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

December 10, 2024

Abstract

Understanding the variability of the environment is essential to function in everyday life. The brain must hence take uncertainty into account when updating its internal model of the world. The basis for updating the model are prediction errors that arise from a difference between the current model and new sensory experiences. Although prediction error neurons have been identified in layer 2/3 of diverse brain areas, how uncertainty modulates these errors and hence learning is, however, unclear. Here, we use a normative approach to derive how uncertainty should modulate prediction errors and postulate that layer 2/3 neurons represent uncertainty-modulated prediction errors (UPE). We further hypothesise that the layer 2/3 circuit calculates the UPE through the subtractive and divisive inhibition by different inhibitory cell types. By implementing the calculation of UPEs in a microcircuit model, we show that different cell types 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, the mechanism enables an organism to optimise its learning strategy via 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 [48, 11]. 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 [58]. 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 [58, 76, 28, 10, 46]. 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, [42, 82, 22, 4, 26], auditory [17, 43], somatosensory [5], and posterior parietal [65]). Importantly, prediction errors could be associated with learning [40]. Prediction error neurons are embedded in neural circuits that 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 [34, 32], and that the prediction is subtracted from the sensory input [see e.g. 66, 4], possibly mediated by so-called somatostatin-positive interneurons (SSTs) [4]. How uncertainty is influencing these computations has not yet been investigated. Prediction error neurons receive inputs from a diversity of inhibitory cell types, 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, first, layer 2/3 prediction error neurons reflect such UPEs. Second, we hypothesise that different inhibitory cell types are involved in calculating the difference between predictions and stimuli, and in the uncertainty modulation, respectively. Based on experimental findings, we suggest that SSTs and PVs play the respective roles. We then derive biologically plausible plasticity rules that enable those cell types 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 suggest 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. Consider a situation in which a person waits for a bus to arrive. If they learned that the bus is not reliable, another late arrival of the bus does not surprise them and does not change their model of the bus (Fig. 1A). If, on the contrary, they learned that the kind of stimulus or context is not very variable (low uncertainty), a prediction error should have a larger impact on their model. For example, if they learned that buses are reliable, they will notice that the bus is late and may use this information to update their model of the bus (Fig. 1A). 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. Here we take the view that uncertainty is computed and represented on each level of the cortical hierarchy, from early sensory areas to higher level brain areas, as opposed to a task-specific uncertainty estimate at the level of decision-making in higher level brain areas (Fig. 1B) [see this review for a comparison of these two accounts: 84].

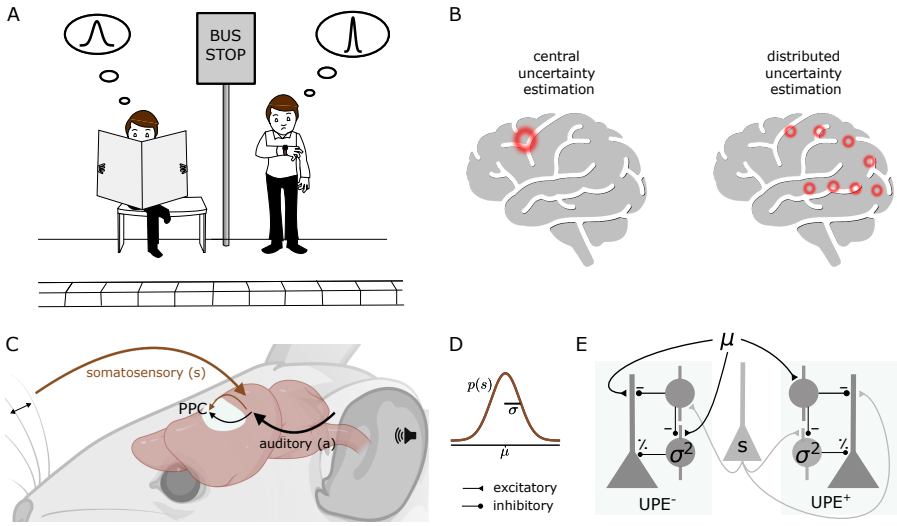

Figure 1: **Distributed uncertainty-modulated prediction error computation in cortical circuits** A: A person who learned that buses are unreliable has a prior expectation, which can be described by a wide Gaussian distribution of expected bus arrival times. When the bus does not arrive at the scheduled time, this person is not surprised and remains calm, as everything happens according to their model of the world. On the other hand, a person who learned that buses are punctual, which can be described by a narrow distribution of arrival times, may notice that the bus is late and get nervous, as they expected the bus to be punctual. This person can learn from this experience. If they always took this particular bus, and their uncertainty estimate is accurate, the prediction error could indicate that the bus schedule changed. B: Models of uncertainty representation in cortex. Some models suggest that uncertainty is only represented in higher-level areas concerned with decision-making (left). In contrast, we propose that uncertainty is represented at each level of the cortical hierarchy (right, shown is the visual hierarchy as an example). C: 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. D: The whisker stimulus intensities are drawn from a Gaussian distribution with mean μ and standard deviation σ. E: Negative (left) and positive (right) 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 in the postive circuit and the stimulus in the negative circuit, and PVs represent the variance.

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. The setting we consider is to predict a somatosensory stimulus based on an auditory stimulus. 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. 1C,D). 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:

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

where L is the likelihood.
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 cortical microcircuits achieve uncertainty modulation? Prediction errors can be positive or negative, but neuronal firing rates are always positive. Because baseline firing rates are low in layer 2/3 pyramidal cells [e.g., 57], positive and negative prediction errors were suggested to be represented by distinct neuronal populations [44, 66], which is in line with experimental data [39]. 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 error neurons compute prediction errors by subtracting the prediction from the stimulus input (or vice versa) [4]. The stimulus input can come from local stimulus-encoding layer 2/3 cells [65]. Inhibitory interneurons provide the subtraction, resulting in an excitation-inhibition balance when they match [34]. To represent a UPE, error neurons need 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. They probably either inherit the mean prediction or calculate it locally. Second, we hypothesise that inhibitory interneurons with divisive inhibitory effects represent the uncertainty σ^2 of the prediction (Fig. 1C,D), which they calculate locally. A layer 2/3 pyramidal cell that receives these sources of inhibition then reflects the uncertainty-modulated prediction error (Fig. 1E,F).
More specifically, we propose, first, that the SSTs are involved in the computation of the difference between predictions and stimuli, as suggested before [4]. This subtraction could happen on the apical dendrite. Second, we propose that the PVs provide the uncertainty modulation. In line with this, prediction error neurons in layer 2/3 receive subtractive inhibition from somatostatin (SST) and divisive inhibition from parvalbumin (PV) interneurons [81]. However, SSTs can also have divisive effects, and PVs can have subtractive effects, dependent on circuit and postsynaptic properties [71, 51, 15].
In the following, we investigate circuits of prediction error neurons and different inhibitory cell types. We start by investigating in local positive and negative prediction error circuits whether the inhibitory cell types can locally learn to predict means and variances, before combining both subcircuits into a recurrent circuit consisting of both positive and negative prediction error neurons.

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 in its current context. Uncertainty estimation requires multiple stimulus samples. Therefore, our brain needs to have a context-dependent mechanism to estimate uncertainty from multiple past instances of the sensory input. Let us consider the simple example from above, in which a sound stimulus represents a context with a particular amount of uncertainty. In terms of neural activity, the context could be encoded in a higher-level representation of the sound. Here, we investigate whether the context representation can elicit activity in the PVs that reflects the expected uncertainty of the situation. To investigate whether the context provided by the sound can cause activity in SSTs and PVs that reflects the mean and the variance of the whisker stimulus distribution, 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

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 Eq. 9 in Methods). The connection weight from the sound representation to the SSTs is plastic according to a local activity-dependent plasticity rule (see Eq. 10). The aim of this rule is to minimise the difference between the activation of the SSTs caused by the sound input (which

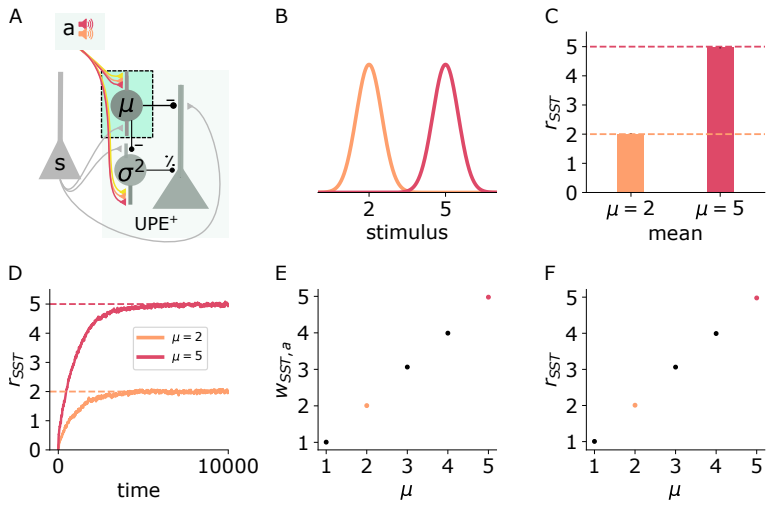

Figure 2: **SSTs learn to represent the mean context-dependently.** Illustration of the changes in the positive prediction error circuit. Thicker lines denote stronger weights. B: Two different tones (red, orange) are associated with two somatosensory stimulus distributions with different means (red: 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. Mean and std were computed over 1000 data points from timestep 9000 to 10000.

has 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 itself 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).

PVs learn to estimate the variance context-dependently

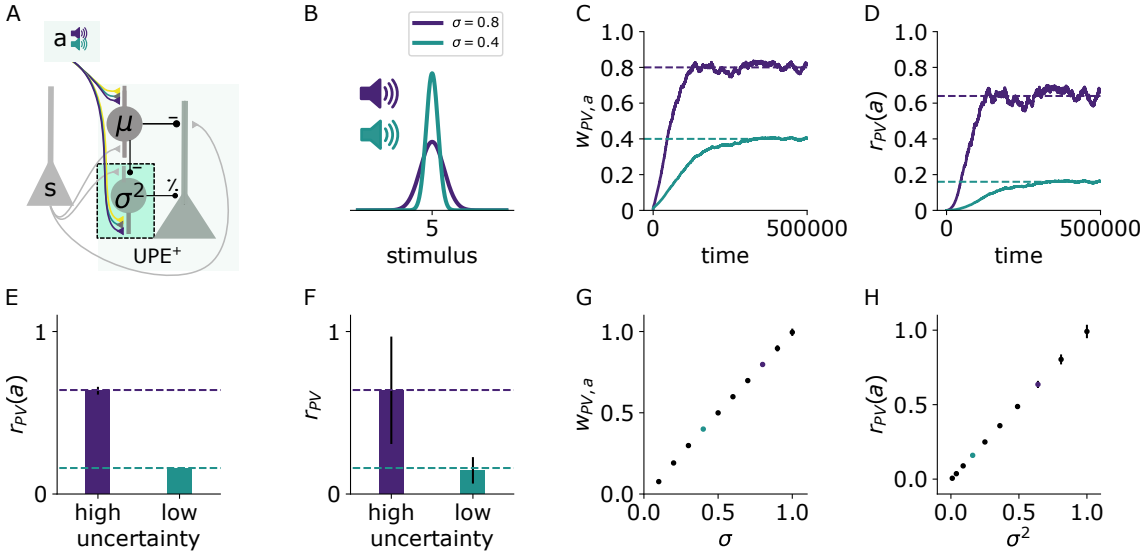

Figure 3: **PVs learn to estimate the variance context-dependently.** A: Illustration of the changes in the positive prediction error circuit. Thicker lines denote stronger weights. B: Two different tones (purple, green) are associated with two somatosensory stimulus distributions with different variances (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 whisker stimulus input) for low (green) and high (purple) stimulus variance. E: PV firing rates (mean and std) given sound input and whisker stimuli 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. Mean and std were computed from 150000 data points from timestep 450000 to 600000.

