Uncertainty-modulated prediction errors in cortical microcircuits

Katharina A. Wilmes[1,*], Mihai A. Petrovici[1], Shankar Sachidhanandam[1], Walter Senn[1]

1 Department of Physiology, University of Bern, Bern, Switzerland
* corresponding author: katharina.wilmes@unibe.ch

July 16, 2023

Abstract

To make contextually appropriate predictions in a stochastic environment, the brain needs to take uncertainty into account. Prediction error neurons have been identified in layer 2/3 of diverse brain areas. How uncertainty modulates prediction error activity and hence learning is, however, unclear. Here, we use a normative approach to derive how prediction errors should be modulated by uncertainty and postulate that such uncertainty-weighted prediction errors (UPE) are represented by layer 2/3 pyramidal neurons. We further hypothesise that the layer 2/3 circuit calculates the UPE through the subtractive and divisive inhibition by different inhibitory cell types, such as somatostatin-positive (SST), and parvalbumin-positive (PV) interneurons. By implementing the calculation of UPEs in a microcircuit model, we show that different cell types in cortical circuits can compute the means and variances of the stimulus distribution. With local activity-dependent plasticity rules, these computations can be learned context-dependently, and allow the prediction of upcoming stimuli and their distribution. Finally, we show that the resulting UPEs allow an organism to optimise its learning strategy with respect to external circumstances by enabling adaptive learning rates.

Introduction

Decades of cognitive research indicate that our brain maintains a model of the world, based on which it can make predictions about upcoming stimuli (Kveraga *et al.*, 2007; Cohen *et al.*, 2011). Predicting the sensory experience is useful for both perception and learning: Perception becomes more tolerant to uncertainty and noise when sensory information and predictions are integrated (Payzan-LeNestour and Bossaerts, 2011). Learning can happen when predictions are compared to sensory information, as the resulting prediction error indicates how to improve the internal model. In both cases, the uncertainties (associated with both the sensory information and the internal model) should determine how much weight we give to the sensory information relative to the predictions, according to theoretical accounts. Behavioural and electrophysiological studies indicate that humans indeed estimate uncertainty and adjust their behaviour accordingly (Payzan-LeNestour and Bossaerts, 2011; Walker *et al.*, 2019; Goris *et al.*, 2018; Cannon *et al.*, 2021). The neural mechanisms underlying uncertainty and prediction error computation are, however, less well understood. Recently, the activity of individual neurons of layer 2/3 cortical circuits in diverse cortical areas of mouse brains has been linked to prediction errors (visual, (Keller *et al.*, 2012; Zmarz and Keller, 2016; Fiser *et al.*, 2016; Attinger *et al.*, 2017; Gillon *et al.*, 2021), auditory (Eliades and Wang, 2008; Keller and Hahnloser, 2009), and somatosensory (Ayaz *et al.*, 2019)). These neural circuits consist of heterogeneous cell types, most of which are inhibitory. It has been suggested that prediction error activity results from an imbalance of excitatory and inhibitory inputs (Hertäg and Sprekeler, 2020; Hertäg and Clopath, 2022), and that the prediction is subtracted from the sensory input (see e.g. Rao and Ballard, 1999; Attinger *et al.*, 2017), possibly mediated by so-called somatostatin-positive interneurons (SSTs) (Attinger *et al.*, 2017). How uncertainty is influencing these computations has not yet been investigated. Prediction error neurons receive inputs from a diversity of inhibitory cell types (Fig. 1), the role of which is not completely understood. Here, we hypothesise that one role of inhibition is to modulate the prediction error neuron activity by uncertainty.

In this study, we use both analytical calculations and numerical simulations of rate-based circuit models with different inhibitory cell types to study circuit mechanisms leading to uncertainty-modulated prediction errors. First, we derive that uncertainty should divisively modulate prediction error activity and introduce uncertainty-modulated prediction errors (UPEs). We hypothesise that layer 2/3 prediction error neurons reflect such UPEs, and that different inhibitory cell types are involved in calculating the difference between predictions and stimuli compared to the uncertainty modulation. Based on experimental findings, we suggest that SSTs and PVs play the respective roles. We then derive biologically plausible plasticity rules for those cell types that allow to learn the means and variances from their inputs. Notably, because the information about the stimulus distribution is stored in the connectivity, single inhibitory cells encode the means and variances of their inputs in a context-dependent manner. Layer 2/3 pyramidal cells in this model hence encode uncertainty-modulated prediction errors context-dependently. We show that error neurons can additionally implement

out-of-distribution detection by amplifying large errors and reducing small errors with a nonlinear fI-curve (activation function). Finally, we demonstrate that UPEs effectively mediate an adjustable learning rate, which allows fast learning in high-certainty contexts, and reduces the learning rate, thus suppressing fluctuations in uncertain contexts.

Results

Normative theories suggests uncertainty-modulated prediction errors (UPEs)

In a complex, uncertain, and hence partly unpredictable world, it is impossible to avoid prediction errors. Some prediction errors will be the result of this variability or noise, other prediction errors will be the result of a change in the environment or new information. Ideally, only the latter should be used for learning, i.e., updating the current model of the world. The challenge our brain faces is to learn from prediction errors that result from new information, and less from prediction errors that result from noise. Hence, intuitively, if we learned that a kind of stimulus or context is very variable (high uncertainty), then a prediction error should have only little influence on our model. If, on the contrary, we learned that the kind of stimulus or context is not very variable (low uncertainty), a prediction error should have a larger impact on our model. This intuition of modulating prediction errors by the uncertainty associated with the stimulus or context is supported by both behavioural studies and normative theories of learning.

Before we suggest how cortical circuits compute such uncertainty-modulated prediction errors, we consider the normative solution to a simple association that a mouse can learn, and which involves a Gaussian-distributed stimulus. The setting we consider is to predict a somatosensory stimulus based on an auditory stimulus (Fig. 1A). The auditory stimulus a is fixed, and the subsequent somatosensory stimulus s is variable and sampled from a Gaussian distribution (s $\sim \mathcal{N}(\mu, \sigma)$, Fig. 1B). The optimal (maximum-likelihood) prediction is given by the mean of the stimulus distribution. Framed as an optimisation problem, the goal is to adapt the internal model of the mean $\hat{\mu}$ such that the probability of observing samples s from the true distribution of whisker deflections is maximised given this model.
Hence, stochastic gradient ascent learning on the log likelihood suggests that with each observation s, the prediction, corresponding to the internal model of the mean, should be updated as follows to approach the maximum likelihood solution:

$$\Delta\hat{\mu} \propto \frac{\partial}{\partial\hat{\mu}}(\log L) = \frac{1}{\sigma^2}(s - \hat{\mu}). \tag{1}$$

According to this formulation, the update for the internal model should be the prediction error scaled inversely by the variance σ^2. Therefore, we propose that prediction errors should be modulated by uncertainty.

Computation of UPEs in cortical microcircuits

How can uncertainty modulation be achieved by cortical microcircuits? Prediction errors can be either positive or negative but neuronal firing rates are always positive. Because baseline firing rates are low in layer 2/3 pyramidal cells (e.g., Niell and Stryker, 2008), positive and negative prediction errors were suggested to be represented by distinct neuronal populations (Keller and Mrsic-Flogel, 2018), which is in line with experimental data (Jordan and Keller, 2020). We, therefore, decompose the UPE into a positive UPE^+ and a negative UPE^- component (Fig. 1C,D):

$$\text{UPE} = \text{UPE}^+ - \text{UPE}^- = \frac{1}{\sigma^2}\lfloor s - \mu \rfloor^+ - \frac{1}{\sigma^2}\lfloor \mu - s \rfloor^+ \tag{2}$$

, where $\lfloor ... \rfloor^+$ denotes rectification at 0.
It has been suggested that prediction errors are computed in error neurons by subtracting the prediction from the stimulus input (or vice versa) (Attinger *et al.*, 2017). Inhibitory interneurons provide the subtraction, resulting in an excitation-inhibition balance when they match (Hertäg and Sprekeler, 2020). To represent a UPE, error neurons need to additionally be divisively modulated by the uncertainty. Depending on synaptic properties, such as reversal potentials, inhibitory neurons can have subtractive or divisive influences on their postsynaptic targets. Therefore, we propose that an inhibitory cell type that divisively modulates prediction error activity represents the uncertainty. We hypothesise, first, that in positive prediction error circuits, inhibitory interneurons with subtractive inhibitory effects represent the mean μ of the prediction. Second, we hypothesise that inhibitory interneurons with divisive inhibitory effects represent the uncertainty σ^2 of the prediction (Fig. 1C,D). A layer 2/3 pyramidal cell that receives these sources of inhibition then reflects the uncertainty-modulated prediction error.
More specifically, we propose that the SSTs are involved in the computation of the difference between predictions and stimuli, as this was suggested before (Attinger *et al.*, 2017), and that the PVs provide the uncertainty modulation. In line with this, prediction error neurons in layer 2/3 were shown to receive subtractive inhibition from somatostatin (SST) and divisive inhibition from parvalbumin (PV) interneurons (Wilson *et al.*, 2012). However, SSTs can also have divisive effects, and PVs can have subtractive effects, dependent on circuit and postsynaptic properties (Seybold *et al.*, 2015; Lee *et al.*, 2012; Dorsett *et al.*, 2021).

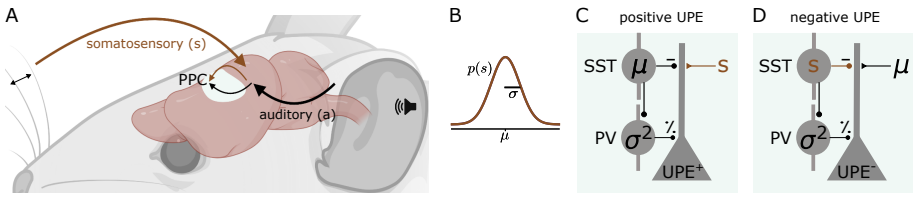

Figure 1: **Computation of UPEs in cortical microcircuits** A: a mouse learns the association between a sound (a) and a whisker deflection (s). The posterior parietal cortex (PPC) receives inputs from both somatosensory and auditory cortex. B: The whisker stimulus intensities are drawn from a Gaussian distribution with mean μ and standard deviation σ. C: Positive prediction error circuit consisting of three cell types: layer 2/3 pyramidal cells (triangle), somatostatin-positive interneurons (SST, circle) and parvalbumin-positive interneurons (PV). SSTs represent the mean prediction, and PVs the variance. D: Negative prediction error circuit, similar to C, SST now represent the stimulus and the mean prediction is an excitatory input.

Local inhibitory cells learn to represent the mean and the variance given an associative cue

As discussed above, how much an individual sensory input contributes to updating the internal model should depend on the uncertainty associated with the sensory stimulus. As the estimation of the uncertainty requires multiple stimulus samples, the uncertainty cannot be estimated from a single input. Additionally, the uncertainty may be context-dependent. Therefore, our brain needs to have a mechanism to learn about the uncertainty from past instances of the sensory input. Let us consider the simple example from above, in which a sound stimulus represents a context, which is associated with a particular amount of uncertainty. Here, we investigate whether the presentation of the sound can elicit activity in the PVs that reflects the expected uncertainty of the situation. To investigate whether a sound can cause activity in SSTs and PVs that reflects the mean and the variance of the whisker stimulus intensities, respectively, we simulated a rate-based circuit model, consisting of pyramidal cells and the relevant inhibitory cell types. This circuit receives both the sound and the whisker stimuli as inputs.

SSTs learn to estimate the mean

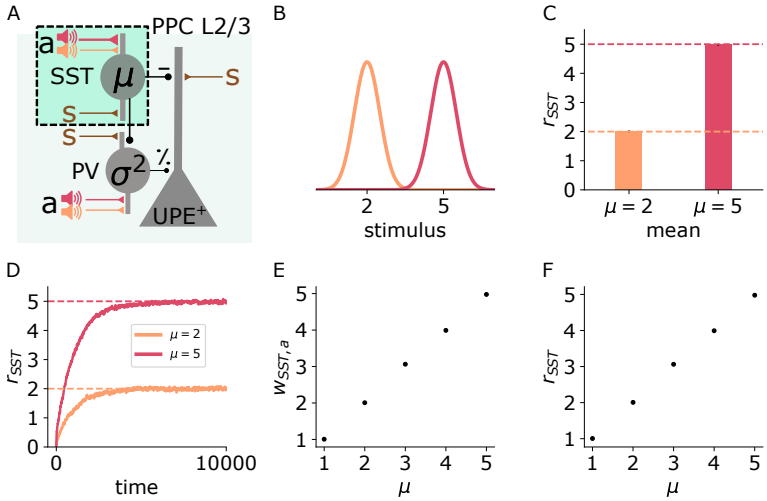

Figure 2: **SSTs learn to represent the mean context-dependently** Illustration of the changes in the positive prediction error circuit. Thicker lines denote stronger weights. B: Two different tones (red, orange) are associated with two somatosensory stimulus distributions with different means (red: high, orange: low). C: SST firing rates (mean and std) during stimulus input. D: SST firing rates over time for low (orange) and high (red) stimulus means. E: Weights (mean and std) from sound a to SST for different values of μ . F: SST firing rates (mean and std) for different values of μ.

