nature communications

Article

https://doi.org/10.1038/s41467-026-70354-x

Global error signal guides local optimization in mismatch calculation

Received: 22 August 2025

Accepted: 24 February 2026

Published online: 12 March 2026

Check for updates

John Hongyu Meng & Xiao-Jing Wang

Corollary discharge denotes internal signals about the expected sensory consequences of one's own actions, leading to reduced sensory responses to self-generated stimulation. To investigate the underlying neural circuit mechanism, here we introduce a biologically plausible three-factor learning rule, where a global signal guides the updating of local inhibitory synapses to enable the computation of mismatches between stimulus and their predictions. We show that our network model, endowed with both positive and negative prediction error neurons, accounts for the salient physiological observations of motor-visual and motor-auditory mismatch responses in mice. Moreover, the model predicts that learning induces a bimodal distribution in activity correlation with stimulus and prediction, supported by our analysis of neural data from a recent experiment. Together, these results link global modulation to local learning for predictive error computation in the sensory areas, and predict how disrupting inhibition impairs mismatch computation in specific ways.

Since Hermann von Helmholtz, an outstanding question in Psychology and Neuroscience has been to elucidate how self-generated movement is monitored in the brain through an internal signal called corollary discharge or efference copy, that conveys (top-down) expectation of sensory consequences of one's own actions, to be compared with the actual stimulus induced (bottom-up) input in a sensory system[1] (Fig. 1a). For instance, we make rapid eye movements several times per second. Yet, our vision is stable thanks to corollary discharge associated with saccades[2]. When we walk, we do not perceive that the world is moving, even though the visual input to our eyes suggests motion[3]. Similarly, when playing an instrument in a band, we may tune out the sound of our own actions relative to those of others[4]. What neural mechanisms underlie this selective suppression? Answering this question is of interest for basic neuroscience as well as psychiatry, since deficits in corollary discharges are believed to be a root cause of hallucination and delusion associated with mental disorders[5,6].

In recent years, mismatch computation between a sensory response and corollary discharge of self-motion has been cast in the general framework of predictive coding[7]. Originally proposed for perception, predictive coding highlights the idea that a stimulus-induced response in a sensory circuit is compared with an internally generated expectation/prediction signal about that stimulus, and is suppressed when expectation is increased[8]. Applied to motor-sensory feedback, the motor cortex sends a copy of the movement signal to sensory areas, which serves as a local prediction of the incoming stimulus. If the prediction and actual input match, their signals cancel out, leading to reduced or no sensory perception (Fig. 1b). For example, we may press a play button and hear the expected tune. But when the prediction and stimulus mismatch, the sensory area responds, signaling a prediction error (Fig. 1c). This mismatch can occur in two ways: either a sound is heard without pressing the button, or we press the button while no sound follows. Both types of mismatch elicit stronger neural responses in mice than expected stimuli[9,10]. Moreover, suppression in the expected case depends on experience, suggesting a critical role for neural plasticity.

Though the mathematical description of mismatch is as simple as computing the absolute difference between stimulus and prediction signals, it is not clear how the brain implements this operation (Fig. 1d). One challenge is that neurons in sensory areas exhibit low baseline firing rates[11–13], making them highly responsive to excitatory input but relatively insensitive to inhibition. A straightforward solution is to introduce two separate neuronal populations: positive prediction error (pPE) neurons, which respond to stimulus minus prediction, and negative prediction error (nPE) neurons, which respond to prediction