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: 69] 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 [12], which could help PVs to approximately estimate 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). PVs receive both the stimulus input as well as the inhibition from the SSTs, which subtracts the prediction of the mean (Eq. 11). 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 (Eq. 13). 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 assuming 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. 10), both when the sound input is presented alone (Fig. 3D,E,H) or when paired with whisker stimulation (Fig. 3F). Notably, a single PV neuron is sufficient for encoding 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 ($w_{\mathrm{PV,SST}}$) match the weights from the stimuli s to the PVs ($w_{\mathrm{PV},s}$). 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. 11 and Suppl. Methods).
The PVs can similarly estimate the uncertainty in negative prediction error circuits (Suppl. Fig. 12). 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

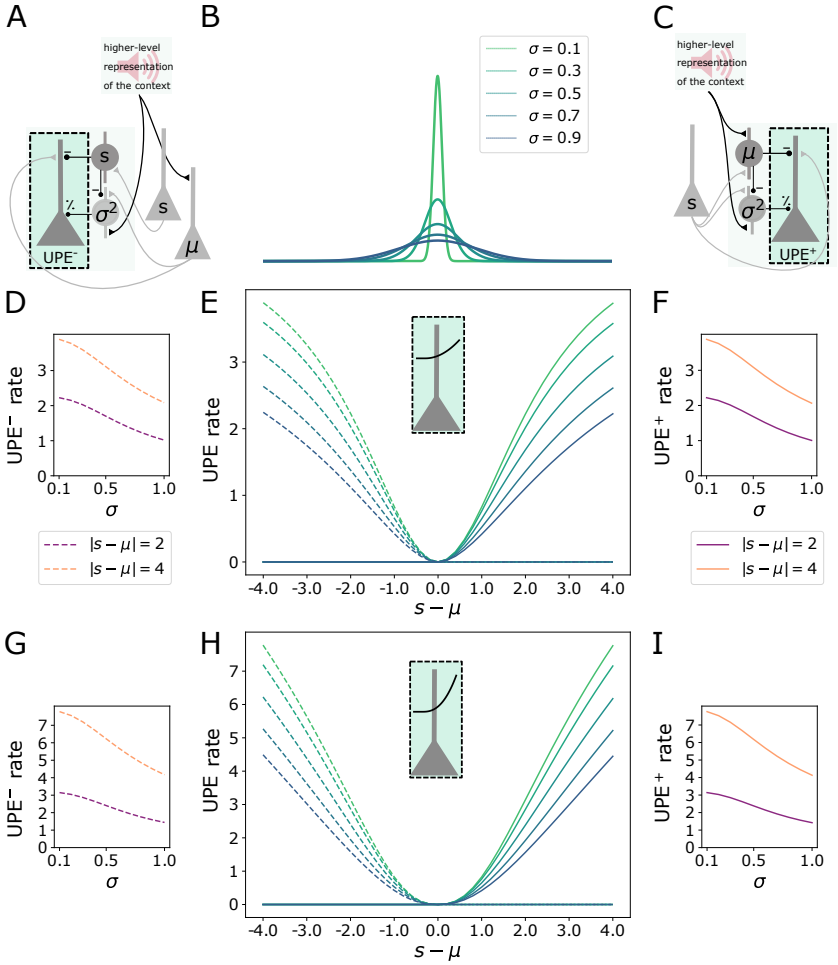

Figure 4: **Calculation of the UPE in layer 2/3 error neurons** A: Illustration of the negative prediction error circuit. B: Distributions with different standard deviations σ. C: Illustration of the positive prediction error circuit. D: Firing rate of the error neuron in the negative prediction error circuit (UPE^-) as a function of σ for two values of $|s-\mu|$ after learning μ and σ. E: Rates of both UPE^+ and UPE^--representing error neurons with a nonlinear activation function, where $k = 2.0$, 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 $|s-\mu|$ after learning μ and σ. G-I: same as D-F for error neurons with $k = 2.5$ with the same legend as in D-F.

Embedded in a circuit with subtractive and divisive interneuron types, layer 2/3 pyramidal cells could first compute the difference between the prediction and the stimulus in their dendrites, before their firing rate is divisively modulated by inhibition close to their soma. Layer 2/3 pyramidal cell dendrites can generate NMDA and calcium spikes, which cause a nonlinear integration of inputs. Hence, we took this into account and modelled the integration of dendritic activity as $I_{\text{dend}} = \lfloor w_{\text{UPE},s}\ s - w_{\text{UPE,SST}}\ r_{\text{SST}} \rfloor^k$ with k determining the non-linearity. The total activity of prediction error neurons was modelled by

$$\tau_{\text{E}} \frac{dr_{\text{UPE}}}{dt} = -r_{\text{UPE}} + \phi\left(\frac{I_{\text{dend}}}{I_0 + w_{\text{UPE,PV}}\ r_{\text{PV}}}\right). \tag{3}$$

The nonlinear integration of inputs is beneficial when the mean input changes and the current prediction differs strongly from the new mean of the stimulus distribution. For example, if the mean input increases strongly, the PV firing rate will increase for larger errors and inhibit error neurons more strongly than indicated by the learned variance estimate. The pyramidal nonlinearity compensates for this increased inhibition by PVs, such that in the end, layer 2/3 cell activity reflects an uncertainty-modulated prediction error (Fig. 4D-F) in both negative (Fig. 4A) and positive (Fig. 4B) prediction error circuits. A stronger nonlinearity (Fig. 4G-I) has the effect that error neurons elicit larger responses to larger prediction errors.

To ensure a comparison between the stimulus and the prediction, the inhibition from the SSTs needs to match the excitation, which it is compared to, in the UPE neurons: In the positive PE circuit, the weights from the SSTs representing the prediction to the UPE neurons need to match the weights from the stimulus s to the UPE neurons. In the negative PE circuit, the weights from SSTs representing the stimulus to the negative UPE neurons need to match the weights from the mean representation to the UPE neurons, respectively. In line with previous proposals, error neuron activity signals the breaking of EI balance [34, 7]. With inhibitory plasticity (target-based, see Suppl. Methods), the weights from the SSTs can learn to match the incoming excitatory weights (Suppl. Fig. 13).

Interactions between representation neurons and error neurons

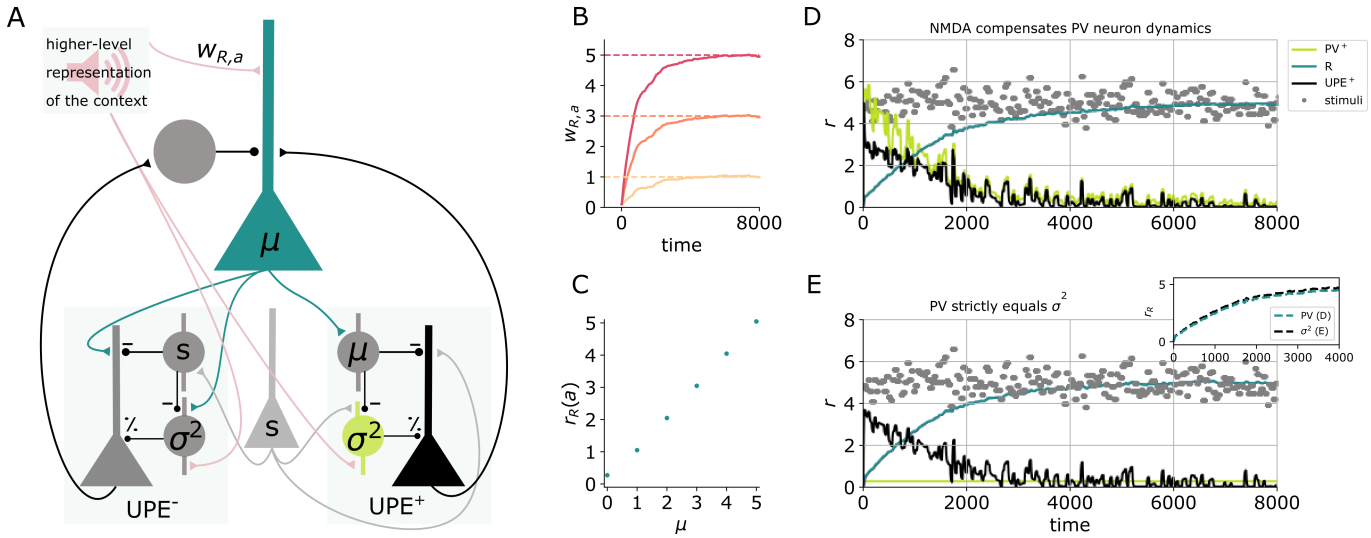

Figure 5: **Learning the mean representation with UPEs** A: Illustration of the circuit. A representation neuron (turquoise) receives input from both positive and negative prediction error circuits (UPE^+ and UPE^-) and projects back to them. The UPE^- has a negative impact on the firing rate of the representation neuron (r_R). A weight $w_{R,a}$ from the higher level representation of the context given by sound a is learned. B: Weights $w_{R,a}$ 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 over 50000 data points from timestep 50000 to 100000, the end of the simulation). D: Activity of the different cell types (PV: light green, R: turquiose, 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. E: Same colour code as in D. Inset shows comparison to 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 representation 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 $w_{R,a}$ from the sound representation to the representation neuron and the average firing rate of this representation neuron reflects the mean of the stimulus distribution (Fig.5B,C).

As discussed earlier, pyramidal cells tend to integrate their dendritic inputs nonlinearly due to NMDA spikes. We here show that a circuit with prediction error neurons with a dendritic nonlinearity (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). The dendritic nonlinearity can hence compensate for PV neuron dynamics. 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. 14).