With our circuit model, we first investigate whether SSTs can learn to represent the mean of the stimulus distribution. In this model, SSTs receive whisker stimulus inputs s, drawn from Gaussian distributions (Fig. 2B), and an input from a higher level representation of the sound a (which is either on or off, see Methods). The connection weight from the sound representation to the SSTs is plastic according to a local activity-dependent plasticity rule. The aim of this rule is to minimise the difference between the activation of the SSTs caused by the sound input (which is to be learned) and the activation of the SSTs by the whisker stimulus (which nudges the SST activity in the right direction). The learning rule ensures that the auditory input alone causes SSTs to fire at the desired rate. After learning, the weight and the average SST firing rate reflect the mean of the presented whisker stimulus intensities (Fig. 2C-F).

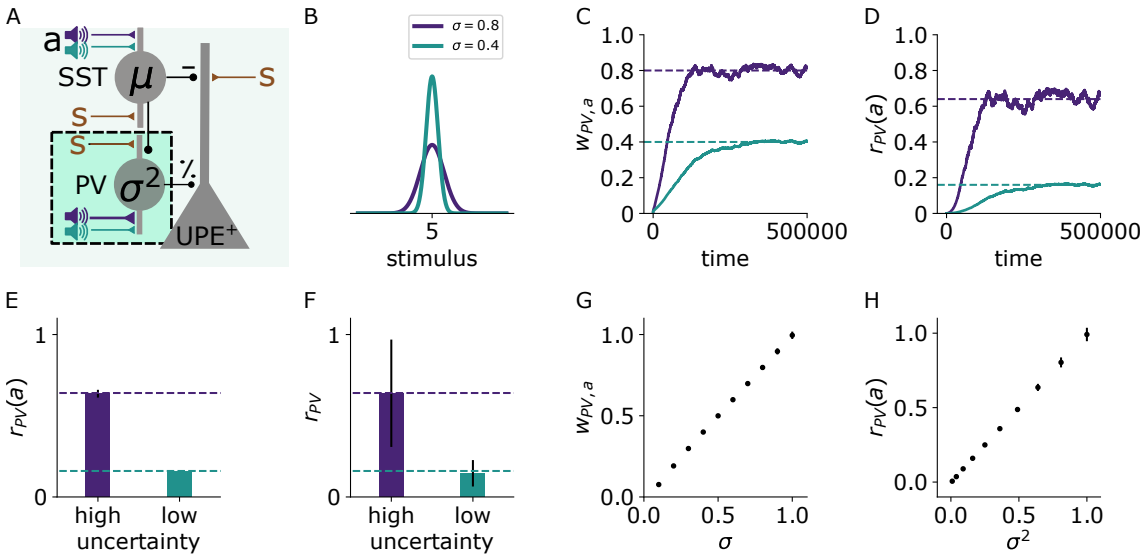

Figure 3: **PVs learn to estimate the variance context-dependently.** A: Illustration of the changes in the positive prediction error circuit. Thicker lines denote stronger weights. B: Two different tones (purple, green) are associated with two somatosensory stimulus distributions with different variances (purple: high, green: low). C: Weights from sound a to PV over time for two different values of stimulus variance (high: $\sigma = 0.8$ (purple), low: $\sigma = 0.4$ (green)). D: PV firing rates over time given sound input (without stimulus input) for low (green) and high (purple) stimulus variance. E: PV firing rates (mean and std) given sound input for low and high stimulus variance. F: PV firing rates (mean and std) during sound and stimulus input. G: Weights (mean and std) from sound a to PV for different values of σ. H: PV firing rates (mean and std) given sound input for different values of σ^2.

PVs learn to estimate the variance context-dependently

We next addressed whether PVs can estimate and learn the variance locally. To estimate the variance of the whisker deflections s, the PVs have to estimate $\sigma^2[s] = \mathbb{E}_s[(s - \mathbb{E}[s])^2] = \mathbb{E}_s[(s - \mu)^2]$. To do so, they need to have access to both the whisker stimulus s and the mean μ. PVs in PPC respond to sensory inputs in diverse cortical areas (S1: Sachidhanandam *et al.*, 2016) and are inhibited by SSTs in layer 2/3, which we assumed to represent the mean. Finally, for calculating the variance, these inputs need to be squared. PVs were shown to integrate their inputs supralinearly (Cornford *et al.*, 2019), which could yield an approximation to the variance.

In our circuit model, we next tested whether the PVs can learn to represent the variance of an upcoming whisker stimulus based on a context provided by an auditory input (Fig. 3A). Two different auditory inputs (Fig. 3B purple, green) are paired with two whisker stimulus distributions that differ in their variances (green: low, purple: high). The synaptic connection from the auditory input to the PVs is plastic according to the same local activity-dependent plasticity rule as the connection to the SSTs. With this learning rule, the weight onto the PV becomes proportional to σ (Fig. 3C), such that the PV firing rate becomes proportional to σ^2 on average (Fig. 3D). The average PV firing rate is exactly proportional to σ^2 with a quadratic activation function $\phi_{PV}(x)$ (Fig. 3D-F,H) and monotonically increasing with σ^2 with other choices of activation functions (Suppl. Fig. 9). This holds both when the sound input is presented alone (Fig. 3D,E,H) or when it is paired with whisker stimulation (Fig. 3F). Notably, a single PV neuron is sufficient to encode variances of different contexts, because the context-dependent σ is stored in the connection weights.

To estimate the variance, the mean needs to be subtracted from the stimulus samples. A faithful mean subtraction is only ensured if the weights from the SSTs to the PVs ($w_{\mathrm{PV,SST}}$) matches the weights from the stimuli s to the PVs ($w_{\mathrm{PV},s}$; see Suppl. Figs. 14,15). The weight $w_{\mathrm{PV,SST}}$ can be learned to match the weight $w_{\mathrm{PV},s}$ with a local activity-dependent plasticity rule (see Suppl. Fig. 10 and Suppl. Methods).
The PVs can similarly estimate the uncertainty in negative prediction error circuits (Suppl. Fig. 11). In these circuits, SSTs represent the current sensory stimulus, and the mean prediction is an excitatory input to both negative prediction error neurons and PVs.

Calculation of the UPE in Layer 2/3 error neurons

Layer 2/3 pyramidal cell dendrites are capable of generating NMDA and calcium spikes, which cause a nonlinear integration of inputs. Such a nonlinear integration of inputs is convenient when the mean input changes and the current prediction differs strongly from the new mean of the stimulus distribution. In this case, the PV firing rate will increase for larger errors, and inhibit error neurons more strongly than indicated by the learned variance estimate. The nonlinearity compensates this increased inhibition by PVs, such that in the end, layer 2/3 cell activity reflects

an uncertainty-modulated prediction error (Fig. 4E) in both negative (Fig. 4A) and positive (Fig. 4B) prediction error circuits. A stronger nonlinearity has an interesting effect: error neurons elicit much larger responses to outliers than to stimuli that match the predicted distribution—a cell-intrinsic form of *out-of-distribution detection*.

To ensure a comparison between the stimulus and the prediction, the weights from the SSTs to the UPE neurons need to match the weights from the stimulus s to the UPE neuron and from the mean representation to the UPE neuron, respectively. With inhibitory plasticity (target-based, see Suppl. Methods), the weights from the SSTs can learn to match the incoming excitatory weights (Suppl. Fig. 12).

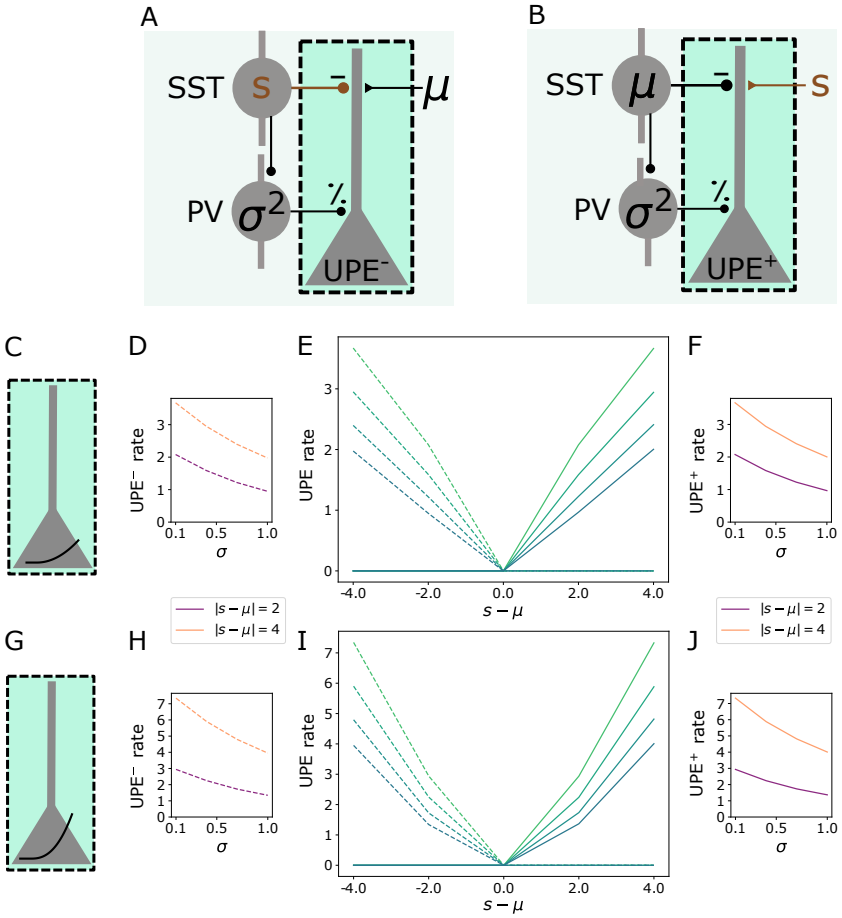

Figure 4: **Calculation of the UPE in layer 2/3 error neurons** A: Illustration of the negative prediction error circuit. B: Illustration of the positive prediction error circuit. C: Illustration of error neuron with a nonlinear integration of inputs ($k = 2$). D: firing rate of the error neuron in the negative prediction error circuit (UPE^-) as a function of σ for two values of $|s - \mu|$. E: Rates of both UPE^+ and UPE^--representing error neurons as a function of the difference between the stimulus and the mean ($s - \mu$). F: firing rate of the error neuron in the positive prediction error circuit (UPE^+) as a function of σ for two values of $|s - \mu|$. G: Illustration of an error neuron with a non-linear activation function $k = 2.5$. H-J: same as D-F for error neurons with $k = 2.5$.