Center for Neural Science, New York University, New York, NY, USA. e-mail: xjwang@nyu.edu

https://doi.org/10.1038/s41467-026-70354-x

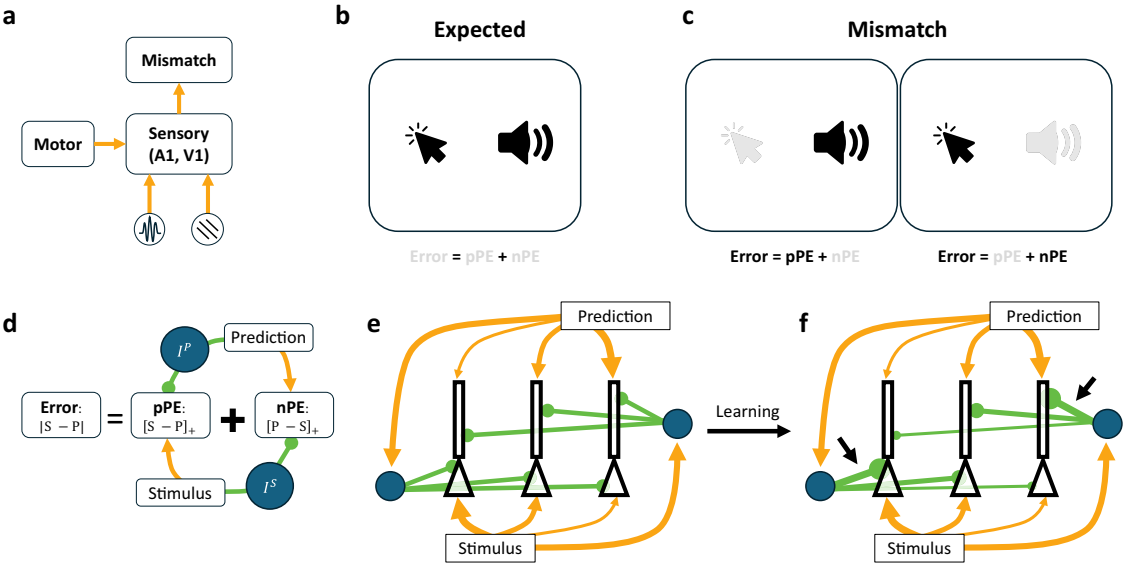

Fig. 1 | Motion-sensory mismatch calculation requires learning. a An efference copy of the motion signal is compared with the incoming auditory or visual stimulus to compute the mismatch. **b** A matched case where the incoming sound aligns with the clicking motion. **c** Mismatch examples: sound is heard without a corresponding motion, or no sound is heard when clicking the button. **d** The mathematical description of error calculation requires two distinct prediction error populations and local interneurons to invert long-range excitation into local inhibition (see main text). **e** Innate connectivity fails to compute errors effectively. **f** Optimal connectivity for error calculation requires inhibitory input from prediction (stimulus) to balance excitation from stimulus (prediction), a pattern that can be achieved through learning. Here, the diagrams in **e, f** assume that the total excitation from the stimulus and the prediction is identical across neurons to simplify the illustration, although this assumption is not required in theory.

minus stimulus[7]. Another challenge is that long-range projections in the brain are predominantly excitatory. Therefore, to compute a difference between stimulus and prediction signals, local interneurons must be involved to convert long-range excitatory inputs into local inhibition.

However, accurate prediction error computation also requires well-tuned local connectivity. With naive connectivity (Fig. 1e), stimulus and motor-related inputs can vary across individual neurons. Yet, these neurons receive dense inhibition from local interneuron populations[14], suggesting that each neuron receives a comparable level of inhibition. As a result, expected stimulus input cannot be effectively canceled by prediction-driven inhibition, and vice versa. This mismatch implies the need for a learning mechanism that tunes local inhibitory connections to accurately align inhibition with its corresponding excitatory drive, such that relayed stimulus inhibition targets prediction-driven excitation and relayed prediction inhibition targets stimulus-driven excitation. (Figure 1f). Notably, excitatory plasticity may not play a major role, as overall excitatory responses to motion or visual input are not significantly altered after learning[9]. Taken together, successful prediction error computation likely depends on inhibitory plasticity. This form of plasticity has been widely observed in cortex[15–17] and is modulated by neuromodulators[18]. In particular, noradrenaline signals broadcast by the locus coeruleus may facilitate local circuit optimization[19,20], though the underlying mechanism remains unknown.