Predictions for different cell types

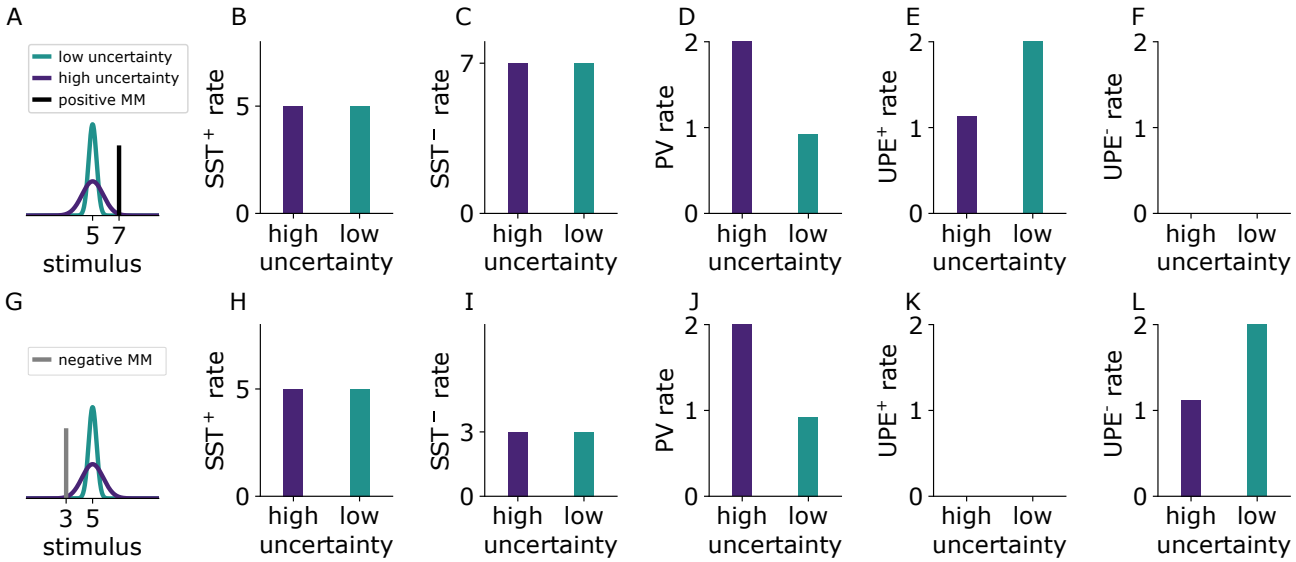

Figure 6: **Cell-type specific experimentally testable predictions** A: Illustration of the two experienced stimulus distributions with different variances that are associated with two different sounds (green, purple). The presented mismatch (MM) stimulus (black) is larger than expected (positive prediction error). B-F: Simulated 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. Because the firing rate predictions are equal for PV^+ and PV^-, we only show the results for PV^+ in the figure.

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 high-uncertainty (purple) versus the low-uncertainty (green) context are associated with different sounds). 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, the existence of so-called internal representation neurons has been proposed [44]. 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 (see the next section and Fig. 7).

The following experimental results are compatible with our predictions: First, putative inhibitory neurons (narrow spiking units) in the macaque anterior cingulate cortex increased their firing rates in periods of high uncertainty [6]. These inhibitory neurons could correspond to the PVs in our model. Second, prediction error activity seems to decrease in less predictable, and hence more uncertain, contexts: in mice reared in a predictable environment [where locomotion

and visual flow match 42], error neuron responses to mismatches in locomotion and visual flow decreased with each day of experiencing these unpredictable mismatches. Third, the responses of SSTs and PVs to mismatches between locomotion and visual flow [4] are in line with our model (note that in this experiment the mismatches are negative prediction errors as visual flow was halted despite ongoing locomotion): In this study, SST responses decreased during mismatch, i.e. when the visual flow was halted, and there was no difference between mice reared in a predictable or unpredictable environment. In line with these observations, the authors concluded that SST responses reflected the actual visual input. In our model negative PE circuit, SSTs also reflect the actual stimulus input, which in our case was a whisker stimulus (SST rates in Fig. 6C and I reflect the stimuli (black and grey bar) in A and G, respectively) and SST rates are the same for high and low uncertainty (corresponding to mice reared in a predictable or unpredictable environment). In the same study, PV responses were absent towards mismatches in animals reared in an unpredictable environment [4]. The authors argued that mice reared in an unpredictable environment did not learn to form a prediction. In our model, the missing prediction corresponds to missing predictive input from the auditory domain (e.g. due to undeveloped synapses from the predictive auditory input). If we removed the predictive input in our model, PVs in the negative PE circuit would also be silent as they would not receive any of the excitatory predictive inputs.

The effective learning rate is automatically adjusted with UPEs

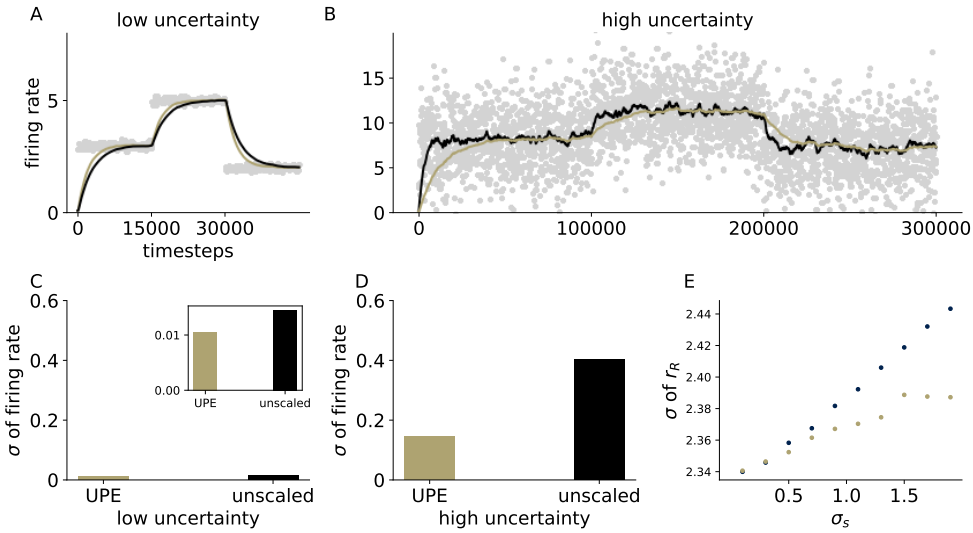

Figure 7: **Effective learning rate is automatically adjusted with UPEs** A,B: Firing rate over time of the representation neuron in a circuit with uncertainty-modulated prediction errors (gold) and in a circuit with unmodulated errors (black) in a low uncertainty setting (A) and a high uncertainty setting (B), C: standard deviation 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: standard deviation of the firing rate of the representation neuron in the high uncertainty setting. E: Standard deviation of the firing rate r_R as a function of the standard deviation of the presented stimulus distribution σ_s. Standard deviations were computed over 100000 data points from timestep 100000 to 200000

To test whether UPEs can automatically adjust the effective learning rate of a downstream neural population, we looked at two contexts that differed in uncertainty and compared how the mean representation evolves with and without UPEs. Indeed, in a low-uncertainty setting, a new 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.

UPEs ensure uncertainty-based weighting of prior and sensory information

Behavioural studies suggest that during perception humans integrate priors or predictions (p) and sensory information (s) in a Bayes-optimal manner [3, 60]. This entails that an internal neural representation (r), which determines perception, is achieved by weighting the two according to their uncertainties:

$$r = \frac{1}{c}\left(\frac{1}{\sigma_s^2}s + \frac{1}{\sigma_p^2}p\right), \tag{4}$$

where $c = \frac{1}{\sigma_\mathrm{s}^2} + \frac{1}{\sigma_\mathrm{p}^2}$, σ_s^2 is the uncertainty of the sensory information, and σ_p^2 is the uncertainty of the prior.

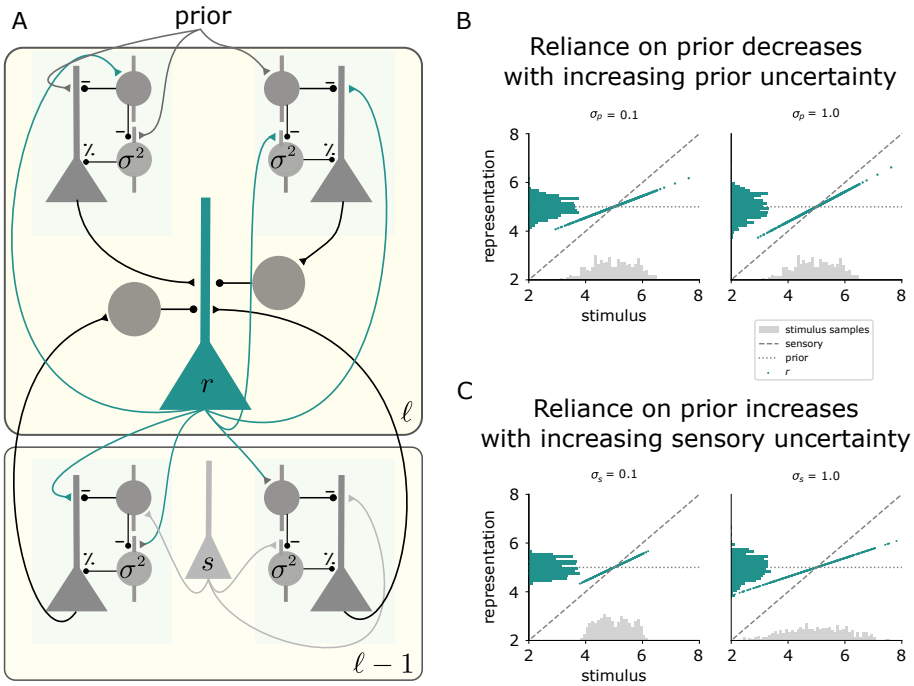

Figure 8: **UPEs ensure uncertainty-based weighting of prior and sensory information.** A: Illustration of the hierarchical two-area model. A representation neuron (r) in area ℓ receives positive and negative UPEs from the area $\ell - 1$ below (sensory prediction error, as in Fig. 5), and positive and negative UPEs from the same area (representation prediction error) with different signs, see Eq. 5. In this example, the uncertainty in area $\ell - 1$ corresponds to the sensory uncertainty σ_s^2, and the uncertainty in the area ℓ above, corresponds to the prior uncertainty σ_p^2. Both uncertainties are represented by PV activity in the respective area. B,C: The simulation results from the hierarchical circuit model yield an uncertainty-based weighting of prior and sensory information as in Eq. 4. B: The activity r of the representation neuron as a function of the stimulus s for different amounts of prior uncertainty σ_p^2, when the sensory uncertainty is fixed $\sigma_s^2 = 0.5$. The histograms show the distribution of the sensory stimulus samples (grey) and the distribution of the activity of the representation neuron (green). C: The activity of the representation neuron as a function of the stimulus for different amounts of sensory uncertainty σ_s^2, when the prior uncertainty is fixed $\sigma_p^2 = 0.5$.

To obtain this weighting in the steady state, prediction errors from the lower area, the sensory prediction error $(s - r)$, and from the local area, the representation prediction error $(r - p)$, can be used to update the current representation [as in 66]. Maximising the log-likelihood (c.f. Eq. 22 in Methods and Eq. 36 in SI) yields an update of the representation by the difference between the bottom-up and top-down prediction errors.

$$\dot{r} = \frac{1}{\sigma_s^2}(s - r) - \frac{1}{\sigma_p^2}(r - p) \tag{5}$$

From this we obtain Eq. 4 by setting $\dot{r} = 0$. Translating the update in Eq. 5 into our framework of UPE circuits reads as

$$\dot{r} = (\mathrm{UPE}_s^+ - \mathrm{UPE}_s^-) - (\mathrm{UPE}_p^+ - \mathrm{UPE}_p^-)\,, \tag{6}$$

illustrated in Fig. 8A.