Interactions between representation neurons and error neurons

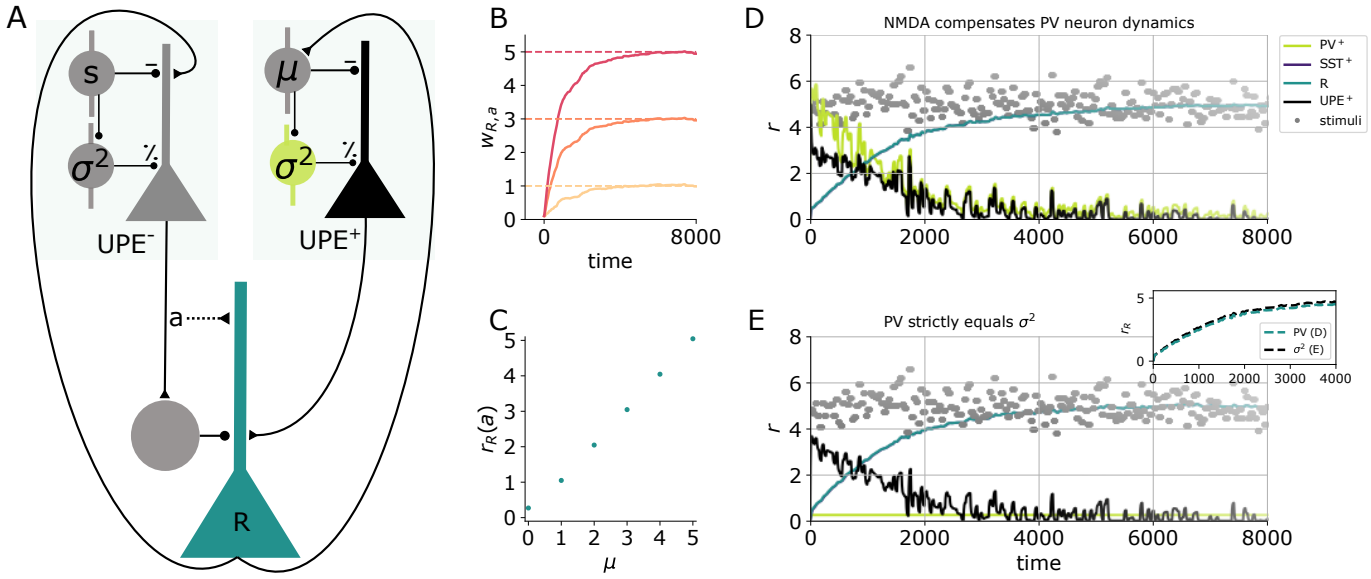

Figure 5: **Learning the mean representation with UPEs** A: Illustration of the circuit. A representation neuron (turquoise, R) receives input from both positive and negative prediction error circuits (UPE^+ and UPE^-) and projects back to them (connectivity is simplified in the illustration, see Methods for the detailed connectivity matrix). The UPE^- has a negative impact on the firing rate of the representation neuron (r_R). A weight $w_{R,a}$ from the higher level representation of the sound a is learned. B: Weights from sound a to R over time for different values of μ ($\mu \in [1, 3, 5]$). C: R firing rates given sound input for different values of μ (mean and std from timestep 100000 to the end of the simulation). D: Activity of the different cell types (PV: light green, R: green, UPE: black) and whisker stimulus samples (grey dots) over time. Learning the mean representation with PVs (light green) reflecting the MSE at the beginning, which is compensated by nonlinear activation of L2/3 neurons (black). The evolution of the mean rate of neuron R (turquoise) is similar to the perfect case in E (inset shows comparison). E: Same colour code as in D. Learning the mean representation assuming PVs (light green) perfectly represent the variance.

The theoretical framework of predictive processing includes both prediction error neurons and representation neurons, the activity of which reflects the internal model and should hence be compared to the sensory information. To make predictions for the activity of representation neurons, we expand our circuit model with this additional cell type. We first show that a representation neuron R can learn a representation of the stimulus mean given inputs from L2/3 error neurons. The representation neuron receives inputs from positive and negative prediction error neurons and from a higher level representation of the sound a (Fig. 5A). It sends its current mean estimate to the error circuits by either targeting the SSTs (in the positive circuit) or the pyramidal cells directly (in the negative circuit). Hence, in this recurrent circuit, the SSTs inherit the mean representation instead of learning it. After learning, the weights from the sound to the representation neuron and the average firing rate of this representation neuron reflects the mean of the stimulus distribution (Fig.5B,C).

Second, we show that a circuit with prediction error neurons that exhibit NMDA spikes (as in Fig. 4) approximates an idealised circuit, in which the PV rate perfectly represents the variance (Fig. 5D,E, see inset for comparison of the two models). Also in this recurrent circuit, PVs learn to reflect the variance, as the weight from the sound representation a is learned to be proportional to σ (Suppl. Fig. 13).

Finally, we studied the link between circuit connectivity and biases in the mean representation. Inhibitory plasticity ensures that the inhibitory connections (from SSTs to PVs and from SSTs to UPEs) match their excitatory counterparts (from the stimulus representation to the PVs and UPEs). This ensures that no biases are introduced in the computation. Interestingly, biases in perception have been linked to numerous psychological conditions (Lieder *et al.*, 2019; Ueda *et al.*, 2022; Karaminis *et al.*, 2016). To better understand how biases depend on the connectivity in our model, we perturbed connections from the SSTs to the prediction error neurons. A 100% increase in the connection from SSTs to both prediction error neurons leads to an underestimation of the mean, which increases with increasing mean (Suppl. Fig. 16). To better understand, why this is the case, we perturbed these connections in each of the two error circuits individually. Increasing this connection only in the positive prediction error circuit has a very similar effect (Suppl. Fig. 17), while increasing this connection only in the negative prediction error circuit has the opposite effect and leads to an overestimation of the mean (Suppl. Fig. 18). The total effect in this setting is hence mostly driven by the perturbation in the positive circuit. This is expected because the total error is dominated by positive prediction errors as the initial mean prediction was too small. Nevertheless, the learned representation is proportional to the mean.

Predictions for different cell types

Our model makes predictions for the activity of different cell types for positive and negative prediction errors (e.g. when a mouse receives whisker stimuli that are larger (Fig. 6A, black) or smaller (Fig. 6G, grey) than expected) in contexts associated with different amounts of uncertainty (e.g., the context could be indicated by sounds that are associated with a high-certainty (green) or an uncertain context (purple)). Our model suggests that there are two types of interneurons that provide subtractive inhibition to the prediction error neurons (presumably SST subtypes): in the positive prediction error circuit (SST^+), they signal the expected value of the whisker stimulus intensity (Fig. 6B,H). in the negative prediction error circuit (SST^-) they signal the whisker stimulus intensity (Fig. 6C,I). We further predict that interneurons that divisively modulate prediction error neuron activity represent the uncertainty (presumably PVs). Those do not differ in their activity between positive and negative circuits, and may even be shared across the two circuits: in both positive and negative prediction error circuits, these cells signal the variance (Fig. 6D,J). L2/3 pyramidal cells that encode prediction errors signal uncertainty-modulated positive prediction errors (Fig. 6E) and uncertainty-modulated negative prediction errors (Fig. 6L), respectively. Finally, it has been proposed that there are so-called internal representation neurons (Keller and Mrsic-Flogel, 2018). In our case, those neurons represent the predicted mean of the associated whisker deflections. Our model predicts that upon presentation of an unexpected whisker stimulus, those internal representation neurons adjust their activity to represent the new whisker deflection depending on the variability of the associated whisker deflections: they adjust their activity more (given equal deviations from the mean) if the associated whisker deflections are less variable.

The following experimental results are in line with our predictions: First, putative inhibitory neurons (narrow spiking units) in macaque anterior cingulate cortex increased their firing rates in periods of high uncertainty (Banaie Boroujeni *et al.*, 2021). These could correspond to the PVs in our model. Second, prediction error activity seems to be indeed smaller for less predictable, and hence more uncertain, contexts: Mice that were trained in a predictable, certain environment (where locomotion and visual flow match) were compared to mice trained in an unpredictable, uncertain environment (Attinger *et al.*, 2017, they saw a video of visual flow that was independent of their locomotion:). It was shown that layer 2/3 activity towards mismatches in locomotion and visual flow is smaller in the mice, which were raised in the unpredictable environment.

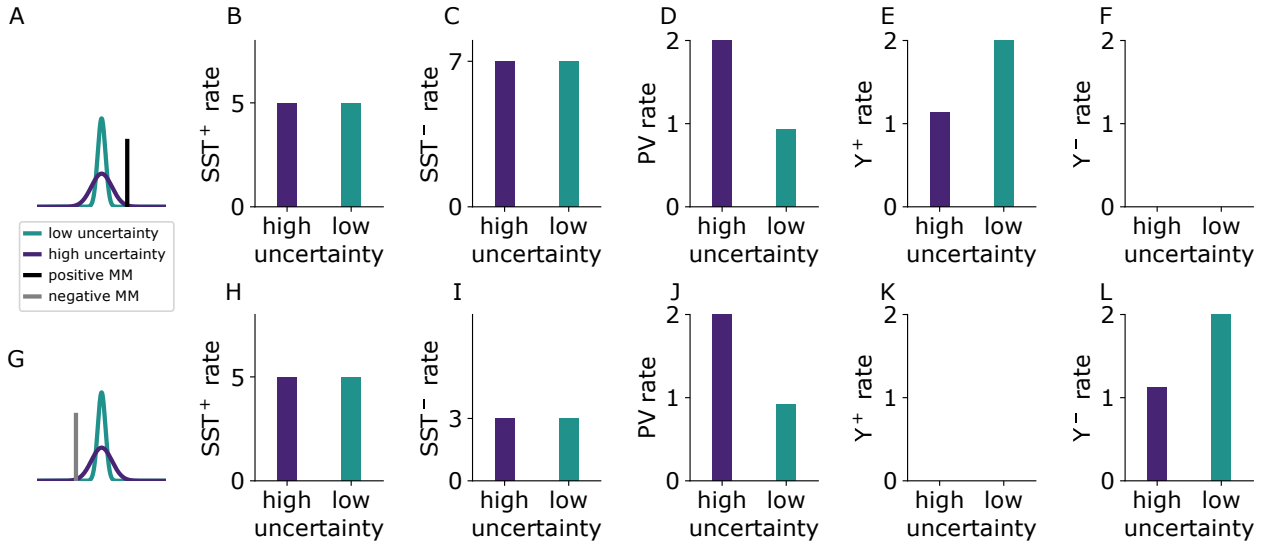

Figure 6: **Cell-type specific experimentally testable predictions** A: Illustration of the two experienced stimulus distributions with different variances that are associated with two different sounds (green, purple). The presented mismatch stimulus (black) is larger than expected (positive prediction error). B-F: Firing rates of different cell types to positive prediction errors when a sound associated with high (purple) or low (green) uncertainty is presented. G: As in A. The presented mismatch stimulus (grey) is smaller than expected (negative prediction error). H-L: Firing rates of different cell types to the negative mismatch when a sound associated with high (purple) or low (green) uncertainty is presented.

The effective learning rate is automatically adjusted with UPEs

To test whether UPEs can automatically adjust the effective learning rate of a downstream neural population, we looked at two contexts that differed in the amount of uncertainty, and compared how the mean representation evolves with and without UPEs. Indeed, in a low uncertainty setting, the mean representation can be learned faster with UPEs (in comparison to unmodulated, Fig. 7A,C). In a high uncertainty setting, the effective learning rate is smaller, and the mean representation is less variable than in the unmodulated case (Fig. 7B,D). The standard deviation of the firing rate

increases only sublinearly with the standard deviation of the inputs (Fig. 7E). In summary, uncertainty-modulation of prediction errors enables an adaptive learning rate modulation.

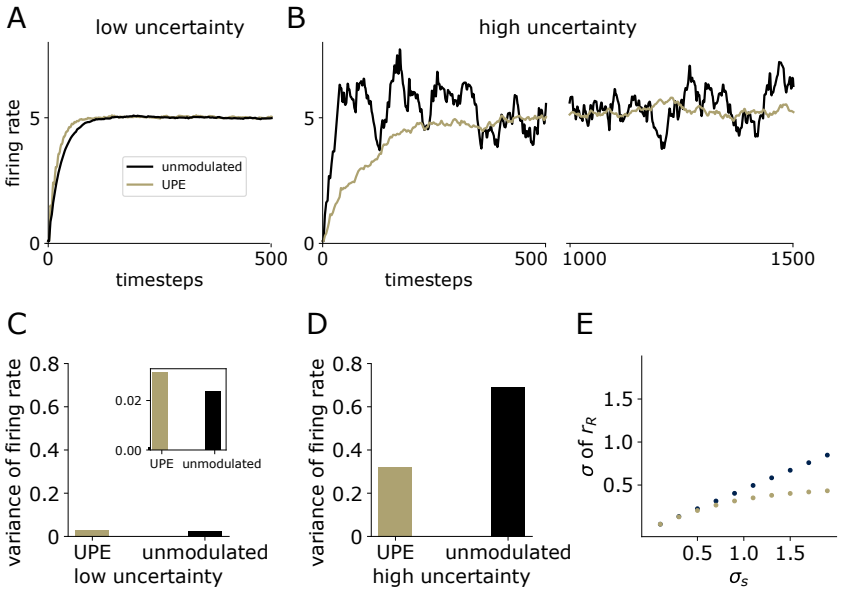

Figure 7: **Effective learning rate is automatically adjusted with UPEs** A,B: Firing rate over time of the representation neuron in a circuit with uncertainty-modulated prediction errors (gold) and in a circuit with unmodulated errors (black) in a low uncertainty setting (A) and a high uncertainty setting (B), C: variance of the firing rate of the representation neuron in the low uncertainty setting (inset has a different scale, outer axis scale matches the one in D), D: variance of the firing rate of the representation neuron in the high uncertainty setting. E: Variability (σ) of the firing rate r_R as a function of the standard deviation of the presented stimulus distribution σ.