In this study, we introduce a biologically plausible three-factor learning rule, inspired by ref. 21, that acts on inhibitory synapses from interneurons to pyramidal cells and enables local circuits to compute prediction errors. We show that this local rule converges to the same optimal connectivity as that learned via gradient descent. We test the rule in models composed of either simplified ReLU or biophysically realistic neuron models. The resulting inhibitory connectivity reproduces key experimental observations, including the selective motion-modulated auditory response[9] and the motion-visual mismatch response[10]. Moreover, we demonstrate that this learned connectivity is functionally optimal: perturbing interneuron activity, either by excitation or inhibition, consistently disrupts the population-level mismatch response. Finally, the model predicts a bimodal distribution of stimulus- and prediction-related correlation across neurons, a feature supported by reanalysis of experimental data.

Results

Three-factor learning rule optimizes prediction-error computation

We implemented a biologically feasible three-factor learning rule to align inhibition relayed from the prediction (or stimulus) with excitation driven by the stimulus (or prediction), such that population responses are suppressed when inputs are predicted, but enhanced during mismatches:

$$\Delta W \sim c(t)(R - R_0)r \tag{1}$$

This learning rule depends on the firing rates of the presynaptic neuron r and postsynaptic neurons R, along with a global third factor $c(t)$ representing the mismatch between prediction and stimulus. Inspired by recent studies[19,20], this third factor may be mediated by noradrenaline, released from the locus coeruleus. Remarkably, the proposed rule converges to the same optimal solution as the traditional gradient descent algorithm, despite relying only on biologically plausible global error signals. The mathematical proof of this equivalence, valid for any number of ReLU neurons, is presented in the Supplementary Notes. We further validate the theory through simulations in a three-dimensional setting (i.e., 3D movement in the home cage leads to a predictable 3D visual flow change) and by replacing ReLU neurons with a realistic conductance-based model. Here, we provide an intuitive explanation in a simplified case that only includes $n = 2$ RELU neurons to illustrate the necessary conditions required for our model.

This simplified case is designed to capture the minimal requirements for computing a prediction error. First, at least two excitatory