In our circuit, the representation neuron (green) receives the bottom-up errors, which are modulated by sensory uncertainty σ_s^2, as in the recurrent circuit model from the previous section. In addition, it receives errors between its current activity and a prior (a higher level expectation of the representation). The prior is set to the learned mean of the stimulus distribution and the errors are modulated by the learned prior uncertainty σ_p^2. We simulated this circuit, to which we presented stimuli drawn from Gaussian distributions as before. We varied the PV activity reflecting the sensory and prior uncertainty, and measured the activity of the representation neuron r to each stimulus. The higher the uncertainty of the prior information σ_s, the more the representation reflects the current stimulus (Fig. 8B, see also Eq. 4). The higher the uncertainty of the sensory information σ_s, the more the representation reflects the prior mean (Fig. 8C). This is a common behavioural effect, which is often referred to as central tendency effect or contraction bias [35, 37, 2, 56].

To give a simple example how the prior uncertainty could come about in a dynamical environment, imagine noisy Gaussian sensory inputs with a fixed variance σ_s^2, and step changes in the mean, as in Fig. 7. On the lowest level 0, after faithful learning, the learned variance σ_0^2 will represent the variance of the sensory input σ_s^2 for a given mean. If in addition, the mean of the sensory input varies (as in Fig. 7), then on the level above σ_1^2 will reflect how much the mean varies. The former corresponds to the sensory uncertainty, the latter to the environmental volatility, which increases the uncertainty about the prediction for the mean (σ_p^2).

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, PV and SST interneurons, in particular, 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 specific to the processing level of a particular area. Experimental evidence is so far insufficient to favour distributed uncertainty representation over the idea that uncertainty is only computed on the level of decisions in higher level brain areas such as the parietal cortex [45], orbitofrontal cortex [55], or prefrontal cortex [68]. Our study provides a concrete suggestion for a canonical circuit implementation and, therefore, experimentally testable predictions. The distributed account has clear computational advantages for task-flexibility, information integration, active sensing, and learning (see [84] for a recent review of the two accounts). Additionally, adding uncertainty-modulated prediction errors from different hierarchical levels according to the predictive coding model [66, 77] yields an uncertainty-based weighting of feedback and feedforward information, similar to a Bayes-optimal weighting, which can be reconciled with human behaviour [58]. Two further important implications result from storing uncertainty 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 compatible with the 80/20 ratio of excitatory to inhibitory cells in the brain. The lower selectivity of interneurons in comparison to pyramidal cells could be a feature in prediction error circuits. Error neurons selective to similar stimuli are more likely to receive similar stimulus information, and hence similar predictions. Therefore, a circuit structure may have developed such that prediction error neurons with similar selectivity may receive inputs from the same inhibitory interneurons.

We claim that the uncertainty represented by PVs in our theoretical framework corresponds to *expected uncertainty* that results from noise or irreducible uncertainty in the stimuli and should therefore decrease the learning rate. Another common source of uncertainty are changes in the environment, also referred to as the *unexpected uncertainty*. In volatile environments with high unexpected uncertainty, the learning rate should increase. We suggest that vasointestinal-peptide-positive interneurons (VIPs) could be responsible for signalling the unexpected uncertainty, as they respond to reward, punishment and surprise [63], which can be indicators of high unexpected uncertainty. They provide potent disinhibition of pyramidal cells [61], and also inhibit PVs in layer 2/3 [62]. Hence, they could increase error activity resulting in a larger learning signal. In general, interneurons are innervated by different kinds of neuromodulators [52, 64] and control pyramidal cell's activity and plasticity [36, 25, 80, 79, 78]. Therefore, neuromodulators could have powerful control over error neuron activity and hence perception and learning.

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 [21], or amplitude of their responses [54], or implicitly via sampling [neural sampling hypothesis; 59, 9, 8], or rather than being represented by a single feature, can be decoded from the activity of an entire population [14]. While we suggest that PVs represent uncertainty to modulate prediction error responses, we do not claim that this is the sole representation of uncertainty in neuronal circuits.

Uncertainty estimation is relevant for Bayes-optimal integration of different sources of information, e.g., different modalities [multi-sensory integration; 18, 19] or priors and sensory information. Here, we present a circuit implementation for weighting sensory information versus priors according to their uncertainties by integrating uncertainty-modulated prediction errors. An alternative solution is to estimate uncertainty from the activity of prediction error neurons and use it to weight priors and sensory information [33], leading to contraction bias Hertäg and Clopath. [32] previously showed that the integration of prediction errors with sensory information in representation neurons can also lead to contraction bias, but without being dependent on uncertainty. Instead of modulating the error on each level by the uncertainty on that level as in our suggestion, one can also obtain a Bayes-optimal weighting by combining an unweighted top-down error with a bottom-error that is multiplicatively modulated by the prior uncertainty, and divisively modulated by the bottom-up uncertainty [29]. It has also been suggested that Bayes-optimal multi-sensory integration could be achieved in single neurons [19, 38]. Our proposal is complementary to these solutions 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 [27, 50, 75, 49, 72]. These disorders are also associated with aberrant inhibition, e.g. ASD is associated with an excitation-inhibition imbalance [67] and reduced inhibition [31, 24]. Interestingly, PV cells, in particular chandelier PV cells, were shown to be reduced in number and synaptic strength in ASD [41]. Our theory provides one possible explanation of how deficits in uncertainty-weighting on the behavioural level could be linked to altered PVs on the circuit level.

Finally, uncertainty-modulated errors could advance deep hierarchical neural networks. In addition to propagating gradients, propagating uncertainty may have advantages for learning. The additional information on uncertainty could enable calculating distances between distributions, which can provide an informative and parameter-independent metric for learning [e.g. natural gradient learning, 47].

To provide experimental predictions that are immediately testable, we suggested specific roles for SSTs and PVs, as they can subtractively and divisively modulate pyramidal cell activity, respectively. In 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.
The model predicts that the divisive interneuron type, which we here suggest to be the PVs, receives a representation of the stimulus as an input. PVs could be pooling the inputs from stimulus-responsive layer 2/3 neurons to estimate uncertainty. The more the stimulus varies, the larger the variability of the pyramidal neuron responses and, hence, the variability of the PV activity. The broader sensory tuning of PVs [13] is in line with the model insofar as uncertainty modulation could be more general than the specific feature, which is more likely for low-level features processed in primary sensory cortices. PVs were shown to connect more to pyramidal cells with similar feature tuning [83]; this would be in line with the model, as uncertainty modulation should be feature-related. In our model, some SSTs deliver the prediction to the positive prediction error neurons. SSTs are already known to be involved in spatial prediction, as they underlie the effect of surround suppression [1], in which SSTs suppress the local activity dependent on a predictive surround.

In the model we propose, SSTs should be subtractive and PVs divisive. However, SSTs can also be divisive, and PVs subtractive dependent on circuit and postsynaptic properties [71, 51, 15]. This does not necessarily contradict our model, as circuits in which SSTs are divisive and PVs subtractive could implement a different function, as not all pyramidal cells are error neurons. Hence, our model suggests that error neurons which can calculate UPEs should have similar physiological properties to the layer 2/3 cells observed in the study by [81].

Our model further posits the existence of two distinct subtypes of SSTs in positive and negative error circuits. Indeed, there are many different subtypes of SSTs [70]. SST is expressed by a large population of interneurons, which can be further subdivided. There is, for example, a type called SST44, which was shown to specifically respond when the animal corrects a movement [30]. Our proposal is hence aligned with the observation of functionally specialised subtypes of SSTs. Importantly, the comparison between stimulus and prediction needs to happen before the divisive modulation. Although our model does not make assumptions about the precise dendritic location of this comparison, we suggest this to happen on the apical dendrite, as top-down inputs and SST synapses arrive there. SSTs receive top-down inputs [53], which could provide the prediction to be subtracted in negative prediction error circuits.

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 assume 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 enable the comparison between predictions and sensory information via subtractive inhibition, we pointed out that the weights of those inputs on the postsynaptic neuron need to match. This essentially means that there needs to be a balance of excitatory and inhibitory inputs. Such an EI balance has been observed experimentally [73]. And it has previously been suggested that error responses are the result of breaking this EI balance [34, 7]. Heterosynaptic plasticity is a possible mechanism to achieve EI balance [20]. For example, spike pairing in pre- and postsynaptic neurons induces long-term potentiation at co-activated excitatory and inhibitory synapses with the degree of inhibitory potentiation depending on the evoked excitation [16], which can normalise EI balance [20].

Conclusion

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 compatible with Bayes-optimal behaviour and makes predictions for future experiments.

Methods

Derivation of the UPE

The goal is to learn $\hat{\mu}$ to maximise the log-likelihood:

$$\begin{aligned}\log \mathcal{L} &= \log p(\mathbf{s}|\hat{\mu}, \sigma) \\ &= \log \prod_{n=1}^{N} \mathcal{N}(s_n|\hat{\mu}, \sigma) \\ &= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (s_n - \hat{\mu})^2 - \frac{N}{2} \log(2\pi\sigma^2)\end{aligned}$$

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

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

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

$$\begin{aligned}\Delta\hat{\mu} &\propto \frac{\partial}{\partial\hat{\mu}}(\log p(s|\hat{\mu}, \sigma)) \\ &= \frac{\partial}{\partial\hat{\mu}}\left(-\frac{1}{2\sigma^2}(s - \hat{\mu})^2 - \frac{1}{2}\log(2\pi\sigma^2)\right) \\ &= \frac{1}{\sigma^2}(s - \hat{\mu}).\end{aligned}$$

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 [81]. We, hence, modelled the activity of prediction error neurons as

$$\tau_{\mathrm{E}} \frac{dr_{\mathrm{UPE}}}{dt} = -r_{\mathrm{UPE}} + \phi\left(\frac{I_{\mathrm{dend}}}{I_0 + w_{\mathrm{UPE,PV}}\, r_{\mathrm{PV}}}\right), \tag{7}$$

where $\phi(x)$ is the activation function defined as:

$$\phi(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 < x < x_{\max} \\ r_{\max} & \text{if } x \geq x_{\max}, \end{cases} \tag{8}$$

$I_{\mathrm{dend}} = \lfloor w_{\mathrm{UPE,s}}\, \mathrm{s} - w_{\mathrm{UPE,SST}}\, r_{\mathrm{SST}} \rfloor^k$ 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 1 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 $\sigma < 1.0$. All firing rates are rectified to ensure that they remain positive.

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

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{SST^+}}}{dt} = -r_{\mathrm{SST^+}} + \phi((1 - \beta) w_{\mathrm{SST^+,a}}\, r_{\mathrm{a}} + \beta s). \tag{9}$$

The SST activity is influenced (nudged with a factor β) by the somatosensory stimuli s, which provide targets for the desired SST activity. 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 [74]:

$$\Delta w_{\mathrm{SST,a}} = \eta(r_{\mathrm{SST}} - \phi(w_{\mathrm{SST,a}}\, r_{\mathrm{a}}))\, r_{\mathrm{a}}, \tag{10}$$

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.
The learning rule ensures that the auditory input alone causes SSTs to fire at their target activity. As in the original proposal [74], 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 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 ($w_{\mathrm{PV^+,s}}\, \mathrm{s} - w_{\mathrm{PV^+,SST^+}}\, r_{\mathrm{SST^+}}$, 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:

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

where $\phi_{\mathrm{PV}}(x)$ is a rectified quadratic activation function, defined as follows:

$$\phi_{\mathrm{PV}}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x^2 & \text{if } 0 < x < x_{\max} \\ r_{\max} & \text{if } x \geq x_{\max}. \end{cases} \tag{12}$$

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 [74]:

$$\Delta w_{\mathrm{PV,a}} = \eta(r_{\mathrm{PV}} - \phi_{\mathrm{PV}}(w_{\mathrm{PV,a}}\, r_{\mathrm{a}}))\, r_{\mathrm{a}}. \tag{13}$$

The weight from the sound representation to the PV $w_{\mathrm{PV^+,a}}$ 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 $\mathbb{E}[\lfloor s - \mu \rfloor^+] \propto \sigma$.