Discussion

Based on normative theories, we propose that the brain uses uncertainty-modulated prediction errors. In particular, we hypothesise that layer 2/3 prediction error neurons represent prediction errors that are inversely modulated by uncertainty. Here we showed that different inhibitory cell types in layer 2/3 cortical circuits can compute means and variances and thereby enable pyramidal cells to represent uncertainty-modulated prediction errors. We further showed that the cells in the circuit are able to learn to predict the means and variances of their inputs with local activity-dependent plasticity rules. Our study makes experimentally testable predictions for the activity of different cell types, in particular, PV and SST interneurons, prediction error neurons and representation neurons. Finally, we showed that circuits with uncertainty-modulated prediction errors enable adaptive learning rates, resulting in fast learning when uncertainty is low and slow learning to avoid detrimental fluctuations when uncertainty is high.

Our theory has the following notable implications: The first implication concerns the hierarchical organisation of the brain. At each level of the hierarchy, we find similar canonical circuit motifs that receive both feedforward (from a lower level) and feedback (from a higher level, predictive) inputs that need to be integrated. We propose that uncertainty is computed on each level of the hierarchy. This enables uncertainty estimates that are specific to the processing level of a particular area. Additionally, adding uncertainty-modulated prediction errors from different hierarchical levels according to the predictive coding model (Rao and Ballard, 1999) yields Bayes-optimal weighting of feedback and feedforward information, which can be reconciled with human behaviour (Payzan-LeNestour and Bossaerts, 2011). Two further important implications result from uncertainty being stored in the afferent connections to the PVs. First, this implies that the same PV cell can store different uncertainties depending on the context, which is encoded in the pre-synaptic activation. Second, fewer PVs than pyramidal cells are required for the mechanism, which is ompatible with the 80/20 ratio of excitatory to inhibitory cells in the brain.

A diversity of proposals about the neural representation of uncertainty exist. For example, it has been suggested that uncertainty is represented in single neurons by the width (Fischer and Peña, 2011), or amplitude of their responses (Ma *et al.*, 2006), or implicitly via sampling (neural sampling hypothesis; Petrovici *et al.*, 2013; Buesing *et al.*, 2011; Berkes *et al.*, 2011), or rather than being represented by a single feature, can be decoded from the activity of an entire population (Dehaene *et al.*, 2021). While we suggest that uncertainty is represented by PVs to modulate prediction error responses, we do not claim that this is the sole representation of uncertainty in neuronal circuits.

We claim that the uncertainty represented by PVs in our theoretical framework corresponds to expected uncertainty that results from noise or uncertainty in the environment, and should therefore decrease the learning rate. The uncertainty-modulation is also related to the Kalman gain in the Kalman filter. The Kalman gain, however, additionally takes into account *estimate uncertainty*—i.e. the uncertainty associated with the internal model—, which decreases with the number of observations, and ideally increases again with unexpected uncertainty. The presence of unexpected rewards, punishments, or surprise, can be viewed as an increase in unexpected uncertainty that should increase the learning rate. VIPs could be responsible for signalling unexpected uncertainty. They were shown to provide potent disinhibition of pyramidal cells, are activated by reward, punishment and surprise (Pi *et al.*, 2013), and inhibit PVs in layer 2/3. Hence, they could increase error activity, which would in turn result in a larger learning signal. In general, interneurons are heavily innervated by different kinds of neuromodulators (Lee *et al.*, 2013; Prönneke *et al.*, 2015). Therefore, neuromodulators could have powerful control over error neuron activity and hence perception and learning.

Uncertainty estimation is relevant for Bayes-optimal integration of different sources of information, e.g., different modalities (Ernst and Banks, 2002; Fetsch *et al.*, 2012) or priors and sensory information. Here, we present a circuit implementation for weighing sensory information according to its uncertainty. It has previously been suggested that Bayes-optimal multi-sensory integration could be achieved in single neurons (Fetsch *et al.*, 2012; Jordan *et al.*, 2022). Our proposal is complementary to this solution in that uncertainty-modulated errors can be forwarded to other cortical and subcortical circuits at different levels of the hierarchy where they can be used for inference and learning. It further allows for a context-dependent integration of sensory inputs.

Multiple neurological disorders, such as autism spectrum disorder or schizophrenia, are associated with maladaptive contextual uncertainty-weighting of sensory and prior information (Goris *et al.*, 2021; Lawson *et al.*, 2014; Van de Cruys *et al.*, 2014; Lawson *et al.*, 2017; Shi *et al.*, 2022). These disorders are also associated with aberrant inhibition, e.g. ASD is associated with an excitation-inhibition imbalance (Rubenstein and Merzenich, 2003) and reduced inhibition (Harada *et al.*, 2011; Gaetz *et al.*, 2014). Interestingly, PV cells, in particular chandelier PV cells, were shown to be reduced in number and synaptic strength in ASD (Juarez and Martínez Cerdeño, 2022). Our theory provides one possible explanation how deficits in uncertainty-weighting on the behavioural level could be linked to altered PVs on the circuit level.

To provide experimental predictions that are immediately testable, we suggested specific roles for SSTs and PVs, as

they were shown to subtractively and divisively modulate pyramidal cell activity, respectively. I principle, our theory more generally posits that any subtractive or divisive inhibition could implement the suggested computations. With the emerging data on inhibitory cell types, subtypes of SSTs and PVs or other cell types may turn out to play the proposed role. To compare predictions and stimuli in a subtractive manner, the encoded prediction/stimulus needs to be translated into a direct variable code. In this framework, we are assuming that this can be achieved by the weight matrix defining the synaptic connections from the neural populations representing predictions and stimuli (possibly in a population code).

To conclude, we proposed that prediction error activity in layer 2/3 circuits is modulated by uncertainty and that the diversity of cell types in these circuits achieves the appropriate scaling of the prediction error activity. The proposed model is in line with Bayes-optimal behaviour, and makes predictions for future experiments.

Methods

Derivation of the UPE

The goal is to learn $\hat{\mu}$ to minimise the negative log likelihood:

$$-\log L = -\log p(\mathbf{s}|\hat{\mu},\sigma) \tag{3}$$

$$= -\log \prod_{n=1}^{N} \mathcal{N}(s_n|\hat{\mu},\sigma) \tag{4}$$

$$= \frac{1}{2\sigma^2}\sum_{n=1}^{N}(s_n-\hat{\mu})^2 + \frac{N}{2}\log\sigma^2 + \frac{N}{2}\log(2\pi) \tag{5}$$

We consider the negative log likelihood for one sample s of the stimulus distribution:

$$-\log p(s|\hat{\mu},\sigma) = \frac{1}{2\sigma^2}(s-\hat{\mu})^2 + \frac{1}{2}\log\sigma^2 + \frac{1}{2}\log(2\pi) \tag{6}$$

Stochastic gradient descent on the negative log likelihood gives the update for $\hat{\mu}$:

$$\Delta\hat{\mu} \propto -\frac{\partial}{\partial\hat{\mu}}(-\log p(s|\hat{\mu},\sigma)) \tag{7}$$

$$= -\frac{\partial}{\partial\hat{\mu}}\left(\frac{1}{2\sigma^2}(s-\hat{\mu})^2 + \frac{1}{2}\log\sigma^2 + \frac{1}{2}\log(2\pi)\right) \tag{8}$$

$$= \frac{1}{\sigma^2}(s-\hat{\mu}). \tag{9}$$

Circuit model

Prediction error circuit We modelled a circuit consisting of excitatory prediction error neurons in layer 2/3, and two inhibitory populations, corresponding to PV, and SST interneurons.
Layer 2/3 pyramidal cells receive divisive inhibition from PVs (Wilson *et al.*, 2012). We, hence, modelled the activity of prediction error neurons as

$$\tau_{\mathrm{E}}\frac{dr_{\mathrm{UPE}}}{dt} = -r_{\mathrm{UPE}} + \phi\left(\frac{I_{\mathrm{dend}}}{I_0 + w_{\mathrm{UPE,PV}}\, r_{\mathrm{PV}}}\right), \tag{10}$$

where $\phi(x)$ is the activation function, defined in Eq. 21, $I_{\mathrm{dend}} = \lfloor w_{\mathrm{UPE,s}}\, r_{\mathrm{s}} - w_{\mathrm{UPE,SST}}\, r_{\mathrm{SST}}\rfloor^k$ is the dendritic input current to the positive prediction error neuron (see section Neuronal dynamics below for r_x and for the negative prediction error neuron, and table for w_x). The nonlinearity in the dendrite is determined by the exponent k, which is by default $k = 2$, unless otherwise specified as in Fig. 4G-J. $I_0 > 1$ is a constant ensuring that the divisive inhibition does not become excitatory, when $\sigma < 1.0$.

The PV firing rate is determined by the input from the sound representation ($w_{\mathrm{PV^+,a}}\, r_{\mathrm{a}}$) and the whisker stimuli, from which their mean is subtracted ($w_{\mathrm{PV^+,s}}\, r_{\mathrm{s}} - w_{\mathrm{PV^+,SST^+}}\, r_{\mathrm{SST^+}}$, where the mean is given by $r_{\mathrm{SST^+}}$). The mean-subtracted whisker stimuli serve as a target for learning the weight from the sound representation to the PV $w_{\mathrm{PV^+,a}}$. The PV firing rate evoles over time according to:

$$\tau_{\mathrm{I}}\frac{dr_{\mathrm{PV^+}}}{dt} = -r_{\mathrm{PV^+}} + \phi_{\mathrm{PV}}((1-\beta)w_{\mathrm{PV^+,a}}\, r_{\mathrm{a}} + \beta(w_{\mathrm{PV^+,s}}\, r_{\mathrm{s}} - w_{\mathrm{PV^+,SST^+}}\, r_{\mathrm{SST^+}}))) \tag{11}$$

where $\phi_{\mathrm{PV}}(x)$ is a rectified square activation function, defined in Eq. 22.

In the positive prediction error circuit, in which the SSTs learn to represent the mean, the SST activity is determined by

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{SST}^+}}{dt} = -r_{\mathrm{SST}^+} + \phi((1-\beta) w_{\mathrm{SST}^+,\mathrm{a}}\, r_{\mathrm{a}} + \beta s). \tag{12}$$

.

Recurrent circuit model In the recurrent circuit, shown in Fig. 5, we added an internal representation neuron to the circuit with firing rate r_{R}. In this circuit the SSTs inherit the mean representation from the representation neuron instead of learning it themselves. In this recurrent circuit, the firing rate of each population r_i where $i \in [\mathrm{SST}^+, \mathrm{SST}^-, \mathrm{PV}^+, \mathrm{PV}^-, \mathrm{UPE}^+, \mathrm{UPE}^-, \mathrm{R}]$ evolves over time according to the following neuronal dynamics. ϕ denotes a rectified linear activation function with saturation, ϕ_{PV} denotes a rectified square activation function with saturation, defined in the section below.

