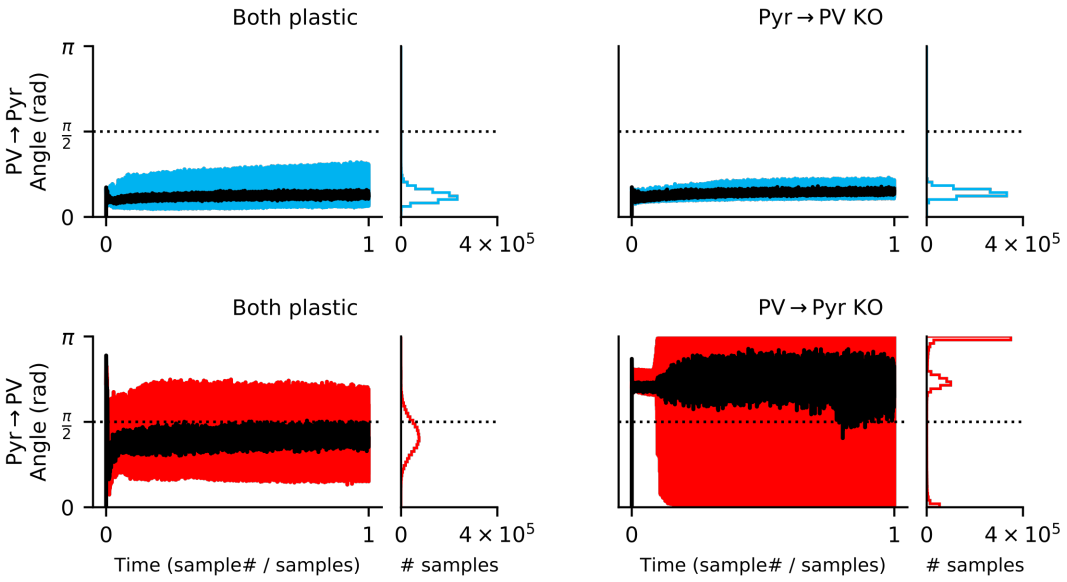

Figure S3: **Both in- and output synapses must be plastic for feedback alignment to occur.** In a network with both local rules (left column), the update to Pyr → PV synapses rapidly align to the gradient (i.e. when the angle between the approximate update and the gradient is below $\pi/2$; bottom left). While updates to the Pyr → PV weights occasionally point away from the gradient, 79% of samples are below $\pi/2$. For the knock-out (KO) experiments (right column), output plasticity closely follows the PV → Pyr gradient even if input plasticity is absent (upper right). In contrast, if output (PV → Pyr) plasticity is absent the approximate Pyr → PV rule does not follow the gradient (lower right).