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 in the positive PE circuit inherit the mean representation from the representation neuron instead of learning it themselves, that is why they receive an input $w_{\mathrm{SST^+,R}} r_{\mathrm{R}}$. The SSTs in the negative circuit inherit the stimulus representation and hence receive an input $w_{\mathrm{SST^-,s}}\, \mathrm{s}$ 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 quadratic activation function with saturation, defined in the section below. All firing rates are rectified to ensure that they remain positive.

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{SST^+}}}{dt} = -r_{\mathrm{SST^+}} + \phi(w_{\mathrm{SST^+,R}}\, r_{\mathrm{R}}), \tag{14}$$

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{PV^+}}}{dt} = -r_{\mathrm{PV^+}} + \phi_{\mathrm{PV}}((1-\beta) w_{\mathrm{PV^+,a}}\, r_{\mathrm{a}} + \beta(w_{\mathrm{PV^+,s}}\, \mathrm{s} - w_{\mathrm{PV^+,SST^+}}\, r_{\mathrm{SST^+}}))), \tag{15}$$

$$\tau_{\mathrm{E}} \frac{dr_{\mathrm{UPE^+}}}{dt} = -r_{\mathrm{UPE^+}} + \phi\left(\frac{\lfloor w_{\mathrm{UPE,s}}\, \mathrm{s} - w_{\mathrm{UPE,SST^+}}\, r_{\mathrm{SST^+}} \rfloor^k}{I_0 + w_{\mathrm{UPE,PV^+}}\, r_{\mathrm{PV^+}}}\right), \tag{16}$$

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{SST^-}}}{dt} = -r_{\mathrm{SST^-}} + \phi(w_{\mathrm{SST^-,s}}\, \mathrm{s}), \tag{17}$$

$$\tau_{\mathrm{I}} \frac{dr_{\mathrm{PV^-}}}{dt} = -r_{\mathrm{PV^-}} + \phi_{\mathrm{PV}}((1-\beta) w_{\mathrm{PV^-,a}}\, r_{\mathrm{a}} + \beta(w_{\mathrm{PV^-,R}}\, r_{\mathrm{R}} - w_{\mathrm{PV^-,SST^-}}\, r_{\mathrm{SST^-}}))), \tag{18}$$

$$\tau_{\mathrm{E}} \frac{dr_{\mathrm{UPE^-}}}{dt} = -r_{\mathrm{UPE^-}} + \phi\left(\frac{\lfloor w_{\mathrm{UPE,R}}\, r_{\mathrm{R}} - w_{\mathrm{UPE,SST^-}}\, r_{\mathrm{SST^-}} \rfloor^k}{I_0 + w_{\mathrm{UPE,PV^-}}\, r_{\mathrm{PV^-}}}\right), \tag{19}$$

$$\tau_{\mathrm{E}} \frac{dr_{\mathrm{R}}}{dt} = -r_{\mathrm{R}} + \phi(w_{\mathrm{R,a}}\, r_{\mathrm{a}} + w_{\mathrm{R,UPE^+}}\, r_{\mathrm{UPE^+}} - w_{\mathrm{R,UPE^-}}\, r_{\mathrm{UPE^-}})). \tag{20}$$

Hierarchical predictive coding The idea behind hierarchical predictive coding is that the brain infers or represents the causes of its sensory inputs using a hierarchical generative model [23]. Each level of the cortical hierarchy provides a prior for the mean of the lower level representation, with the top level representation r_L being determined by the context. Noise enters in the sensory area by sampling a stimulus $s = r_0$. In the sensory area, the variance σ_0^2 is learned to

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.**

match the variability of the external stimulus σ_s^2. The goal is to infer the set of latent representations $\mathbf{r} = (r_1, .., r_{L-1})$, given the synaptic weights w_ℓ of the internal model and the top activity r_L, that best reproduces the sensory inputs s. To this end, we minimise the following energy:

$$E = \frac{1}{2}\sum_{\ell=0}^{L-1}\frac{1}{\sigma_\ell^2}\|r_\ell - \rho(w_\ell r_{\ell+1})\|_2^2\,, \tag{21}$$

where ρ is a transfer function.
We obtain the update for r_ℓ with gradient descent on the energy with respect to r_ℓ:

$$\begin{aligned}\dot{r_\ell} = -\frac{\partial E}{\partial r_\ell} &= \frac{1}{\sigma_{\ell-1}^2}(r_{\ell-1} - \rho(w_{\ell-1}r_\ell))w_{\ell-1}\rho'(r_\ell) - \frac{1}{\sigma_\ell^2}(r_\ell - \rho(w_\ell r_{\ell+1})) \\ &= \frac{1}{\sigma_{\ell-1}^2}\rho'(w_{\ell-1}r_\ell)e_{\ell-1} - \frac{1}{\sigma_\ell^2}e_\ell\end{aligned} \tag{22}$$

In our model, w_ℓ are scalars as they denote single weights.

We obtain the steady-state representation r_ℓ by setting its derivative in Eq. 21 to 0:

$$0 \stackrel{!}{=} \frac{w_{\ell-1}\rho'(w_{\ell-1}r_\ell)}{\sigma_{\ell-1}^2}(r_{\ell-1} - \rho(w_{\ell-1}r_\ell)) - \frac{1}{\sigma_\ell^2}(r_\ell - \rho(w_\ell r_{\ell+1})) \tag{23}$$

We next consider, for simplicity, a threshold-linear transfer function $\rho(w_{\ell-1}r_\ell)$ such that if $w_{\ell-1}r_\ell < 0$, then $\rho(w_{\ell-1}r_\ell) = 0$ and its derivative $\rho' = 0$ or if $w_{\ell-1}r_\ell \geq 0$ then $\rho(w_{\ell-1}r_\ell) = w_{\ell-1}r_\ell$ and $\rho' = 1$.
Solving Eq. 24 for r_ℓ and assuming $r_\ell \geq 0$, we get:

$$r_\ell = \frac{\frac{w_{\ell-1}^2}{\sigma_{\ell-1}^2}}{\frac{w_{\ell-1}^2}{\sigma_{\ell-1}^2} + \frac{1}{\sigma_\ell^2}}\frac{r_{\ell-1}}{w_{\ell-1}} + \frac{\frac{1}{\sigma_\ell^2}}{\frac{w_{\ell-1}^2}{\sigma_{\ell-1}^2} + \frac{1}{\sigma_\ell^2}}w_\ell r_{\ell+1}. \tag{24}$$

See appendix for the general case with any transfer function and weight matrix.

Hierarchical circuit model In the hierarchical circuit model, the representation neuron does not only receive UPEs from the area below, but also from the current area.

$$\tau_{\mathrm{E}}\frac{dr_{\mathrm{R}}}{dt} = -r_{\mathrm{R}} + \phi(w_{\mathrm{R,a}}\, r_{\mathrm{a}} + w_{\mathrm{R},\mathrm{UPE}^+_{\ell-1}}\, r_{\mathrm{UPE}^+_{\ell-1}} - w_{\mathrm{R},\mathrm{UPE}^-_{\ell-1}}\, r_{\mathrm{UPE}^-_{\ell-1}} - w_{\mathrm{R},\mathrm{UPE}^+_{\ell}}\, r_{\mathrm{UPE}^+_{\ell}} + w_{\mathrm{R},\mathrm{UPE}^-_{\ell}}\, r_{\mathrm{UPE}^-_{\ell}}). \tag{25}$$

The UPEs from the area below are as defined in the recurrent circuit model, and the UPEs from the current area are defined accordingly as:

$$\tau_{\mathrm{E}} \frac{dr_{\mathrm{UPE}_\ell^+}}{dt} = -r_{\mathrm{UPE}_\ell^+} + \phi\left(\frac{\lfloor w_{\mathrm{UPE},r_{\mathrm{R}}}\ \mathrm{R} - w_{\mathrm{UPE,SST}^+}\ r_{\mathrm{SST}^+}\rfloor^k}{I_0 + w_{\mathrm{UPE,PV}^+}\ r_{\mathrm{PV}^+}}\right), \tag{26}$$

$$\tau_{\mathrm{E}} \frac{dr_{\mathrm{UPE}_\ell^-}}{dt} = -r_{\mathrm{UPE}_\ell^-} + \phi\left(\frac{\lfloor w_{\mathrm{UPE,prior}}\ \mu_{\mathrm{prior}} - w_{\mathrm{UPE,SST}^-}\ r_{\mathrm{SST}^-}\rfloor^k}{I_0 + w_{\mathrm{UPE,PV}^-}\ r_{\mathrm{PV}^-}}\right), \tag{27}$$

The computations and parameters in each area are the same as for the recurrent circuit model above and in Fig. 5.

Parameter	Value	Description
$w_{\mathrm{R,UPE}_\ell^+}$	1.0	weight from UPE_ℓ^+ to R
$w_{\mathrm{R,UPE}_\ell^-}$	1.0	weight from UPE_ℓ^- to R
$w_{\mathrm{UPE}_\ell^-,\mathrm{R}}$	1.0	weight from R to UPE_ℓ^-
$w_{\mathrm{UPE}^-,\mathrm{prior}}$	1.0	weight from R to UPE^-

Table 2: **Additional Parameters of the hierarchical network.**

Parameter	Value	Description
a	$\{0.0, 1.0\}$	auditory stimulus (on/off)
s	$\sim \mathcal{N}(\mu, \sigma)$	somatosensory (whisker) stimulus
N	1000-20000	number of whisker stimulus samples
D	$\{10, 100\}$	stimulus duration (Figs. 1-5,7; Fig. 7)

Table 3: **Inputs.**

Synapses from the higher level representation of the sound a to R were plastic according to the following activity-dependent plasticity rules [74].

$$\Delta w_{\mathrm{R,a}} = \eta_{\mathrm{R}}(r_{\mathrm{R}} - \phi(w_{\mathrm{R,a}} r_{\mathrm{a}}))r_{\mathrm{a}}, \tag{28}$$

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

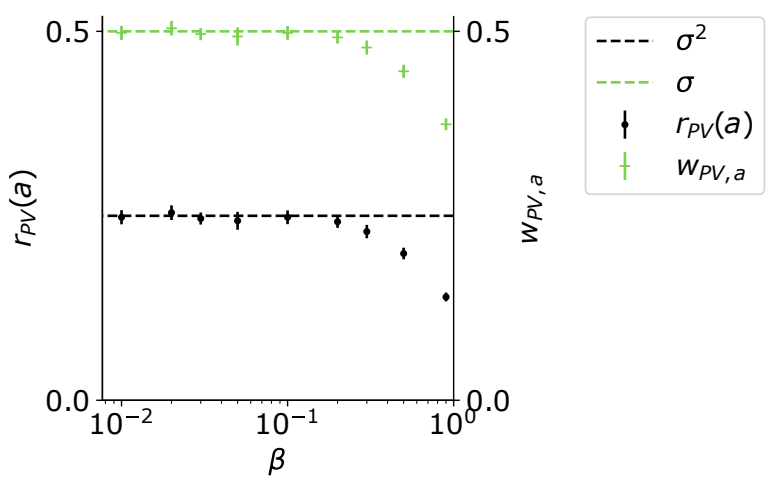

Figure 9: For small β, and $w_s = \sqrt{\frac{2-\beta}{\beta}}$, the weight from a to PV approaches σ and the PV firing rate approaches σ^2.