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{SST}^+}}{dt} = -r_{\mathrm{SST}^+} + \phi(w_{\mathrm{SST}^+,\mathrm{R}}\, r_{\mathrm{R}}), \tag{13}$$

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{PV}^+}}{dt} = -r_{\mathrm{PV}^+} + \phi_{\mathrm{PV}}((1-\beta) w_{\mathrm{PV}^+,\mathrm{a}}\, r_{\mathrm{a}} + \beta(w_{\mathrm{PV}^+,\mathrm{s}}\, r_{\mathrm{s}} - w_{\mathrm{PV}^+,\mathrm{SST}^+}\, r_{\mathrm{SST}^+}))), \tag{14}$$

$$\tau_{\mathrm{E}} \frac{dr_{\mathrm{UPE}^+}}{dt} = -r_{\mathrm{UPE}^+} + \phi\left(\frac{\lfloor w_{\mathrm{UPE,s}}\, r_{\mathrm{s}} - w_{\mathrm{UPE,SST}^+}\, r_{\mathrm{SST}^+} \rfloor^k}{I_0 + w_{\mathrm{UPE,PV}^+}\, r_{\mathrm{PV}^+}}\right), \tag{15}$$

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{SST}^-}}{dt} = -r_{\mathrm{SST}^-} + \phi(w_{\mathrm{SST}^-,\mathrm{s}}\, r_{\mathrm{s}}), \tag{16}$$

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{PV}^-}}{dt} = -r_{\mathrm{PV}^-} + \phi_{\mathrm{PV}}((1-\beta) w_{\mathrm{PV}^-,\mathrm{a}}\, r_{\mathrm{a}} + \beta(w_{\mathrm{PV}^+,\mathrm{R}}\, r_{\mathrm{R}} - w_{\mathrm{PV}^-,\mathrm{SST}^-}\, r_{\mathrm{SST}^-}))), \tag{17}$$

$$\tau_{\mathrm{E}} \frac{dr_{\mathrm{UPE}^-}}{dt} = -r_{\mathrm{UPE}^-} + \phi\left(\frac{\lfloor w_{\mathrm{UPE,R}}\, r_{\mathrm{R}} - w_{\mathrm{UPE,SST}^-}\, r_{\mathrm{SST}^-} \rfloor^k}{I_0 + w_{\mathrm{UPE,PV}^-}\, r_{\mathrm{PV}^-}}\right), \tag{18}$$

$$\tau_{\mathrm{E}} \frac{dr_{\mathrm{R}}}{dt} = -r_{\mathrm{R}} + \phi(w_{\mathrm{R,a}}\, r_{\mathrm{a}} + w_{\mathrm{R,UPE}^+}\, r_{\mathrm{UPE}^+} - w_{\mathrm{R,UPE}^-}\, r_{\mathrm{UPE}^-})). \tag{19}$$

$$\tag{20}$$

Activation functions

$$\phi(x) = \begin{cases} 0 & \text{if } x \le 0 \\ x & \text{if } 0 < x < x_{\max} \\ r_{\max} & \text{if } x \ge x_{\max} \end{cases} \tag{21}$$

and

$$\phi_{\mathrm{PV}}(x) = \begin{cases} 0 & \text{if } x \le 0 \\ x^2 & \text{if } 0 < x < x_{\max} \\ r_{\max} & \text{if } x \ge x_{\max} \end{cases} \tag{22}$$

Inputs The inputs to the circuit were the higher level representation of the sound a, which was either on (1.0) or off (0.0), and N samples from the Gaussian distribution of whisker stimulus intensities. Each whisker stimulus intensity was presented for D timesteps (see Table).

Synaptic dynamics / Plasticity rules Synapses from the higher level representation of the sound a to the SSTs, PVs, and to R were plastic according to the following activity-dependent plasticity rules (Urbanczik and Senn, 2014).

$$\Delta w_{\mathrm{SST,a}} = \eta_{\mathrm{SST}}(r_{\mathrm{SST}} - \phi(w_{\mathrm{SST,a}} r_{\mathrm{a}})) r_{\mathrm{a}}, \tag{23}$$

$$\Delta w_{\mathrm{PV,a}} = \eta_{\mathrm{PV}}(r_{\mathrm{PV}} - \phi_{\mathrm{PV}}(w_{\mathrm{PV,a}} r_{\mathrm{a}})) r_{\mathrm{a}}, \tag{24}$$

$$\Delta w_{\mathrm{R,a}} = \eta_{\mathrm{R}}(r_{\mathrm{R}} - \phi(w_{\mathrm{R,a}} r_{\mathrm{a}})) r_{\mathrm{a}}, \tag{25}$$

$$\tag{26}$$

where $\eta_{\mathrm{PV}} = 0.01\eta_R$.

Parameter	Value	Description
$w_{\mathrm{PV}^+,\mathrm{SST}^+}$	$\sqrt{\frac{2-\beta}{\beta}}$	weight from SST^+ to PV^+
$w_{\mathrm{PV}^+,\mathrm{s}}$	$\sqrt{\frac{2-\beta}{\beta}}$	weight from s to PV^+
$w_{\mathrm{PV}^-,\mathrm{SST}^-}$	$\sqrt{\frac{2-\beta}{\beta}}$	weight from SST^- to PV^-
$w_{\mathrm{PV}^-,\mathrm{R}}$	$\sqrt{\frac{2-\beta}{\beta}}$	weight from R to PV^-
$w_{\mathrm{SST}^+,\mathrm{R}}$	1.0	weight from R to SST^+
$w_{\mathrm{SST}^-,\mathrm{s}}$	1.0	weight from s to SST^-
$w_{\mathrm{UPE}^+,\mathrm{SST}^+}$	1.0	weight from SST^+ to UPE^+
$w_{\mathrm{UPE}^+,\mathrm{s}}$	1.0	weight from s to UPE^+
$w_{\mathrm{UPE}^-,\mathrm{R}}$	1.0	weight from R to UPE^-
$w_{\mathrm{UPE}^-,\mathrm{SST}^-}$	1.0	weight from SST^- to UPE^-
$w_{\mathrm{R},\mathrm{UPE}^+}$	0.1(1.0)	weight from UPE^+ to R (Fig. 6)
$w_{\mathrm{R},\mathrm{UPE}^-}$	0.1(1.0)	weight from UPE^- to R (Fig. 6)
$x_{\max}$	20	limits neuronal activity
β	0.1	nudging parameter

Table 1: **Parameters of the network.**

Parameter	Value	Description
a	$\{0.0, 1.0\}$	auditory stimulus (on/off)
s	$\sim \mathcal{N}(\mu, \sigma)$	somatosensory (whisker) stimulus
N	200000	number of whisker stimulus samples
D	$\{10, 100\}$	stimulus duration (Figs. 1-5,7; Fig. 7)

Table 2: **Inputs.**

Explanation of the synaptic dynamics The connection weight from the sound representation to the SSTs $w_{\mathrm{SST,a}}$ is plastic according to the following local activity-dependent plasticity rule (Urbanczik and Senn, 2014):

$$\Delta w_{\mathrm{SST,a}} = \eta(r_{\mathrm{SST}} - \phi(w_{\mathrm{SST,a}}\, r_{\mathrm{a}}))\, r_{\mathrm{a}}, \tag{27}$$

where η is the learning rate, r_{a} is the pre-synaptic firing rate, r_{SST} is the post-synaptic firing rate of the SSTs, $\phi(x)$ is a rectified linear activation function of the SSTs, and the SST activity is determined by

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{SST}}}{dt} = -r_{\mathrm{SST}} + \phi((1-\beta) w_{\mathrm{SST,a}}\, r_{\mathrm{a}} + \beta s). \tag{28}$$

The SST activity is influenced (nudged with a factor β) by the somatosensory stimuli s, which provide targets for the desired SST activity. The learning rule ensures that the auditory input alone causes SSTs to fire at their target activity. As in the original proposal (Urbanczik and Senn, 2014), the terms in the learning rule can be mapped to local neuronal variables, which could be represented by dendritic ($w_{\mathrm{SST,a}}\, r_{\mathrm{a}}$) and somatic ($r_{\mathrm{SST}}$) activity.

The connection weight from the sound representation to the PVs $w_{\mathrm{PV,a}}$ is plastic according to the same local activity-dependent plasticity rule as the SSTs (Urbanczik and Senn, 2014):

$$\Delta w_{\mathrm{PV,a}} = \eta(r_{\mathrm{PV}} - \phi_{\mathrm{PV}}(w_{\mathrm{PV,a}}\, r_{\mathrm{a}}))\, r_{\mathrm{a}}. \tag{29}$$

The weight from the sound representation to the PV $w_{\mathrm{PV}^+,\mathrm{a}}$ approaches σ (instead of μ as the weight to the SSTs), because the PV activity is a function of the mean-subtracted whisker stimuli (instead of the whisker stimuli as the SST activity), and for a Gaussian-distributed stimulus $s \sim \mathcal{N}(s|\mu, \sigma)$, it holds that $\mathbb{E}(\lfloor s - \mu \rfloor) \propto \sigma$.

Estimating the variance correctly The PVs estimate the variance of the sensory input from the variance of the teaching input $(s - \mu)$, which nudges the membrane potential of the PVs with a nudging factor β. The nudging factor reduces the effective variance of the teaching input, such that in order to correctly estimate the variance, this reduction needs to be compensated by larger weights from the SSTs to the PVs ($w_{\mathrm{PV,SST}}$) and from the sensory input to the PVs

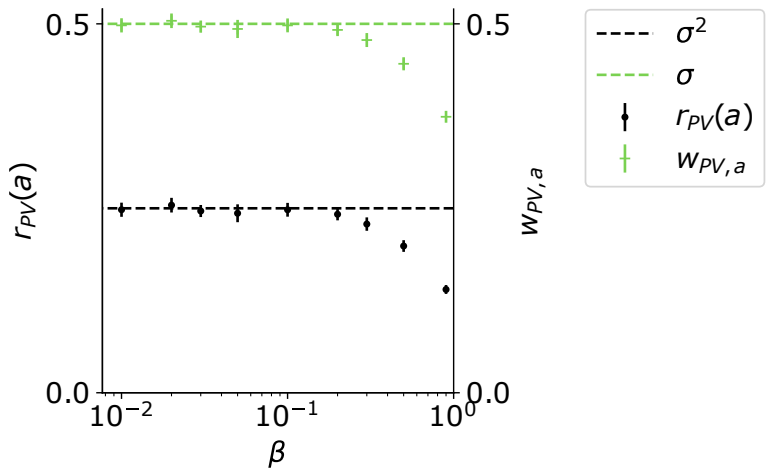

Figure 8: For small β, and $w_s = \sqrt{\frac{2-\beta}{\beta}}$, the weight from a to PV approaches σ and the PV firing rate approaches σ^2.

($w_{\mathrm{PV},s}$). To determine how strong the weights $w_s = w_{\mathrm{PV,SST}} = w_{\mathrm{PV},s}$ need to be to compensate for the downscaling of the input variance by β, we require that $\langle wa \rangle^2 = \sigma^2$ when the average weight change $\langle \Delta w \rangle = 0$. The learning rule for w is as follows:

$$\Delta w = \eta[r_{\mathrm{PV}} - \phi(wa)]a \tag{30}$$
$$= \eta[\phi((1-\beta)wa + \beta w_s \tilde{s}) - \phi(wa)]a \tag{31}$$
$$\tag{32}$$

where $r_{PV} = \phi((1-\beta)wa + \beta w_s \tilde{s})$ and $\tilde{s} = (s - \mu) \sim \mathcal{N}(0, \sigma)$.

Using that $\phi(u) = u^2$, the average weight change becomes:

$$\langle \Delta w \rangle = \langle (1-\beta)^2 w^2 a^2 + \beta^2 w_s^2 \tilde{s}^2 + 2(1-\beta) wa \beta w_s \tilde{s} - w^2 a^2 \rangle a \quad | \quad \langle \tilde{s} \rangle = 0 \tag{33}$$
$$= \langle -2\beta w^2 a^2 + \beta^2 w^2 a^2 + \beta^2 w_s^2 \tilde{s}^2 \rangle a \tag{34}$$
$$= \beta(\langle w^2 \rangle a^2 (\beta - 2) + \beta w_s^2 \sigma^2) \tag{35}$$
$$\tag{36}$$

Given our objective $\langle wa \rangle^2 = \sigma^2$, we can write:

$$\langle \Delta w \rangle = (\beta - 2 + \beta w_s^2) \tag{37}$$
$$\tag{38}$$

Then for $\langle \Delta w \rangle = 0$:

$$0 = (\beta - 2 + \beta w_s^2) \tag{39}$$
$$\Rightarrow w_s = \sqrt{\frac{2-\beta}{\beta}} \tag{40}$$

Here, we assumed that $\phi(u) = u^2$ instead of $\phi(u) = \lfloor u \rfloor^2$. To test how well this approximation holds, we simulated the circuit for different values of β and hence w_s, and plotted the PV firing rate $r_{\mathrm{PV}}(a)$ given the sound input a and the weight from a to PV, $w_{\mathrm{PV},a}$, for different values of β (Fig. 8). This analysis shows that the approximation holds for small β up to a value of $\beta = 0.2$.

Simulation We initialised the circuit with the initial weight configuration in Tables 1 and 3 and neural firing rates were initialised to be 0 ($r_i(0) = 0$ with $i \in [\mathrm{SST}^+, \mathrm{SST}^-, \mathrm{PV}^+, \mathrm{PV}^-, \mathrm{UPE}^+, \mathrm{UPE}^-, \mathrm{R}]$). We then paired a constant tone input (a1 or a2) with N samples from the whisker stimulus distribution, the parameters of which we varied and are indicated in each Figure. Each whisker stimulus intensity was presented for D timesteps (see Table 2). All simulations were written in Python. Differential equations were numerically integrated with a time step of $dt = 1$.