https://doi.org/10.1038/s41467-026-70354-x

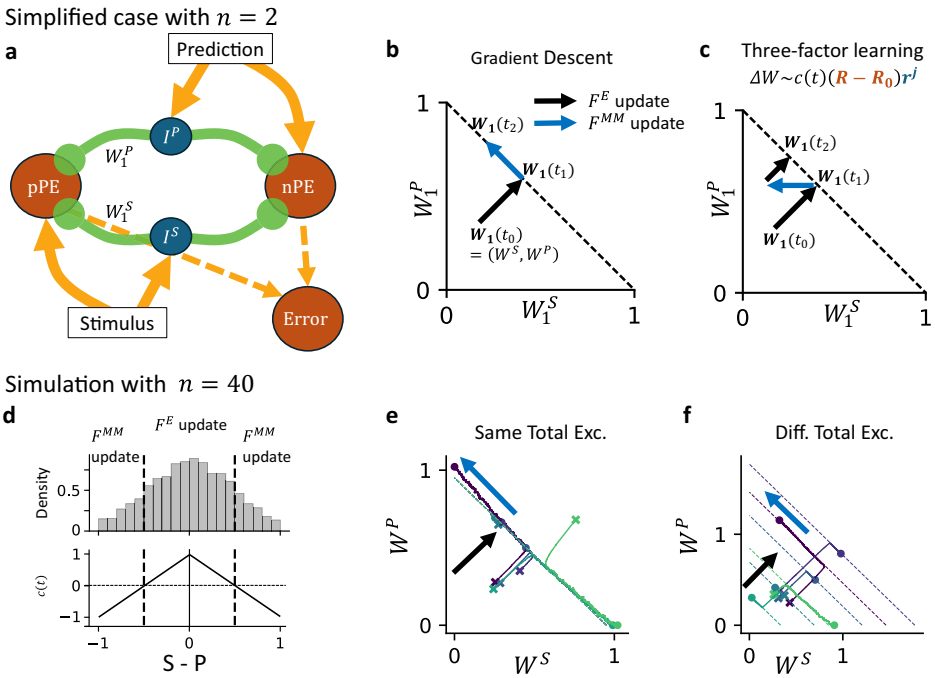

Fig. 2 | Mechanisms of the three-factor learning rule. a Diagram of network connectivity in a simplified case with $n = 2$, consisting of one nPE and one pPE neuron. **b** Weight updates under gradient descent. When trained on the expected set F^E, where prediction matches stimulus, weights converge to the slow learning manifold $W^S + W^P = 1$. Constrained by the slow learning manifold, training on the mismatch set F^{MM} shifts the weights along it. Eventually, inhibitory selectivity emerges as $(W^S, W^P) \to (0, 1)$. **c** Weight updates under the local three-factor learning rule. Inhibitory selectivity emerges through a zig-zag trajectory guided by the local rule. $c(t)$ represents the third factor. See "Methods" for details. **d** The sign of the third factor $c(t)$ depends on the true error. When the mismatch is small ($|S - P| < 0.5$), weights update as in the expected set F^E; when large ($|S - P| > 0.5$), weights update as in the mismatch set F^{MM}. (Top) The error distribution is drawn from a truncated Gaussian distribution. (Bottom) Corresponding third factor $c(t)$. **e** Weight evolution when total excitatory input from stimulus and prediction is the same across neurons. The dashed line indicates the theoretical slow-learning manifold. Crosses mark initial weights; circles mark final weights; colors mark different post-synaptic neurons. The black arrow highlights fast initial learning; the blue arrow indicates slower refinement along a slow learning manifold. **f** Same as **e**, but the total excitatory input from stimulus and prediction varies across neurons. In this case, the total inhibitory input after learning diverges across neurons.

neurons (n = 2) are needed, since a single neuron cannot encode bidirectional deviations from a low baseline firing rate. Likewise, two interneurons are included to represent distinct inhibitory populations. Second, one excitatory neuron is assumed to receive stronger sensory (stimulus-driven) input, whereas the other receives stronger top-down (prediction) input; an analogous arrangement applies to the two interneurons. Third, we retain only the dominant excitatory synapse onto each pyramidal neuron and each interneuron, neglecting weaker projections for simplicity. Fourth, all neurons are modeled as rectified linear units (ReLUs). These assumptions together define a minimal circuit consisting of two excitatory neurons (n = 2) and their corresponding interneurons (Fig. 2a). Last, we assume both neurons project equally to an output neuron whose activity encodes the difference between stimulus and prediction. The objective is that, through learning, these two neurons evolve into one pPE and one nPE neuron, as shown in (Fig. 2a).

We then trained the inhibitory connections from two interneurons I^S, I^P to both the pPE and nPE neurons to minimize the difference between the output neuron's activity and the true error, defined as the absolute difference between the stimulus and prediction. The optimal solution in this setting requires the connectivity configured as in Fig. 1d. In this case, the pPE neuron only receives relayed prediction inhibition but none of the relayed stimulus inhibition, and its firing rate reflects the positive prediction error $R^{\mathrm{pPE}} = [S - P]_+$. Conversely, the nPE neuron only receives relayed stimulus inhibition but not prediction inhibition, and its firing rate reflects the negative prediction error $R^{\mathrm{nPE}} = [P - S]_+$.