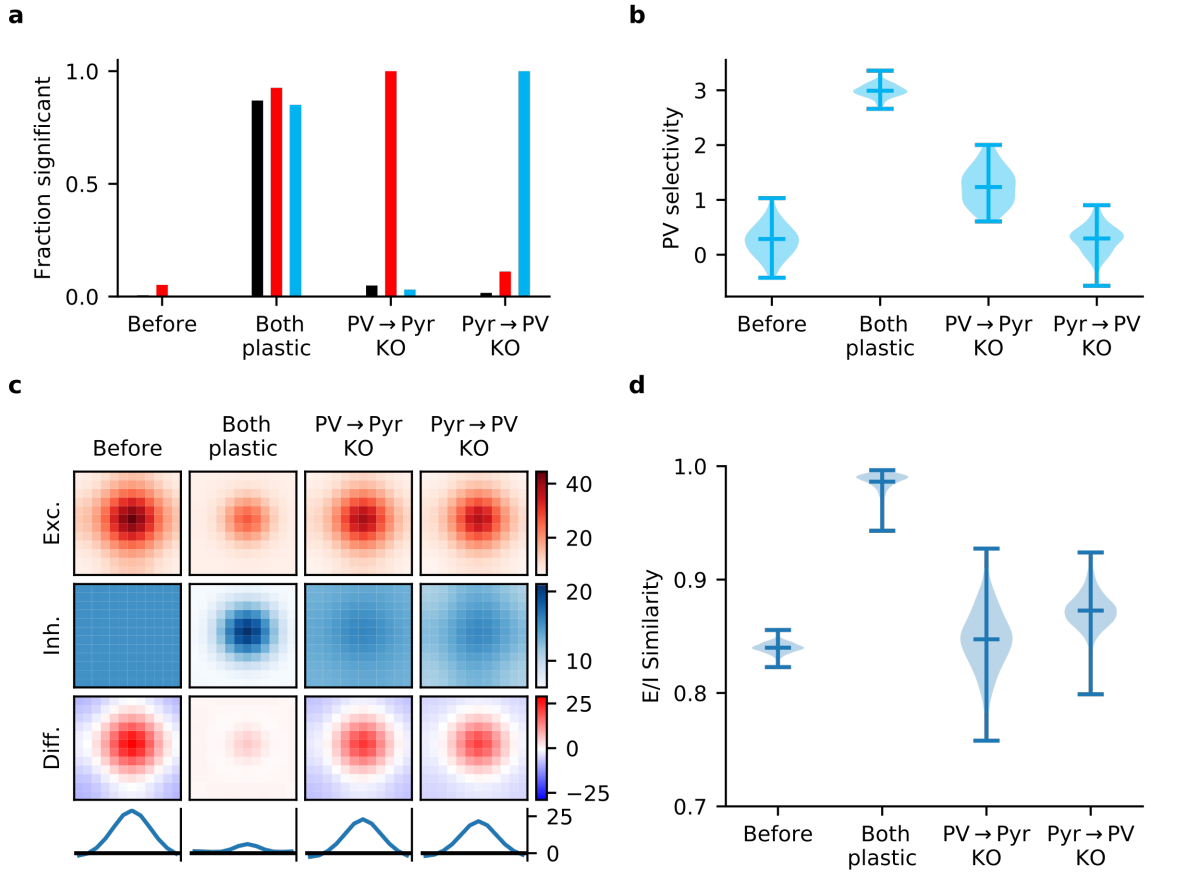

Figure S4: **Gradient rules also require plasticity of both in- and output synapses of PV interneurons.** **a.** In a network learning with the derived gradient rules of Eq. (15), significant correlations are reliably detected between response similarity (RS) and excitatory weights (red bars), RS and inhibitory weights (blue bars), and excitatory & inhibitory weights for reciprocally connected Pyr-PV cell pairs (black bars) only if both synapse types are plastic. **b.** Interneurons fail to develop stimulus selectivity if their input weights do not change according to the gradient rule of Eq. (15a). **c.** Synaptic currents onto Pyr neurons only develop reliable, strong excitatory-inhibitory (E/I) co-tuning if both in- and output synapses are updated using the gradient rules. Currents are averaged across all Pyr neurons after centering according to the neuron's preferred stimulus. The bottom row is a slice through the difference (third row) of the average excitatory (first row) and inhibitory currents (second row). **d.** Violin plot of the distribution of E/I synaptic current similarity values for all Pyr neurons in the network (see Methods).