Parameter	Value	Description
η_{SST}	0.1	learning rate for $w_{\mathrm{SST}^{+/-},\mathrm{a}}$
η_{PV}	$0.01 * \eta_{\mathrm{R}} = 0.001$	learning rate for $w_{\mathrm{PV}^{+/-},\mathrm{a}}$
η_{R}	0.1	learning rate for $w_{\mathrm{R,a}}$
$w^{\mathrm{initial}}_{\mathrm{SST,a}}$	0.01	initial value for $w_{\mathrm{SST}^{+/-},\mathrm{a}}$
$w^{\mathrm{initial}}_{\mathrm{PV,a}}$	0.01	initial value for $w_{\mathrm{PV}^{+/-},\mathrm{a}}$
$w^{\mathrm{initial}}_{\mathrm{R,a}}$	0.01	initial value for $w_{\mathrm{R,a}}$

Table 3: **Parameters of the plasticity rules.**

Parameter	Value	Description
T	$N * D$	simulation time
dt	0.1	simulation time step
τ_E	1.0	excitatory membrane time constant
τ_I	1.0	inhibitory membrane time constant

Table 4: **Parameters of simulations in Figs. 2-5.**

Parameter	Value	Description
T	$N * D$	simulation time
dt	0.1	simulation time step
τ_E	10.0	excitatory membrane time constant
τ_I	2.0	inhibitory membrane time constant
η_R	0.01	learning rate of $w_{R,A}$

Table 5: **Parameters of the simulation in Fig. 6.**

Comparing the UPE circuit with an unmodulated circuit (Fig. 7) To ensure a fair comparison, the unmodulated control has an effective learning rate that is the mean of the two effective learning rates in the uncertainty-modulated case.

Supplementary Information

Supplementary Methods

Synaptic dynamics/plasticity rules

$$\Delta w_{\mathrm{PV}^+,\mathrm{SST}^+} = \eta_{\mathrm{PV}}(w_{\mathrm{PV}^+,s} r_{\mathrm{s}} - w_{\mathrm{PV}^+,\mathrm{SST}^+} r_{\mathrm{SST}^+}) r_{\mathrm{SST}^+}, \tag{41}$$

$$\Delta w_{\mathrm{PV}^-,\mathrm{SST}^-} = \eta_{\mathrm{PV}}(w_{\mathrm{PV}^-,R} r_{\mathrm{R}} - w_{\mathrm{PV}^-,\mathrm{SST}^-} r_{\mathrm{SST}^-}) r_{\mathrm{SST}^-}, \tag{42}$$

$$\Delta w_{\mathrm{UPE}^+,\mathrm{SST}^+} = \eta_{\mathrm{UPE}}(w_{\mathrm{UPE}^+,r_{\mathrm{s}}} r_{\mathrm{s}} - w_{\mathrm{UPE}^+,\mathrm{SST}^+} r_{\mathrm{SST}^+}) r_{\mathrm{SST}^+}, \tag{43}$$

$$\Delta w_{\mathrm{UPE}^+,\mathrm{SST}^-} = \eta_{\mathrm{UPE}}(w_{\mathrm{UPE}^-,\mathrm{R}} r_{\mathrm{R}} - w_{\mathrm{UPE}^-,\mathrm{SST}^-} r_{\mathrm{SST}^-}) r_{\mathrm{SST}^-}, \tag{44}$$

$$\tag{45}$$

Different choice of supralinear activation function for PV

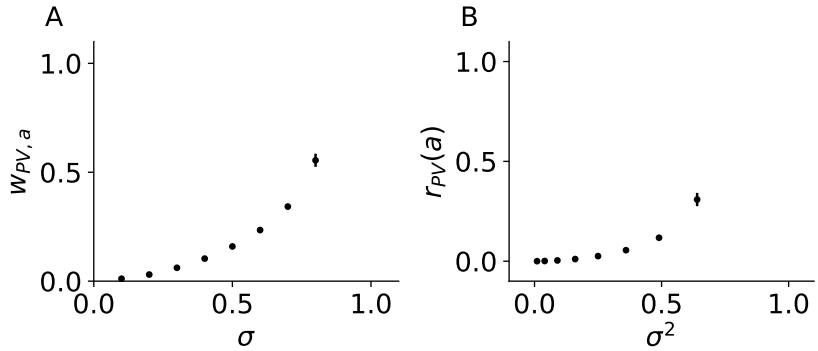

Figure 9: Learning the variance in the positive prediction error circuit with PVs with a power activation function (exponent = 3.0). The figure is analogous to Fig. 3G,H, and the circuit is the same except that the activation function of the PVs ($\phi_{PV}(x)$) has an exponent of 3.0 instead of 2.0.

Plastic weights from SST to PV learn to match weights from s to PV

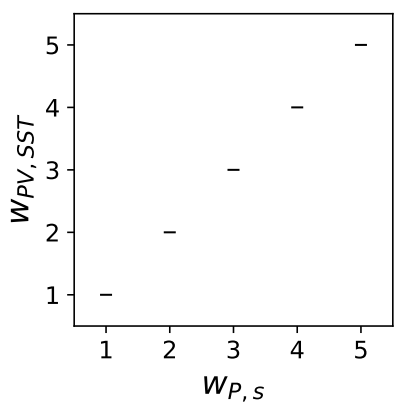

Figure 10: With inhibitory plasticity, weights from SST to PV can be learned. This figure shows that the weight from SST to PV ($w_{\mathrm{PV,SST}}$) is equal to the weight from s to PV ($w_{\mathrm{PV},s}$). The inhibitory plasticity rule is described in the Supplementary Methods.

PVs learn the variance in the negative prediction error circuit

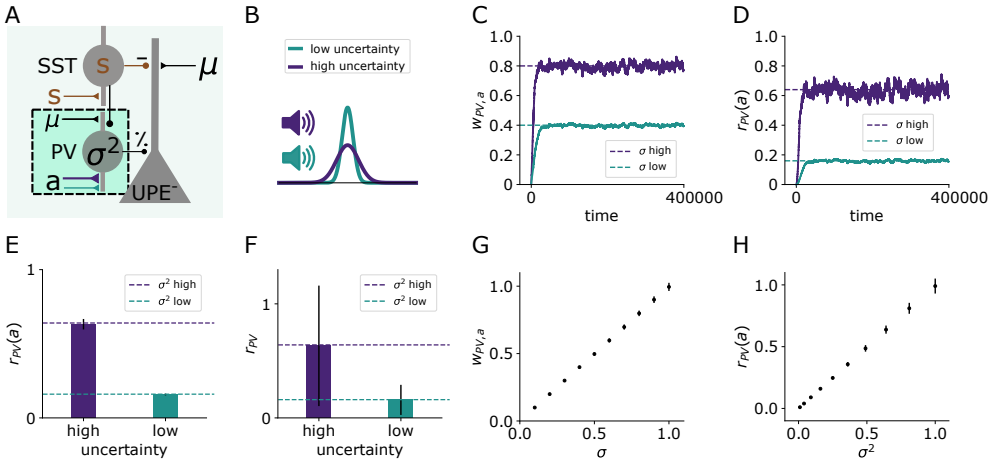

Figure 11: **PVs learn to represent the variance given an associative cue in the negative prediction error circuit.** A: Illustration of the changes in the negative prediction error circuit. Thicker lines denote stronger weights. B: Two different tones (purple, green) are associated with two somatosensory stimulus distributions with different variances (purple: high, green: low). C: Weights from sound a to PV over time for two different values of stimulus variance (high: $\sigma = 0.8$ (purple), low: $\sigma = 0.4$ (green)). D: PV firing rates over time given sound input (without stimulus input) for low (green) and high (purple) stimulus variance. E: PV firing rates (mean and std) given sound input for low and high stimulus variance. F: PV firing rates (mean and std) during sound and stimulus input. G: Weights from sound a to PV for different values of σ (mean and std from timestep 100000 to the end of the simulation). H: PV firing rates given sound input for different values of σ^2 (mean and std from timestep 100000 to the end of the simulation).

Learning SST to E weights

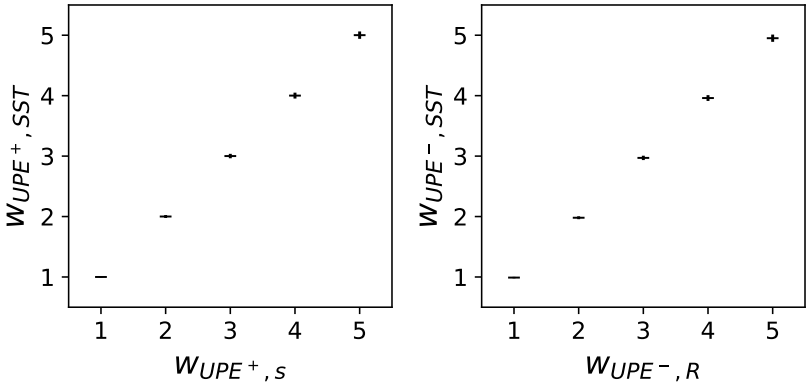

Figure 12: Learning the weights from the SSTs to the UPE neurons. This figure shows that the weights from the SSTs to the UPEs in both the positive (left) and the negative (right) prediction error circuit can be learned with inhibitory plasticity to match the weights from the stimulus representation s to the UPEs. The inhibitory plasticity rule is described in the supplementary methods.

PV activity is proportional to the variance in the recurrent circuit

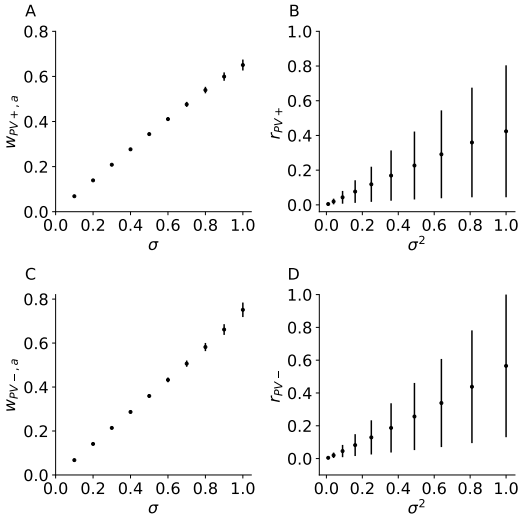

Figure 13: PV firing rates are proportional to the variance in the recurrent circuit model. Weights from a to PV as a function of σ in the positive (A) and negative (C) prediction error subcircuit. PV firing rates as a function of σ^2 in the positive (B) and negative (D) prediction error circuit.

Perturbation experiments

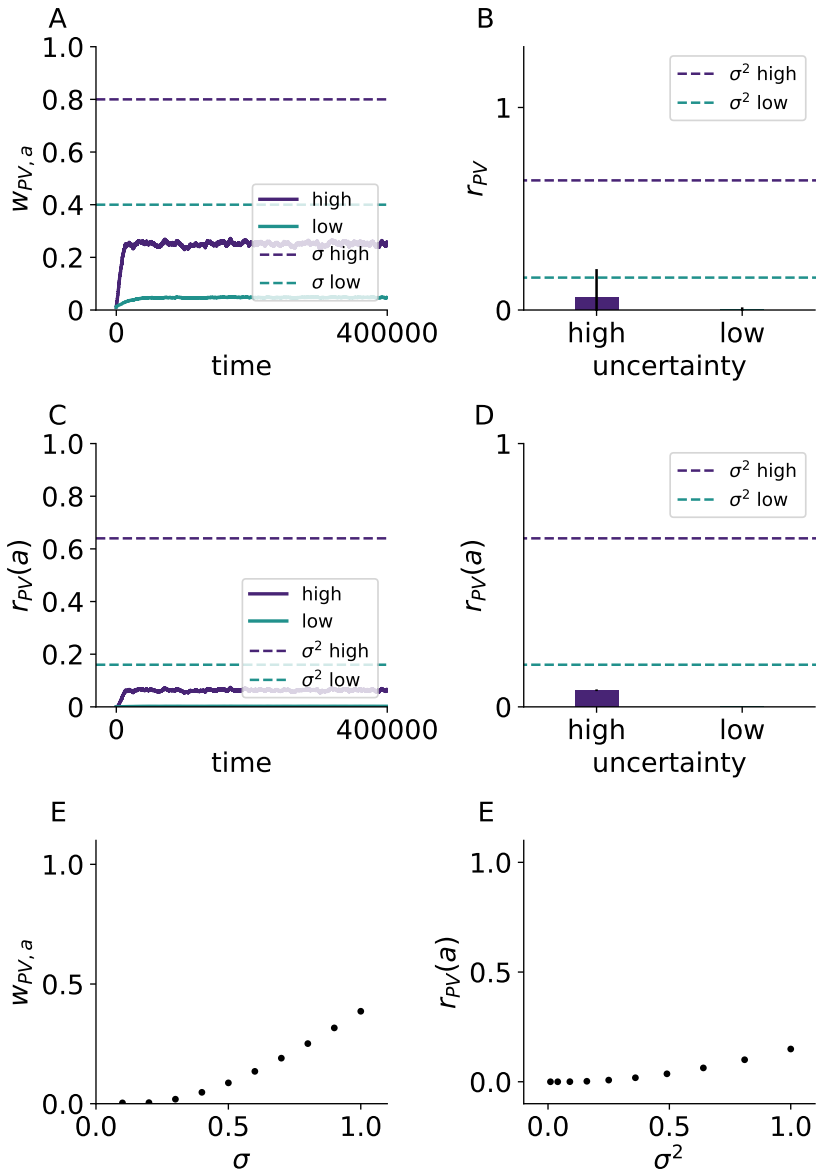

Figure 14: The panels are analogous to the panels in Fig. 3. Here, we used the same positive prediction error circuit as inf Fig. 3, and perturbed the SS-to-PV weight in the positive prediction error circuit by +10%.