Learning how to compute prediction error accurately requires a model or an animal to use the prediction signal to cancel the incoming stimulus in the expected cases when both the prediction and stimulus signals are present (Expected set F^E: $|S - P| \approx 0$). However, this result may be from training in both the expected case (F^E and the mismatch set F^{MM}: $|S - P| > 0$). Recent experiments have suggested that the neuromodulation signal may be different in the F^E and F^{MM}[19,20]. To understand the updating difference in these two sets, we first examine the behavior of the gradient descent algorithm within the expected training set F^E. Within the expected training set F^E, the total inhibitory input converges to match the total excitation (Fig. 2b, black arrow). Yet, in this condition, it does not matter whether the inhibition originates from the prediction or the stimulus, resulting in a family of equally valid solutions rather than a unique one. This family of solutions is referred to as the slow learning manifold in the following. Indeed, the optimal response in the expected case is zero, indicating that excessive inhibition can suppress neural activity trivially, thereby satisfying the objective without yielding a meaningful solution. To resolve this ambiguity, training on the mismatch set F^{MM}, where stimulus and prediction inputs are not matched, is necessary. In this case, the source of inhibition becomes critical for accurate mismatch computation. As a result, the synaptic weights move along the slow learning manifold to the unique optimal solution (Fig. 2b, blue arrow), where the circuit correctly associates inhibition with its corresponding excitatory drive.

The same optimal solution emerges from our three-factor learning rule, assuming expected training samples are more frequent than

mismatched ones, as is typical in experimental paradigms. The third factor, $c(t)$, presumably mediated by a neuromodulator such as noradrenaline[19], controls the sign of learning. Indeed[22], demonstrated that different concentrations of neuromodulators could control whether paired pre- and post-synaptic activity led to long-term potentiation or depression. During training on expected samples, the rule behaves like Hebbian learning, with weight updates proportional to the product of pre- and post-synaptic activity, causing total inhibition to match excitation and thereby defining a slow-learning manifold. For mismatched samples, the rule becomes anti-Hebbian, with weight updates negatively correlated with the product of activity, driving weights toward the optimal configuration. Although inhibition temporarily deviates from the slow-learning manifold, the dominance of expected samples pulls total inhibition toward the manifold, keeping it close throughout learning. This alternation guides convergence along a zigzag trajectory (Fig. 2c).

Next, we test our algorithm with a network of $n = 40$ ReLU neurons. At each learning step, the true prediction error is sampled from a truncated Gaussian distribution (Fig. 2d), and the third factor $c(t)$ reflects the true error. Here, the third factor is only required to differentiate expected cases from mismatched ones by its sign, while the exact form may vary. Here, we choose a linear function for simplicity, while other choice of the functions works (see below). For analytical convenience, we first assume that the sum of stimulus and prediction inputs is constant across neurons. Under this assumption, different neurons share the same slow learning manifold, which is defined by the total excitatory input, and they can be ordered along a one-dimensional axis according to their input affinity to the stimulus (Fig. 2e). We further assume that the utility of mismatch signals saturates beyond a certain threshold (see "Methods"), which causes synaptic weights to different postsynaptic neurons to converge at distinct positions along the slow learning manifold. However, when considering two or more stimuli modulated by the same prediction signal, the sum of stimulus and prediction inputs will naturally differ across neurons, since enforcing identical sums would collapse the representation of distinct stimuli. Indeed, our learning rule still works, where the slow learning manifold differs for each neuron (Fig. 2f). The corresponding activity and synaptic weights across neurons are shown in Supplementary Fig. 1.

So far, we have shown that the proposed learning rule minimizes neuronal responses in the one-dimensional (1D) expected condition, where the prediction strength matches the stimulus strength (e.g., movement on a treadmill produces corresponding visual flow in one direction). However, animals operate in a three-dimensional (3D) environment, where free movement (e.g., a mouse exploring a cage) generates visual-flow changes along multiple axes. To test the generality of our learning rule, we extended the model to include three stimulus-prediction pairs, such that changes in each stimulus are canceled by changes in the corresponding prediction (Supplementary Fig. 2a; see "Methods"). All stimuli and predictions project to the same excitatory neuron pool, but their synaptic strengths are shuffled to remove correlations between any pair. We further assume that distinct stimuli and prediction signals are relayed through separate interneuron (IN) pools. This assumption is consistent with the reported IN selectivity[23,24] and/or the plasticity of excitatory synapses onto INs[25,26].