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 ($w_{\mathrm{PV},s}$). 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 $\mathbb{E}[wa]^2 = \sigma^2$ when the average weight change $\mathbb{E}[\Delta w] = 0$. The learning rule for w is as follows:

$$\begin{aligned}\Delta w &= \eta[r_{\mathrm{PV}} - \phi(wa)]a \\ &= \eta[\phi((1-\beta)wa + \beta w_s \tilde{s}) - \phi(wa)]a\end{aligned}$$

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:

$$\begin{aligned}\mathbb{E}[\Delta w] &= \mathbb{E}[\eta((1-\beta)^2 w^2 a^2 + \beta^2 w_s^2 \tilde{s}^2 + 2(1-\beta)wa\beta w_s \tilde{s} - w^2 a^2)a] \\ &= \mathbb{E}[\eta((1+\beta^2-2\beta)w^2 a^2 + \beta^2 w_s^2 \tilde{s}^2 + 2(1-\beta)wa\beta w_s \tilde{s} - w^2 a^2)a] \quad | \quad \mathbb{E}[\tilde{s}] = 0 \\ &= \mathbb{E}[\eta(\beta^2 w^2 a^2 - 2\beta w^2 a^2 + \beta^2 w_s^2 \tilde{s}^2)a] \\ &= \mathbb{E}[\eta\beta(\beta w^2 a^2 - 2w^2 a^2 + \beta w_s^2 \tilde{s}^2)a] \\ &= \eta\beta((\beta-2)\mathbb{E}[(wa)^2] + \beta w_s^2 \mathbb{E}[\tilde{s}^2])a \quad | \quad \mathbb{E}[\tilde{s}^2] = \mathbb{E}[(s-\mu)^2] = \sigma^2 \\ &= \eta\beta((\beta-2)\mathbb{E}[(wa)^2] + \beta w_s^2 \sigma^2)a\end{aligned}$$

Given our objective $\mathbb{E}[(wa)^2] = \sigma^2$, we can write:

$$\mathbb{E}[\Delta w] = \eta\beta((\beta-2)\sigma^2 + \beta w_s^2 \sigma^2)a$$

Then for $\mathbb{E}[\Delta w] = 0$:

$$\begin{aligned}0 &= \beta - 2 + \beta w_s^2 \\ \Rightarrow w_s &= \sqrt{\frac{2-\beta}{\beta}}\end{aligned}$$

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. 9). This analysis shows that the approximation holds for small β up to a value of $\beta = 0.2$.

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

Table 4: **Parameters of the plasticity rules.**

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 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 $dt = 0.1$.

Eliciting responses to mismatches (Fig. 4 and Fig. 6) We first trained the circuit with 10000 stimulus samples to learn the variances in the a-to-PV weights. Then we presented different mismatch stimuli to calculate the error magnitude for each mismatch of magnitude $s - \mu$.

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 5: **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 6: **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.

References

1. H. Adesnik, W. Bruns, H. Taniguchi, Z. J. Huang, M. Scanziani, *Nature* **490**, 226 EP – (Oct. 2012).
2. A. Akrami, C. D. Kopec, M. E. Diamond, C. D. Brody, eng, *Nature* **554**, 368–372 (2018).
3. P. Ashourian, Y. Loewenstein, *PLOS ONE* **6**, 1–8, (`https://doi.org/10.1371/journal.pone.0019551`) (May 2011).
4. A. Attinger, B. Wang, G. B. Keller, *Cell* **169**, 1291–1302.e14, (`https://doi.org/10.1016/j.cell.2017.05.023`) (2017).
5. A. Ayaz *et al.*, *Nature Communications* **10**, 2585, (`https://doi.org/10.1038/s41467-019-10564-8`) (2019).
6. 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).
7. M. L. L. R. Barry, W. Gerstner, *PLOS Computational Biology* **20**, 1–41, (`https://doi.org/10.1371/journal.pcbi.1011839`) (Feb. 2024).
8. 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).
9. L. Buesing, J. Bill, B. Nessler, W. Maass, eng, *PLoS Comput Biol* **7**, e1002211 (2011).
10. J. Cannon, A. M. O'Brien, L. Bungert, P. Sinha, eng, *Autism Res* **14** (2021).
11. M. X. Cohen, K. Wilmes, I. v. d. Vijver, eng, *Trends Cogn Sci* **15**, 558–566 (2011).
12. 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).
13. J. C. H. Cottam, S. L. Smith, M. Häusser, *Journal of Neuroscience* **33**, 19567–19578 (2013).
14. G. P. Dehaene, R. Coen-Cagli, A. Pouget, *PLOS Computational Biology* **17**, 1–30, (`https://doi.org/10.1371/journal.pcbi.1008138`) (Feb. 2021).
15. C. Dorsett, B. D. Philpot, S. L. Smith, I. T. Smith, eng, *eNeuro* **8** (2021).
16. J. A. D'amour, R. C. Froemke, *Neuron* **86**, 514–528, (`https://www.sciencedirect.com/science/article/pii/S089662731500210X`) (2015).
17. S. J. Eliades, X. Wang, *Nature* **453**, 1102–1106, (`https://doi.org/10.1038/nature06910`) (2008).
18. M. O. Ernst, M. S. Banks, *Nature* **415**, 429–433, (`https://doi.org/10.1038/415429a`) (2002).
19. C. R. Fetsch, A. Pouget, G. C. DeAngelis, D. E. Angelaki, *Nature Neuroscience* **15**, 146–154, (`https://doi.org/10.1038/nn.2983`) (2012).

20. R. E. Field *et al.*, eng, *Neuron* **106**, 842–854 (2020).
21. B. J. Fischer, J. Peña, *Nature Neuroscience* **14**, 1061–1066, (https://doi.org/10.1038/nn.2872) (2011).
22. A. Fiser *et al.*, *Nature Neuroscience* **19**, 1658 EP, (https://doi.org/10.1038/nn.4385) (Sept. 2016).
23. K. Friston, eng, *Philos Trans R Soc Lond B Biol Sci* **360**, 815–836 (2005).
24. W Gaetz *et al.*, eng, *Neuroimage* **86**, 1–9 (2014).
25. A. Gidon, I. Segev, *Neuron* **75**, 330–341, (https://www.sciencedirect.com/science/article/pii/S0896627312004813) (2012).
26. 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).
27. J. Goris *et al.*, *Autism* **25**, PMID: 33030041, 440–451, eprint: https://doi.org/10.1177/1362361320962237, (https://doi.org/10.1177/1362361320962237) (2021).
28. J. Goris *et al.*, eng, *Biol Psychiatry Cogn Neurosci Neuroimaging* **3**, 667–674 (2018).
29. A. Granier, M. A. Petrovici, W. Senn, K. A. Wilmes, *Confidence and second-order errors in cortical circuits*, 2024, arXiv: 2309.16046 [id='q-bio.NC' full$_name =' NeuronsandCognition'is_active = Truealt_name = Nonein_archive =' q-bio'is_general = Falsedescription =' Synapse, cortex, neuronaldynamics, neuralnetwork, sensorimot$
30. J. Green *et al.*, *Nature* **620**, 366–373, (https://doi.org/10.1038/s41586-023-06357-1) (2023).
31. M. Harada *et al.*, eng, *J Autism Dev Disord* **41**, 447–454 (2011).
32. L. Hertäg, C. Clopath, eng, *Proc Natl Acad Sci U S A* **119**, e2115699119 (2022).
33. L. Hertäg, K. A. Wilmes, C. Clopath, *bioRxiv*, eprint: https://www.biorxiv.org/content/early/2023/12/14/2023.12.13.571410.full.pdf, (https://www.biorxiv.org/content/early/2023/12/14/2023.12.13.571410) (2023).
34. L. Hertäg, H. Sprekeler, *eLife* **9**, ed. by S. Ostojic, R. B. Ivry, S. Ostojic, e57541, (https://doi.org/10.7554/eLife.57541) (2020).
35. H. L. Hollingworth, *The Journal of Philosophy, Psychology and Scientific Methods* **7**, 461–469, (2024; http://www.jstor.org/stable/2012819) (1910).
36. J. S. Isaacson, M. Scanziani, *Neuron* **72**, 231–243, (https://www.sciencedirect.com/science/article/pii/S0896627311008798) (2011).
37. M. Jazayeri, M. N. Shadlen, eng, *Nat Neurosci* **13**, 1020–1026 (2010).
38. 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].
39. R. Jordan, G. B. Keller, *Neuron* **108**, 1194–1206.e5, (https://www.sciencedirect.com/science/article/pii/S0896627320307480) (2020).
40. R. Jordan, G. B. Keller, (https://doi.org/10.7554/elife.85111.2) (2023).
41. P. Juarez, V. Martínez Cerdeño, eng, *Front Psychiatry* **13**, 913550 (2022).
42. G. B. Keller, T. Bonhoeffer, M. Hübener, *Neuron* **74**, 809 –815, (http://www.sciencedirect.com/science/article/pii/S0896627312003844) (2012).
43. G. B. Keller, R. H. R. Hahnloser, *Nature* **457**, 187–190, (https://doi.org/10.1038/nature07467) (2009).
44. G. B. Keller, T. D. Mrsic-Flogel, *Neuron* **100**, 424 –435 (2018).
45. R. Kiani, M. N. Shadlen, eng, *Science* **324**, 759–764 (2009).
46. K. P. Körding, D. M. Wolpert, *Nature* **427**, 244–247, (https://doi.org/10.1038/nature02169) (2004).
47. E. Kreutzer, W. Senn, M. A. Petrovici, *eLife* **11**, ed. by P. Latham, J. R. Huguenard, e66526, (https://doi.org/10.7554/eLife.66526) (2022).
48. K. Kveraga, A. S. Ghuman, M. Bar, *Brain and cognition* **65**, 145–168, (https://pubmed.ncbi.nlm.nih.gov/17923222) (Nov. 2007).
49. R. P. Lawson, C. Mathys, G. Rees, *Nature Neuroscience* **20**, 1293–1299, (https://doi.org/10.1038/nn.4615) (2017).
50. R. P. Lawson, G. Rees, K. J. Friston, *Frontiers in human neuroscience* **8**, 302–302, (https://pubmed.ncbi.nlm.nih.gov/24860482) (May 2014).
51. S.-H. Lee *et al.*, *Nature* **488**, 379–383, (https://doi.org/10.1038/nature11312) (2012).