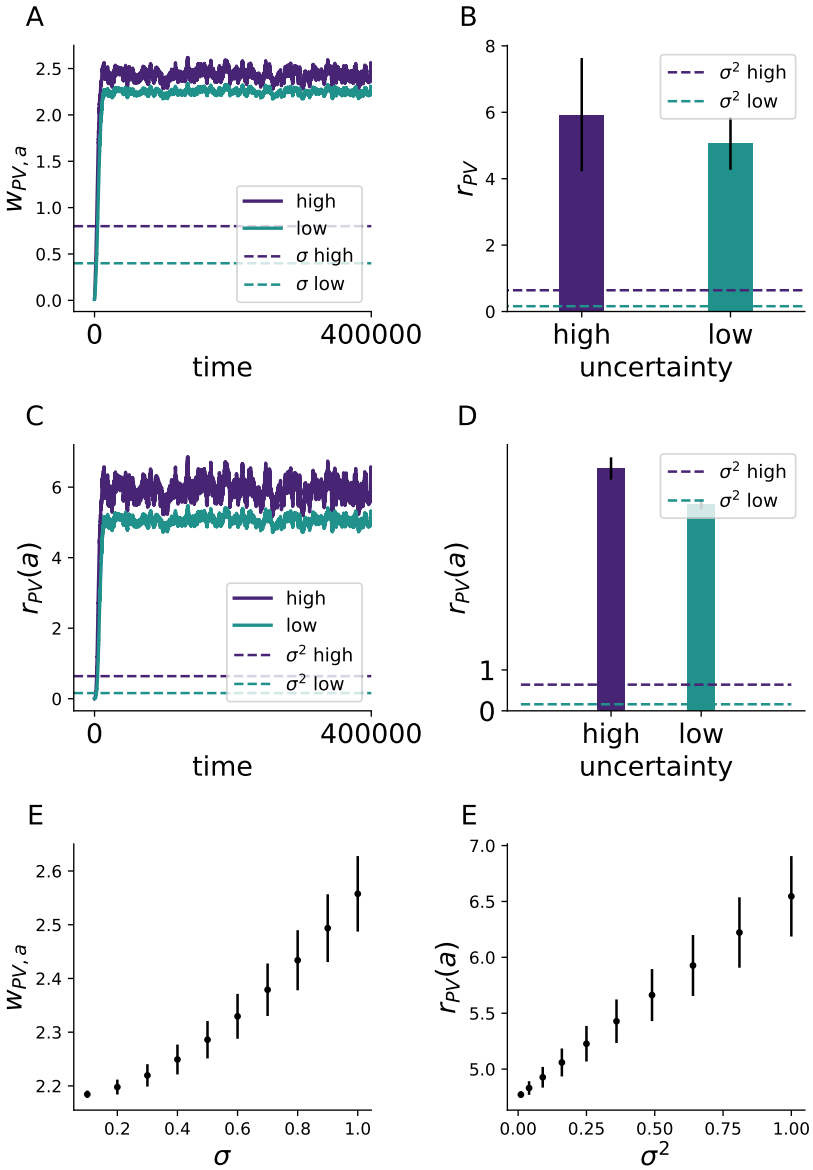

Figure 15: The panels are analogous to the panels in Fig. 3. Here, we used the same positive prediction error circuit as inf Fig. 3, and perturbed the SS-to-PV weight in the positive prediction error circuit by -10%.

.

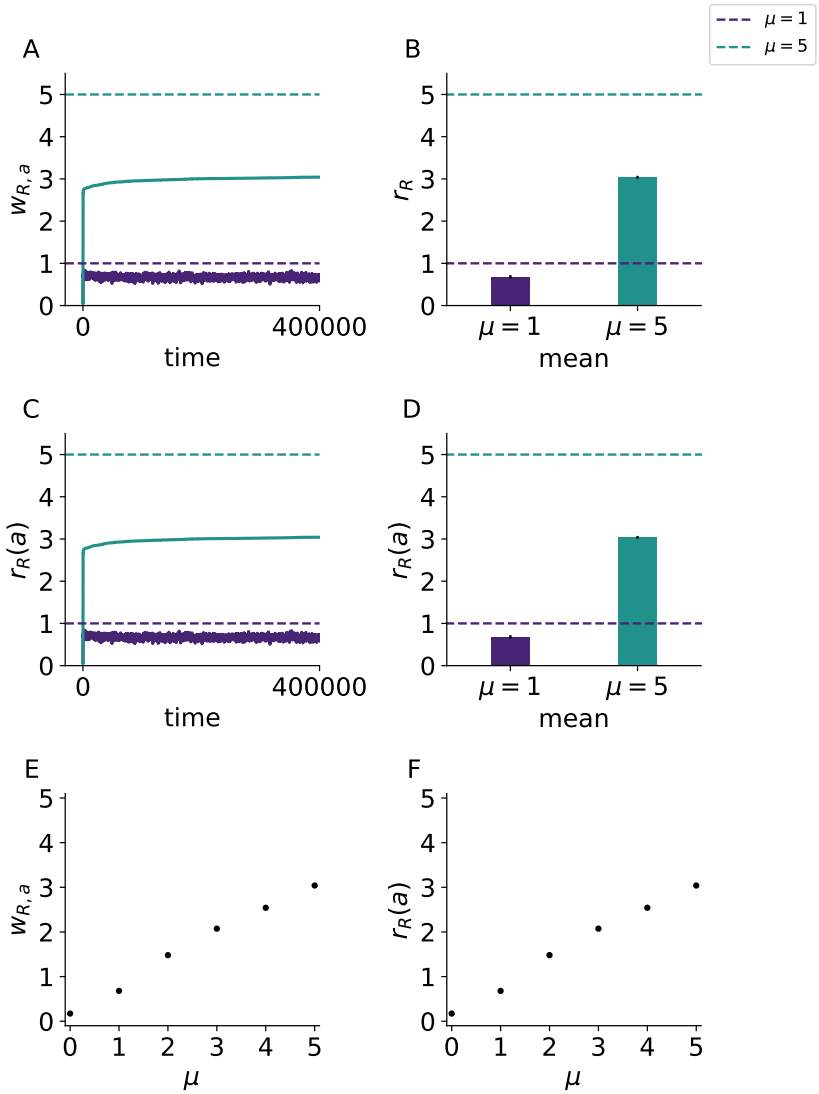

Figure 16: **Perturbation of SST$^+$ to UPE$^+$ and from SST$^-$ to UPE$^-$ weights by +100%.** Here we used the recurrent circuit model from Fig. 5 and perturbed the weights from SST$^+$ to UPE$^+$ and SST$^-$ to UPE$^-$ by +100%. Two different tones (purple, green) are associated with two somatosensory stimulus distributions with different means (purple: small, green: large). A: Weights from sound a to R over time for two different values of stimulus mean (small: $\mu = 1.0$ (purple), large: $\mu = 5.0$ (green)). B: R firing rates (mean and std) given sound input and whisker input for small and large stimulus means. C: R firing rates over time given sound input (without whisker stimulus input) for small (purple) and large (green) stimulus means. D: R firing rates (mean and std) given sound input (without whisker stimulus input). E: Weights from sound a to R for different values of μ (mean and std from timestep 100000 to the end of the simulation). H: R firing rates given sound input (without whisker stimulus input) for different values of μ^2 (mean and std from timestep 100000 to the end of the simulation).

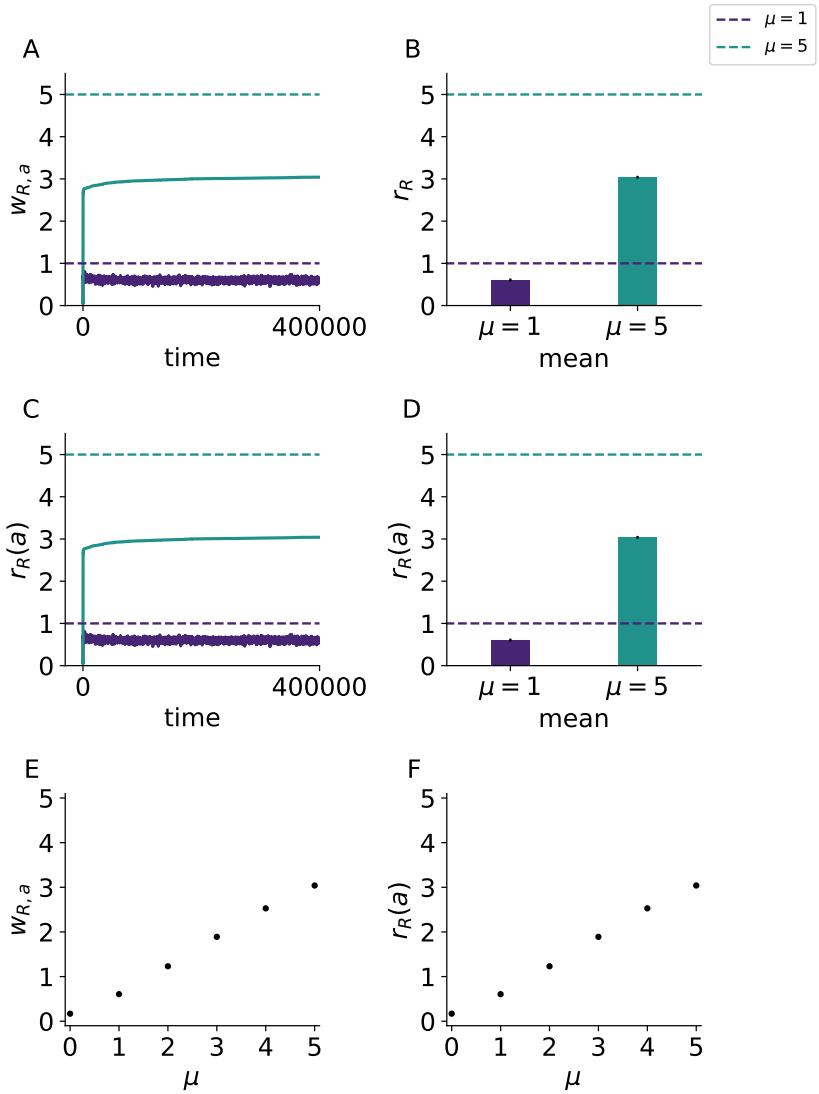

Figure 17: **Perturbation of SST$^+$ to UPE$^+$ weights by +100%.** Here we used the recurrent circuit model from Fig. 5 and perturbed the weights from SST$^+$ to UPE$^+$ by +100%. Two different tones (purple, green) are associated with two somatosensory stimulus distributions with different means (purple: small, green: large). A: Weights from sound a to R over time for two different values of stimulus mean (small: $\mu = 1.0$ (purple), large: $\mu = 5.0$ (green)). B: R firing rates (mean and std) given sound input and whisker input for small and large stimulus means. C: R firing rates over time given sound input (without whisker stimulus input) for small (purple) and large (green) stimulus means. D: R firing rates (mean and std) given sound input (without whisker stimulus input). E: Weights from sound a to R for different values of μ (mean and std from timestep 100000 to the end of the simulation). H: R firing rates given sound input (without whisker stimulus input) for different values of μ^2 (mean and std from timestep 100000 to the end of the simulation).

.