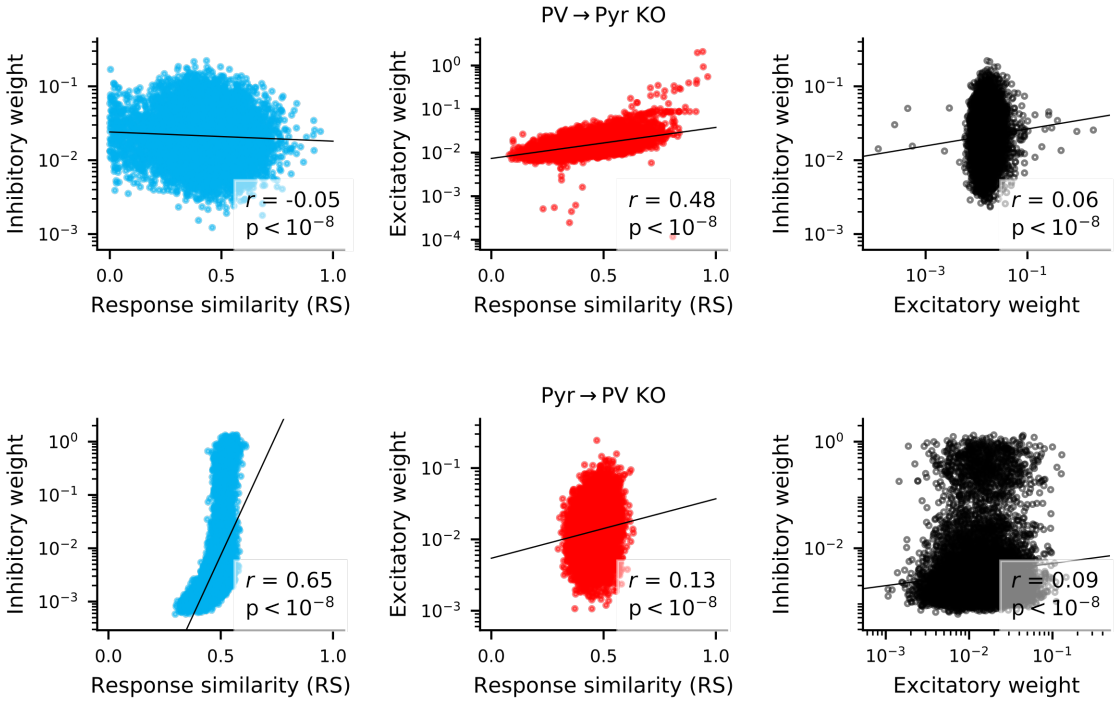

Figure S5: **Correlation between weights and response similarity.** Scatter plots containing every synapse in networks without PV → Pyr plasticity (top), or without Pyr → PV plasticity (bottom). Pearson correlation is always highly significant, though sometimes weak.

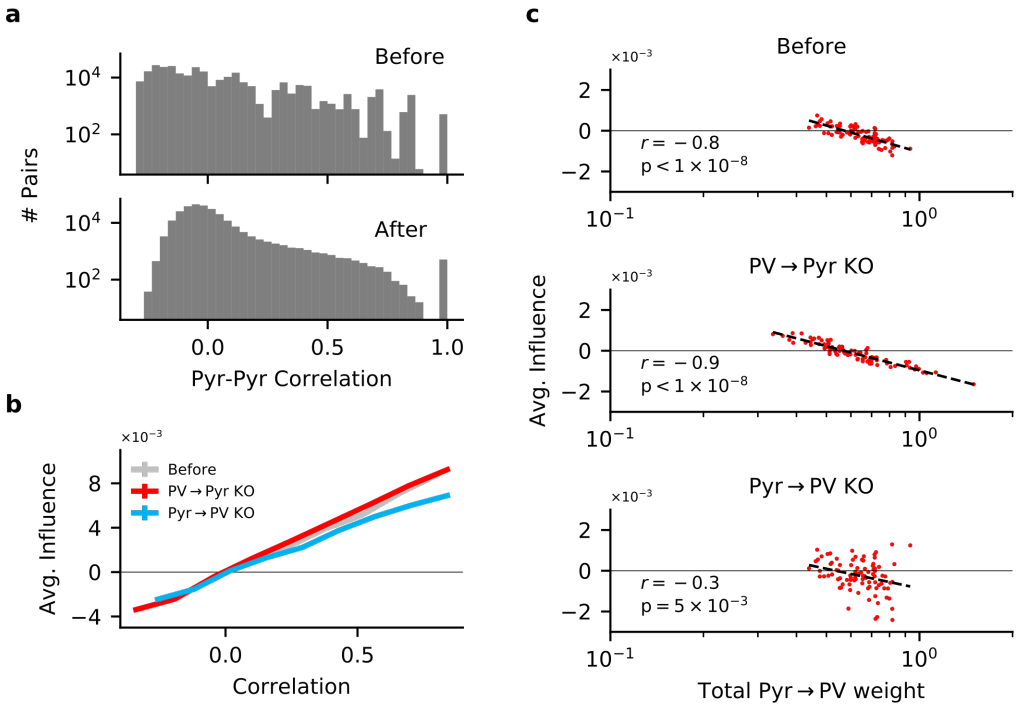

Figure S6: **In- and output plasticity together change correlations between pyramidal (Pyr) neurons, while plasticity knock-out (KO) eliminates feature competition. a.** Receptive-field correlations (Pearson) between Pyr neurons, before (top) and after (bottom) learning with both PV → Pyr and Pyr → PV synaptic plasticity. **b.** The effect of perturbing a Pyr neuron on the response of other Pyr neurons (to random stimuli) as a function of their receptive-field correlation (see Materials & Methods). On their own, both Pyr → PV and PV → Pyr plasticity have little effect on the feature amplification observed prior to learning. **c.** Despite the absence of feature competition on average in the KO networks, the total strength of Pyr → PV synapses from a given Pyr neuron is still predictive of its influence on the rest of the network: The stronger its total weight, the more likely a Pyr is to suppressing the response of other Pyr neurons.