52. S. Lee, I. Kruglikov, Z. J. Huang, G. Fishell, B. Rudy, *Nature Neuroscience* **16**, 1662–1670 (2013).
53. D. Liu *et al.*, *Nature Communications* **11**, 2784, (https://doi.org/10.1038/s41467-020-16609-7) (2020).
54. W. J. Ma, J. M. Beck, P. E. Latham, A. Pouget, *Nature Neuroscience* **9**, 1432–1438, (https://doi.org/10.1038/nn1790) (2006).
55. P. Masset, T. Ott, A. Lak, J. Hirokawa, A. Kepecs, *Cell* **182**, 112–126.e18, (https://www.sciencedirect.com/science/article/pii/S0092867420306176) (2020).
56. N. Meirhaeghe, H. Sohn, M. Jazayeri, *Neuron* **109**, 2995–3011.e5, (https://www.sciencedirect.com/science/article/pii/S089662732100622X) (2021).
57. C. M. Niell, M. P. Stryker, *Journal of Neuroscience* **28**, 7520–7536 (2008).
58. E. Payzan-LeNestour, P. Bossaerts, *PLOS Computational Biology* **7**, 1–14, (https://doi.org/10.1371/journal.pcbi.1001048) (Jan. 2011).
59. M. A. Petrovici, J. Bill, I. Bytschok, J. Schemmel, K. Meier, *Stochastic inference with deterministic spiking neurons*, 2013, arXiv: 1311.3211 [q-bio.NC].
60. F. H. Petzschner, S. Glasauer, *Journal of Neuroscience* **31**, 17220–17229, eprint: https://www.jneurosci.org/content/31/47/17220.full.pdf, (https://www.jneurosci.org/content/31/47/17220) (2011).
61. C. K. Pfeffer, *Current Biology* **24**, 18–20 (2013).
62. C. K. Pfeffer, M. Xue, M. He, Z. J. Huang, M. Scanziani, *Nature neuroscience* **16**, 1068–1076 (Aug. 2013).
63. H.-J. Pi *et al.*, *Nature* **503**, 521–524 (2013).
64. A. Prönneke *et al.*, *Cerebral Cortex* **25**, 4854–4868 (2015).
65. C. Raltschev, S. Kasavica, B. Leonardon, T. Nevian, S. Sachidhanandam, *bioRxiv*, eprint: https://www.biorxiv.org/content/early/2023/05/17/2023.05.11.540431.full.pdf, (https://www.biorxiv.org/content/early/2023/05/17/2023.05.11.540431) (2023).
66. R. P. N. Rao, D. H. Ballard, *Nature Neuroscience* **2**, 79–87, (https://doi.org/10.1038/4580) (1999).
67. 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).
68. M. F. S. Rushworth, T. E. J. Behrens, *Nature Neuroscience* **11**, 389–397, (https://doi.org/10.1038/nn2066) (2008).
69. S. Sachidhanandam, B. S. Sermet, C. C. Petersen, *Cell Reports* **15**, 700–706, (https://www.sciencedirect.com/science/article/pii/S2211124716303345) (2016).
70. C. M. Schneider-Mizell *et al.*, eng, *bioRxiv* (2024).
71. 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).
72. Z. Shi *et al.*, *bioRxiv*, (https://www.biorxiv.org/content/early/2022/01/25/2022.01.21.477218) (2022).
73. A. Tan, M. Wehr, *Neuroscience* **163**, 1302–1315, (https://www.sciencedirect.com/science/article/pii/S0306452209012093) (2009).
74. R. Urbanczik, W. Senn, *Neuron* **81**, 521–528, (https://www.sciencedirect.com/science/article/pii/S0896627313011276) (2014).
75. S. Van de Cruys *et al.*, eng, *Psychol Rev* **121**, 649–675 (2014).
76. 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).
77. J. C. R. Whittington, R. Bogacz, eng, *Neural Comput* **29**, 1229–1262 (2017).
78. K. A. Wilmes, C. Clopath, *Nature Communications* **10**, 5055, (https://doi.org/10.1038/s41467-019-12972-2) (2019).
79. K. A. Wilmes, J.-H. Schleimer, S. Schreiber, *European Journal of Neuroscience* **45**, 1032–1043, eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.13326 (2017).
80. K. A. Wilmes, H. Sprekeler, S. Schreiber, *PLoS Computational Biology* **12**, e1004768 (2016).
81. N. R. Wilson, C. A. Runyan, F. L. Wang, M. Sur, eng, *Nature* **488**, 343–348 (2012).
82. P. Zmarz, G. B. Keller, eng, *Neuron* **92**, 766–772 (2016).
83. P. Znamenskiy *et al.*, eng, *Neuron* **112**, 991–1000 (2024).
84. Ádám Koblinger, J. Fiser, M. Lengyel, *Current Opinion in Behavioral Sciences* **38**, Computational cognitive neuroscience, 150–162, (https://www.sciencedirect.com/science/article/pii/S2352154621000577) (2021).

Acknowledgements

We would like to thank Loreen Hertäg and Jakob Jordan for helpful discussions and 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.

Competing Interests Statement

The authors declare that they have no competing interests.

Code availability

All simulation code used for this paper will be made available on GitHub upon publication (`https://github.com/k47h4/UPE`) and is attached to the submission as supplementary file for the reviewers.

Supplementary Information

Supplementary Methods

Synaptic dynamics/plasticity rules

$$\Delta w_{\mathrm{PV}^+,\mathrm{SST}^+} = \eta_{\mathrm{PV}}(w_{\mathrm{PV}^+,s}\mathrm{s} - w_{\mathrm{PV}^+,\mathrm{SST}^+} r_{\mathrm{SST}^+}) r_{\mathrm{SST}^+}, \tag{29}$$

$$\Delta w_{\mathrm{PV}^-,\mathrm{SST}^-} = \eta_{\mathrm{PV}}(w_{\mathrm{PV}^-,R} r_{\mathrm{R}} - w_{\mathrm{PV}^-,\mathrm{SST}^-} r_{\mathrm{SST}^-}) r_{\mathrm{SST}^-}, \tag{30}$$

$$\Delta w_{\mathrm{UPE}^+,\mathrm{SST}^+} = \eta_{\mathrm{UPE}}(w_{\mathrm{UPE}^+,s}\mathrm{s} - w_{\mathrm{UPE}^+,\mathrm{SST}^+} r_{\mathrm{SST}^+}) r_{\mathrm{SST}^+}, \tag{31}$$

$$\Delta w_{\mathrm{UPE}^+,\mathrm{SST}^-} = \eta_{\mathrm{UPE}}(w_{\mathrm{UPE}^-,\mathrm{R}} r_{\mathrm{R}} - w_{\mathrm{UPE}^-,\mathrm{SST}^-} r_{\mathrm{SST}^-}) r_{\mathrm{SST}^-}, \tag{32}$$

$$\tag{33}$$

Different choice of supralinear activation function for PV

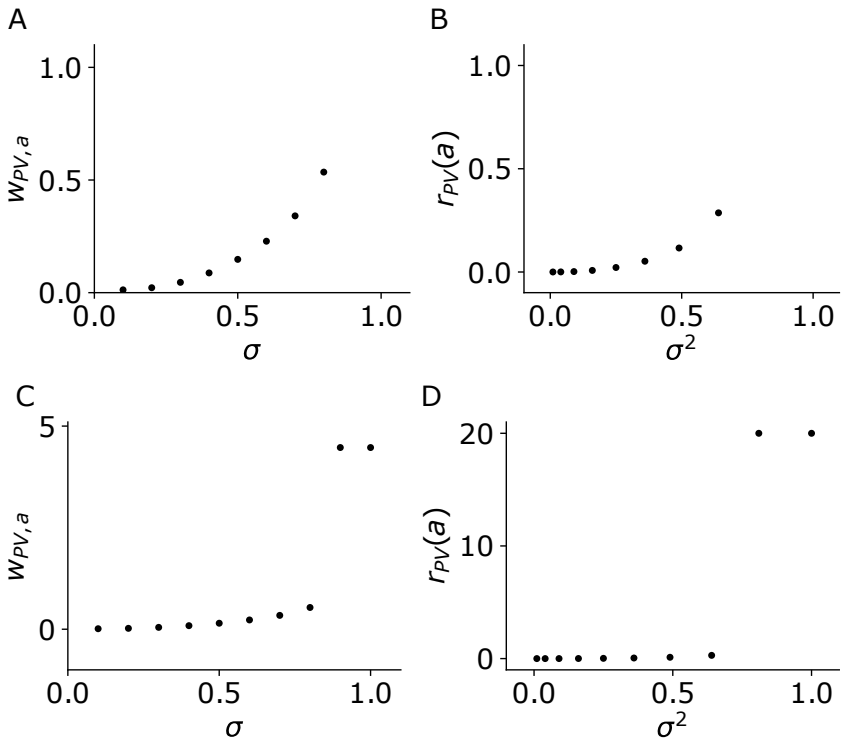

Figure 10: Learning the variance in the positive prediction error circuit with PVs with a power activation function (exponent = 3.0). A and B are analogous to Fig. 3G and H, and the circuit is the same except that the activation function of the PVs ($\phi_{PV}(x)$) has an exponent of 3.0 instead of 2.0. C and D are zoomed-out versions of A and B.

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

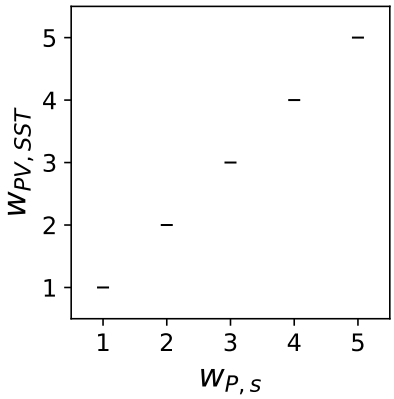

Figure 11: With inhibitory plasticity, weights from SST to PV can be learned. This figure shows that the weight from SST to PV ($w_{\mathrm{PV},\mathrm{SST}}$) is equal to the weight from s to PV ($w_{\mathrm{PV},s}$). The inhibitory plasticity rule is described in the Supplementary Methods.

PVs learn the variance in the negative prediction error circuit

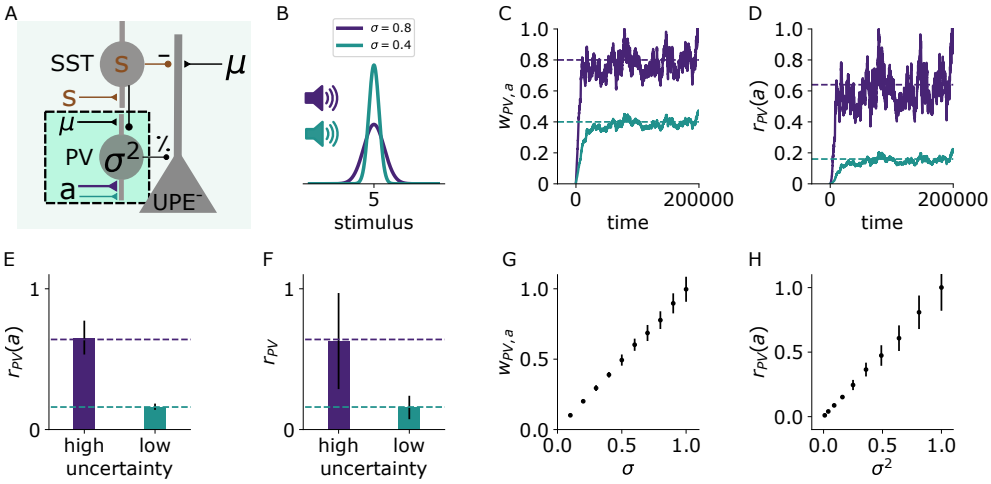

Figure 12: **PVs learn to represent the variance given an associative cue in the negative prediction error circuit.** A: Illustration of the changes in the negative prediction error circuit. Thicker lines denote stronger weights. B: Two different tones (purple, green) are associated with two somatosensory 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). H: PV firing rates given sound input for different values of σ^2 (mean and std).