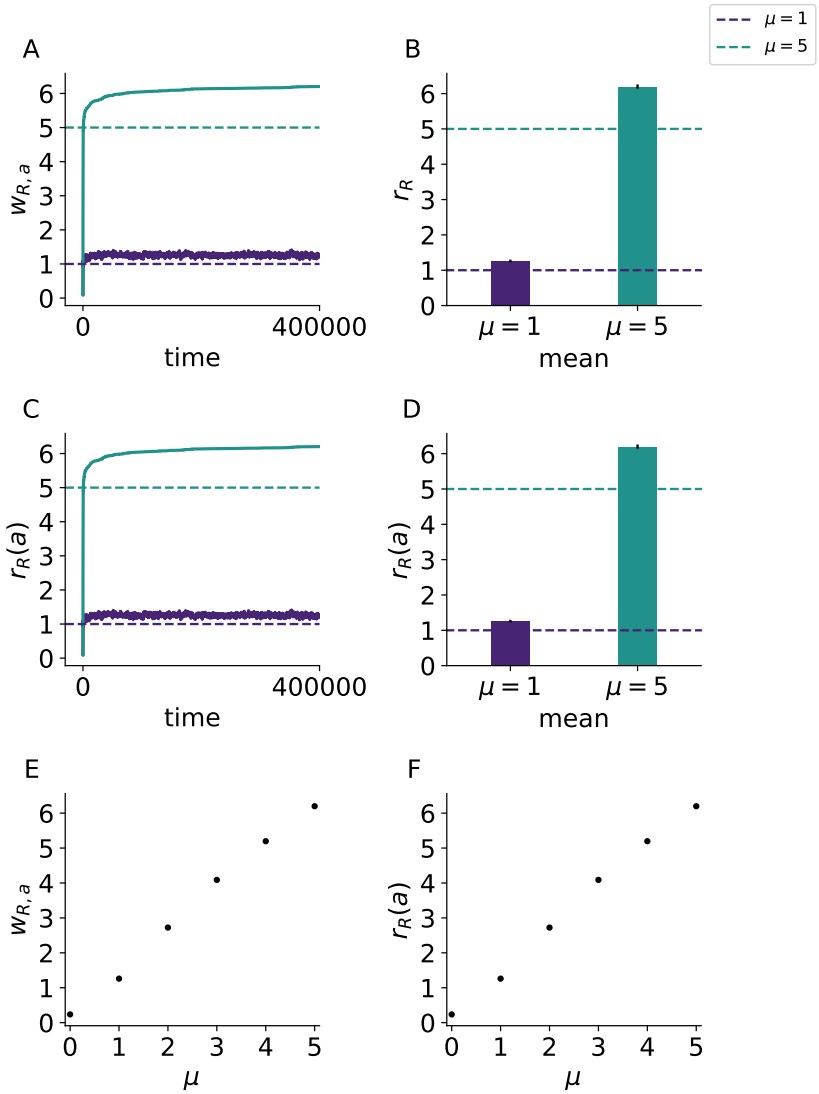

Figure 18: **Perturbation of SST$^-$ to UPE$^-$ weights by +100%.** Here we used the recurrent circuit model from Fig. 5 and perturbed the weights from SST$^-$ to UPE$^-$ weights by +100%. Two different tones (purple, green) are associated with two somatosensory stimulus distributions with different means (purple: small, green: large). A: Weights from sound a to R over time for two different values of stimulus mean (small: $\mu = 1.0$ (purple), large: $\mu = 5.0$ (green)). B: R firing rates (mean and std) given sound input and whisker input for small and large stimulus means. C: R firing rates over time given sound input (without whisker stimulus input) for small (purple) and large (green) stimulus means. D: R firing rates (mean and std) given sound input (without whisker stimulus input). E: Weights from sound a to R for different values of μ (mean and std from timestep 100000 to the end of the simulation). H: R firing rates given sound input (without whisker stimulus input) for different values of μ^2 (mean and std from timestep 100000 to the end of the simulation).

.

References

1. A. Attinger, B. Wang, G. B. Keller, *Cell* **169**, 1291–1302.e14, (https://doi.org/10.1016/j.cell.2017.05.023) (2017).
2. A. Ayaz *et al.*, *Nature Communications* **10**, 2585, (https://doi.org/10.1038/s41467-019-10564-8) (2019).
3. K. Banaie Boroujeni, P. Tiesinga, T. Womelsdorf, *eLife* **10**, ed. by S. Haegens, M. J. Frank, e69111, (https://doi.org/10.7554/eLife.69111) (2021).
4. P. Berkes, G. Orbán, M. Lengyel, J. Fiser, *Science* **331**, 83–87, eprint: https://www.science.org/doi/pdf/10.1126/science.1195870, (https://www.science.org/doi/abs/10.1126/science.1195870) (2011).
5. L. Buesing, J. Bill, B. Nessler, W. Maass, eng, *PLoS Comput Biol* **7**, e1002211 (2011).
6. J. Cannon, A. M. O'Brien, L. Bungert, P. Sinha, eng, *Autism Res* **14** (2021).
7. M. X. Cohen, **Wilmes, K.**, I. v. d. Vijver, eng, *Trends Cogn Sci* **15**, 558–566 (2011).
8. J. H. Cornford *et al.*, *eLife* **8**, ed. by M. Bartos, G. L. Westbrook, C.-C. Lien, J. C. Poncer, e49872, (https://doi.org/10.7554/eLife.49872) (2019).
9. G. P. Dehaene, R. Coen-Cagli, A. Pouget, *PLOS Computational Biology* **17**, 1–30, (https://doi.org/10.1371/journal.pcbi.1008138) (Feb. 2021).
10. C. Dorsett, B. D. Philpot, S. L. Smith, I. T. Smith, eng, *eNeuro* **8** (2021).
11. S. J. Eliades, X. Wang, *Nature* **453**, 1102–1106, (https://doi.org/10.1038/nature06910) (2008).
12. M. O. Ernst, M. S. Banks, *Nature* **415**, 429–433, (https://doi.org/10.1038/415429a) (2002).
13. C. R. Fetsch, A. Pouget, G. C. DeAngelis, D. E. Angelaki, *Nature Neuroscience* **15**, 146–154, (https://doi.org/10.1038/nn.2983) (2012).
14. B. J. Fischer, J. Peña, *Nature Neuroscience* **14**, 1061–1066, (https://doi.org/10.1038/nn.2872) (2011).
15. A. Fiser *et al.*, *Nature Neuroscience* **19**, 1658 EP, (https://doi.org/10.1038/nn.4385) (Sept. 2016).
16. W Gaetz *et al.*, eng, *Neuroimage* **86**, 1–9 (2014).
17. C. J. Gillon *et al.*, *bioRxiv*, eprint: https://www.biorxiv.org/content/early/2021/01/16/2021.01.15.426915.full.pdf, (https://www.biorxiv.org/content/early/2021/01/16/2021.01.15.426915) (2021).
18. J. Goris *et al.*, *Autism* **25**, PMID: 33030041, 440–451, eprint: https://doi.org/10.1177/1362361320962237, (https://doi.org/10.1177/1362361320962237) (2021).
19. J. Goris *et al.*, eng, *Biol Psychiatry Cogn Neurosci Neuroimaging* **3**, 667–674 (2018).
20. M. Harada *et al.*, eng, *J Autism Dev Disord* **41**, 447–454 (2011).
21. L. Hertäg, C. Clopath, eng, *Proc Natl Acad Sci U S A* **119**, e2115699119 (2022).
22. L. Hertäg, H. Sprekeler, *eLife* **9**, ed. by S. Ostojic, R. B. Ivry, S. Ostojic, e57541, (https://doi.org/10.7554/eLife.57541) (2020).
23. J. Jordan, J. Sacramento, W. A. M. Wybo, M. A. Petrovici, W. Senn, *Learning Bayes-optimal dendritic opinion pooling*, 2022, arXiv: 2104.13238 [q-bio.NC].
24. R. Jordan, G. B. Keller, *Neuron* **108**, 1194–1206.e5, (https://www.sciencedirect.com/science/article/pii/S0896627320307480) (2020).
25. P. Juarez, V. Martínez Cerdeño, eng, *Front Psychiatry* **13**, 913550 (2022).
26. T. Karaminis *et al.*, eng, *Sci Rep* **6**, 28570 (2016).
27. G. B. Keller, T. Bonhoeffer, M. Hübener, *Neuron* **74**, 809 –815, (http://www.sciencedirect.com/science/article/pii/S0896627312003844) (2012).
28. G. B. Keller, R. H. R. Hahnloser, *Nature* **457**, 187–190, (https://doi.org/10.1038/nature07467) (2009).
29. G. B. Keller, T. D. Mrsic-Flogel, *Neuron* **100**, 424 –435 (2018).
30. K. Kveraga, A. S. Ghuman, M. Bar, *Brain and cognition* **65**, 145–168, (https://pubmed.ncbi.nlm.nih.gov/17923222) (Nov. 2007).
31. R. P. Lawson, C. Mathys, G. Rees, *Nature Neuroscience* **20**, 1293–1299, (https://doi.org/10.1038/nn.4615) (2017).
32. R. P. Lawson, G. Rees, K. J. Friston, *Frontiers in human neuroscience* **8**, 302–302, (https://pubmed.ncbi.nlm.nih.gov/24860482) (May 2014).

33. S.-H. Lee *et al.*, *Nature* **488**, 379–383, (https://doi.org/10.1038/nature11312) (2012).

34. S. Lee, I. Kruglikov, Z. J. Huang, G. Fishell, B. Rudy, *Nature Neuroscience* **16**, 1662–1670 (2013).

35. I. Lieder *et al.*, *Nature Neuroscience* **22**, 256–264, (https://doi.org/10.1038/s41593-018-0308-9) (2019).

36. W. J. Ma, J. M. Beck, P. E. Latham, A. Pouget, *Nature Neuroscience* **9**, 1432–1438, (https://doi.org/10.1038/nn1790) (2006).

37. C. M. Niell, M. P. Stryker, *Journal of Neuroscience* **28**, 7520–7536 (2008).

38. E. Payzan-LeNestour, P. Bossaerts, *PLOS Computational Biology* **7**, 1–14, (https://doi.org/10.1371/journal.pcbi.1001048) (Jan. 2011).

39. M. A. Petrovici, J. Bill, I. Bytschok, J. Schemmel, K. Meier, *Stochastic inference with deterministic spiking neurons*, 2013, arXiv: 1311.3211 [q-bio.NC].

40. H.-J. Pi *et al.*, *Nature* **503**, 521–524 (2013).

41. A. Prönneke *et al.*, *Cerebral Cortex* **25**, 4854–4868 (2015).

42. R. P. N. Rao, D. H. Ballard, *Nature Neuroscience* **2**, 79–87, (https://doi.org/10.1038/4580) (1999).

43. J. L. R. Rubenstein, M. M. Merzenich, eng, *Genes, brain, and behavior* **2**, 255–267, (https://pubmed.ncbi.nlm.nih.gov/14606691https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6748642/) (2003).

44. S. Sachidhanandam, B. S. Sermet, C. C. Petersen, *Cell Reports* **15**, 700–706, (https://www.sciencedirect.com/science/article/pii/S2211124716303345) (2016).

45. B. A. Seybold, E. A. Phillips, C. E. Schreiner, A. R. Hasenstaub, *Neuron* **87**, 1181–1192, (https://www.sciencedirect.com/science/article/pii/S0896627315007709) (2015).

46. Z. Shi *et al.*, *bioRxiv*, (https://www.biorxiv.org/content/early/2022/01/25/2022.01.21.477218) (2022).

47. N. Ueda *et al.*, *Schizophrenia Research: Cognition* **28**, 100229, (https://www.sciencedirect.com/science/article/pii/S2215001321000366) (2022).

48. R. Urbanczik, W. Senn, *Neuron* **81**, 521–528, (https://www.sciencedirect.com/science/article/pii/S0896627313011276) (2014).

49. S. Van de Cruys *et al.*, eng, *Psychol Rev* **121**, 649–675 (2014).

50. A. R. Walker, D. Luque, M. E. Le Pelley, T. Beesley, *Psychonomic Bulletin & Review* **26**, 1911–1916, (https://doi.org/10.3758/s13423-019-01653-2) (2019).

51. N. R. Wilson, C. A. Runyan, F. L. Wang, M. Sur, eng, *Nature* **488**, 343–348 (2012).

52. P. Zmarz, G. B. Keller, eng, *Neuron* **92**, 766–772 (2016).

Acknowledgements

We would like to thank Loreen Hertäg and Sadra Sadeh for feedback on the manuscript. This work has received funding from the European Union 7th Framework Programme under grant agreement 604102 (HBP), the Horizon 2020 Framework Programme under grant agreements 720270, 785907 and 945539 (HBP) and the Manfred Stärk Foundation.