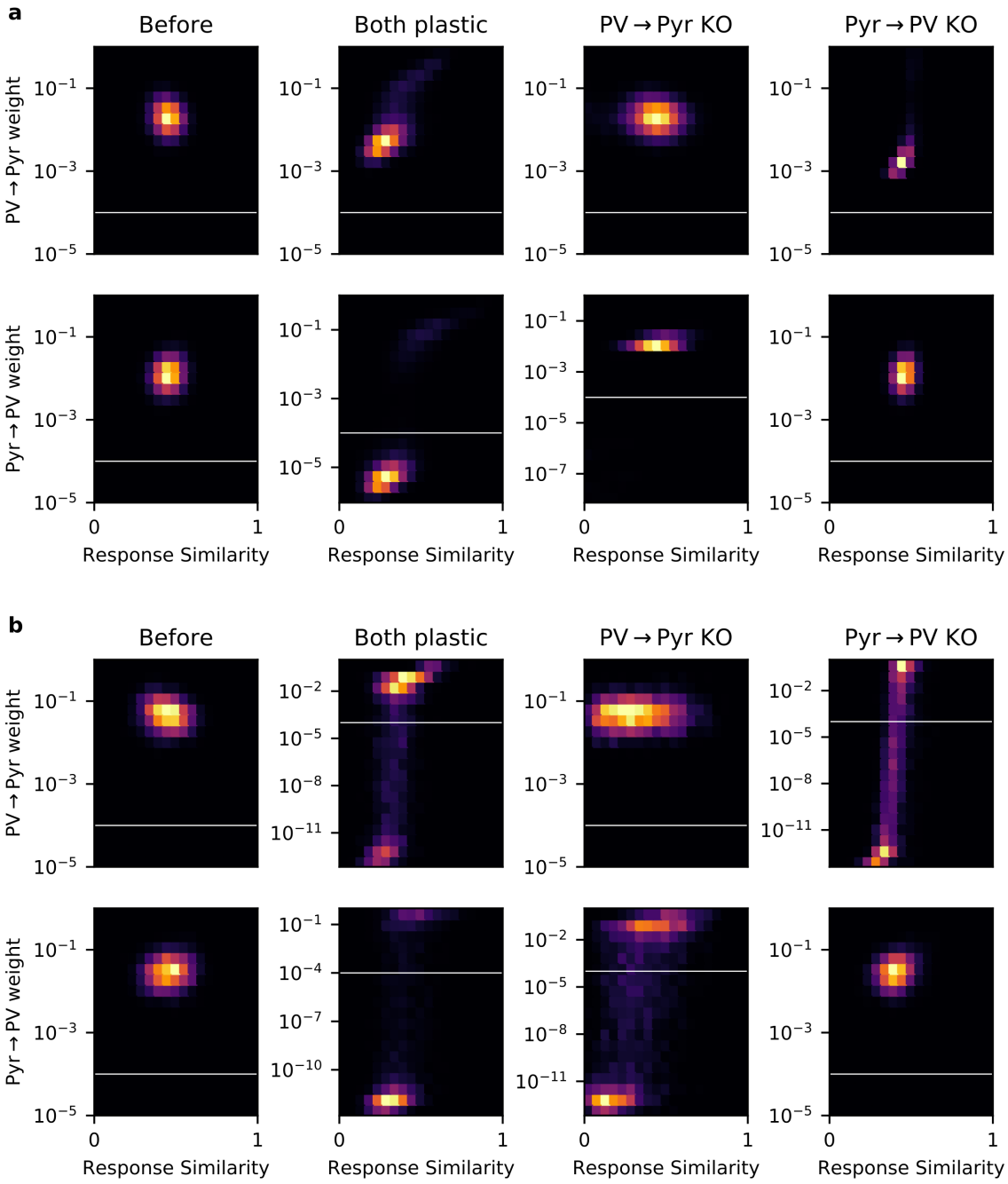

Figure S7: **Some networks contain experimentally undetectable weights. a.** Plots of 2D histograms for PV → Pyr (top) and Pyr → PV (bottom) weight versus response similarity (RS), in different networks trained with the local plasticity rules (columns). White lines indicate the threshold of experimental detectability. Any weight $< 10^{-4}$ is not included when computing Pearson's correlation between RS and synaptic weight, or weight-weight correlations. **b.** Same plots as (a), but for networks trained with the gradient rules.