Learning the weights from the SSTs to the prediction error neurons

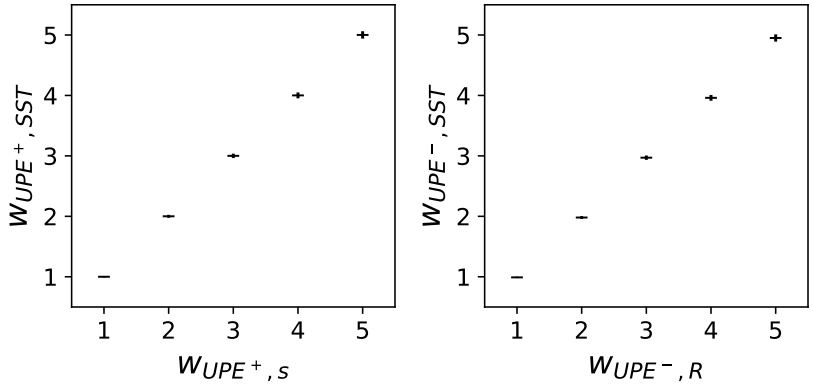

Figure 13: Learning the weights from the SSTs to the UPE neurons. This figure shows that the weights from the SSTs to the UPEs in both the positive (left) and the negative (right) prediction error circuit can be learned with inhibitory plasticity to match the weights from the stimulus representation 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

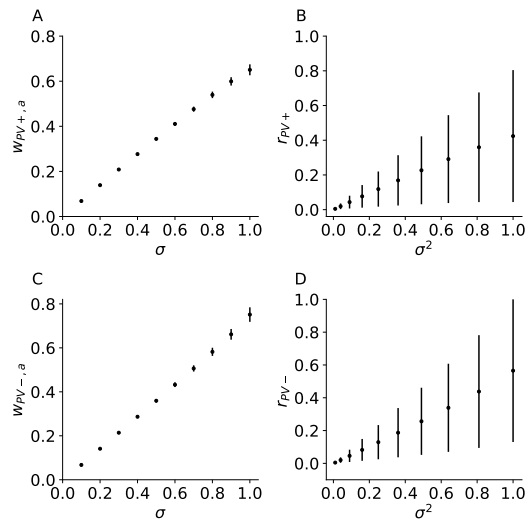

Figure 14: PV firing rates are proportional to the variance in the recurrent circuit model. Weights from a to PV as a function of σ in the positive (A) and negative (C) prediction error subcircuit. PV firing rates as a function of σ^2 in the positive (B) and negative (D) prediction error circuit.

Hierarchical predictive coding with uncertainties

The total energy across the L layers is

$$E = \frac{1}{2}\sum_{\ell=0}^{L-1} \left\| \frac{e_\ell}{\sigma_\ell} \right\|^2 , \tag{34}$$

where σ_ℓ and e_ℓ are vectors, and the division is taken component-wise. The error at layer ℓ is

$$e_\ell = r_\ell - \rho(W_\ell r_{\ell+1}) , \tag{35}$$

where, $r_0 = s$ is the stochastic sensory input vector, and r_L represents a given context vector at the very top layer L. The only source of stochasticity is the stimulus s, for which the top-down input, $\rho(W_0 r_1)$, is an estimate of its mean.

Dynamics for uncertainty-weighted prediction errors. We first consider the gradient dynamics $\dot{r}_\ell = -\frac{\partial E}{\partial r_\ell}$ for a fixed stimulus s, and assume that this converges to a critical point of E given by $\frac{\partial E}{\partial r_\ell} = 0$. In our simulations, s is sampled from a Gaussian with constant mean, or a mean that changes in steps across time. We calculate (assuming that the variances σ_ℓ^2 are constant),

$$\dot{r}_\ell = -\frac{\partial E}{\partial r_\ell} = W_{\ell-1}^T \left(\rho'(W_{\ell-1} r_\ell) \cdot \frac{e_{\ell-1}}{\sigma_{\ell-1}^2} \right) - \frac{e_l}{\sigma_l^2} , \tag{36}$$

where $\cdot$ denotes the component-wise product. To reveal the relation to our uncertainty-weighted prediction errors UPEs, we rewrite this equation as

$$\dot{r}_\ell = M_\ell \frac{e_{\ell-1}}{\sigma_{\ell-1}^2} - \frac{e_l}{\sigma_l^2} = M_\ell \, \mathrm{UPE}_{\ell-1} - \mathrm{UPE}_\ell , \tag{37}$$

with matrix M_ℓ mapping from the lower layer $\ell - 1$ to layer ℓ,

$$M_\ell = W_{\ell-1}^T \, \mathrm{diag}\big(\rho'(W_{\ell-1} r_\ell)\big) . \tag{38}$$

Eq. 37 is the general rate dynamics in a hierarchical predictive network. It represents a reformulation of the classical predictive coding model by Rao & Ballard [66], but with a noise model that is restricted to Gaussian noise only in the stimulus s. Upper layer representations r_ℓ will inherit the noise from the sensory input by a propagation of the stochastic error $e_0 = s - \rho(W_0 r_1)$ to higher layers ℓ. We also consider a fixed prior on the context rate r_L at the top layer L that is not included in the energy.

Going beyond [66], we show how a microcircuit can explicitly learn the uncertainties σ_ℓ^2, and how they can scale the error representations. Notice that the dynamics of Eq. 37 only contains divisions by σ^2 in the UPEs, and hence will lead to a simple neuronal implementation as shown in Fig. 8 for two layers.

From Eq. 37 we obtain the special case of Eq. 6 in the main text, i.e. $\dot{r} = (\mathrm{UPE}_s^+ - \mathrm{UPE}_s^-) - (\mathrm{UPE}_p^+ - \mathrm{UPE}_p^-)$, by choosing $\ell = 1$, M_ℓ the identity, and decomposing the uncertainty-modulated prediction error into the positive and negative part, $\mathrm{UPE}_\ell = \mathrm{UPE}_\ell^+ - \mathrm{UPE}_\ell^-$.

TODOOpposite of r_ℓ scaling by upper and lower uncertainties. We next give two expressions for the steady state rate reached at $\dot{r}_\ell = 0$ with a given stimulus s. The steady state condition $\dot{r}_\ell = -\frac{\partial E}{\partial r_\ell} = 0$ writes according to Eq. 37 as

$$M_\ell \frac{e_{\ell-1}}{\sigma_{\ell-1}^2} - \frac{e_l}{\sigma_l^2} = 0 \,, \tag{39}$$

where M_ℓ is defined in Eq. 38. To lighten the notation, we abbreviate

$$\hat{r}_\ell = \rho(W_\ell r_{\ell+1}) \,. \tag{40}$$

$\hat{r}_\ell$ is the top-down estimate of the representation r_ℓ, based on the prior $r_{\ell+1}$ in the upper layer. The error defined in Eq. 35 then writes as $e_l = r_\ell - \hat{r}_\ell$. Plugging this into Eq. 39, yields a self-consistency equation for r_ℓ

$$r_\ell = \hat{r}_\ell + \sigma_\ell^2 \cdot \Big(M_\ell \frac{e_{\ell-1}}{\sigma_{\ell-1}^2} \Big) \,. \tag{41}$$

Here, the divisive modulation of the error by the lower-layer uncertainty $\sigma_{\ell-1}^2$, and the multiplicative modulation by the upper-layer uncertainty σ_ℓ^2 becomes visible. This feature is shared by the full model of uncertainty-coding that considers independent Gaussian noise at each layer, and that takes the rate-dependency of σ_ℓ^2 into account (resulting in second-order error driving the dynamics of r_ℓ, see [29]). Crucially, again, the neuronal implementation of the dynamics in Eq. 37 only requires the division by $\sigma_{\ell-1}^2$ and σ_ℓ^2, see also Fig. 8.

Uncertainty representation as a convex combinations of rates. According to the dynamics in Eq. 37, the dynamics of the representation is given by a combination of bottom-up and top-down errors. The steady state is characterised by balanced errors, $M_\ell \,\mathrm{UPE}_{\ell-1} = \mathrm{UPE}_\ell$. We next show that also on the level of the rates, the steady state can be written as a combination of bottom-up and top-down rates.
For this, we introduce the bottom-up error-corrected representation $\check{r}_\ell$, which is the representation r_ℓ updated by the uncertainty-weighted error from the lower layer,

$$\check{r}_\ell = r_\ell + M_\ell \frac{e_{\ell-1}}{\sigma_{\ell-1}^2} \,. \tag{42}$$

From this we obtain $\check{r}_\ell - r_\ell = M_\ell \frac{e_{\ell-1}}{\sigma_{\ell-1}^2}$, while $r_\ell - \hat{r}_\ell = e_l$ is the error from Eq. 35. Plugging these two expressions into the dynamics underlying Eq. 39 yields

$$(\check{r}_\ell - r_\ell) + \frac{\hat{r}_\ell - r_\ell}{\sigma_l^2} = 0 \,. \tag{43}$$

This Eq. 43 now yields a self-consistency equation for r_ℓ, where the representation r_ℓ is a convex combination of the top-down prior $\hat{r}_\ell$ and the bottom-up update $\check{r}_\ell$,

$$r_\ell = \frac{1}{c}\Big(\frac{\hat{r}_\ell}{\sigma_l^2} + \check{r}_\ell \Big) \,. \tag{44}$$

Here, $c = \frac{1}{\sigma_\ell^2} + \mathbf{1}$ is the sum of the certainty vector $1/\sigma_\ell^2$ at layer ℓ and the vector $\mathbf{1}$ of $\dim(r_\ell)$ filled by 1's. Hence, the integration of the UPEs according to Eq. 37 converges to the convex combination of $\hat{r}_\ell$ and $\check{r}_\ell$ given by Eq. 44. This is the generalization of Eq. 4 in the main text.

Interpretation of the convex combination. Eq. 44 can be interpreted in different ways: It gives

- the posterior rate r_ℓ as convex combination of the top-down prior $\hat{r}_\ell$ (Eq. 42) and the bottom-up error-corrected representation $\check{r}_\ell$ (Eq. 40). Notice that 'prior' and 'posterior' here do not imply the classical Bayesian inversion since the noise at the various layers ℓ, inherited from the stochastic stimulus s, is not independent.
- a self-consistency equation for the stationary rate r_ℓ satisfying $\frac{\partial E}{\partial r_\ell} = 0$, with r_ℓ appearing also on the right-hand side of Eq. 44 via Eqs 42 and 38.
- an iterative scheme to calculate the steady-state representation r_ℓ, starting from the old value of r_ℓ on the right-hand side, and obtaining the new value on the left-hand side, if the iteration converges. The iteration typically converges due to the gradient descent construction.