In this scenario, the expected condition is defined as the L2-norm of the error vector being below a threshold, while the mismatch condition includes all remaining cases (Supplementary Fig. 2b). Similarly, the sign of the third factor reverses between expected and mismatch conditions. As before, we sampled the prediction error at each learning step from a truncated 3D Gaussian distribution. After learning, we observed qualitatively similar outcomes: neural responses in the expected condition were suppressed, whereas those in the stimulus-only or prediction-only conditions were amplified (Supplementary Fig. 2c, d). This arises because the relayed inhibition aligns with the excitation of the same pair (e.g., Supplementary Fig. 2e, f). Similar cancellation patterns were observed for the other two stimulus-prediction pairs (Supplementary Fig. 2g, h). In summary, although limited experimental evidence supports this approach in high-dimensional cases as far as we know, our model offers a biologically plausible method for calculating high-dimensional prediction error.

Prediction response depends on context after learning

After establishing our learning rule, we examine neural responses to stimulus and prediction signals before and after learning. From this point on, we use more realistic models for both pyramidal neurons and interneurons (see "Methods"), though ReLU neurons yield similar results (Supplementary Fig. 1). Our pyramidal neuron model consists of two compartments and includes a dendritic voltage variable. With weak coupling between the somatic and dendritic compartments, the model captures both subtractive and divisive dendritic inhibition, depending on the level of dendritic depolarization (Supplementary Fig. 3). Here, we use the model with strong dendro-somatic coupling, reflecting a model of Layer 2/3 pyramidal neurons. We further select the single-neuron parameters based on data recorded from Layer 2/3 in the mouse primary visual cortex[27]. Neurons receive a baseline input that maintains low but nonzero firing rates. Each neuron receives input from both the stimulus and prediction signals, with varying strengths. They also receive inhibition that is relayed through local interneuron populations. Given the high density of local inhibitory connectivity[14], we assume that the strength of inhibition is similar across neurons. For simplicity, there is no recurrent connectivity among pyramidal neurons or from pyramidal neurons to interneurons in this model.

Before learning, the majority of neurons exhibit stronger responses when both the stimulus and prediction are present, compared to when only one input is provided (Fig. 3a). Here, we assume each neuron receives the same total input in the expected case, and sort neurons based on their affinity to the stimulus. At the population level, the response is highest when both inputs are present. Under our learning rule, inhibitory connectivity evolves from uniform inhibition into selective inhibition, such that excitatory input from the stimulus is canceled by inhibition relayed from the prediction (Fig. 3b, top), and excitatory input from the prediction is canceled by inhibition relayed from the stimulus (Fig. 3b, bottom).

After learning, population neuronal activity reflects prediction error computation (Fig. 3c). The left-most neuron receives the strongest stimulus excitation and the least inhibition from the stimulus, resulting in stronger responses when the stimulus is present alone. The right-most neuron shows a similar pattern when only the prediction input is provided. However, when both stimulus and prediction inputs are present, all neurons exhibit low firing rates, slightly above baseline. At the population level, the average response in the expected condition, when both stimulus and prediction inputs are present, is lower than the response to either input alone. Furthermore, comparing activity distributions before and after learning, the average activity decreases only in the expected condition (Fig. 3d). The model has similar qualitative results when each neuron receives a different total input in the expected case (Supplementary Fig. 4).

From another perspective, the net effect of the prediction signal on population responses becomes context-dependent after learning. Before learning, adding prediction increases activity, both with the stimulus (Fig. 3a, blue vs. gray) and without it (Fig. 3a, red vs. green). After learning, however, prediction reduces activity when comparing the expected case to the stimulus-only case (Fig. 3c, red vs. green), indicating that prediction-modulated inhibition is learned and context-specific. We will examine this further in the next section.

Here, the exact form of the third factor $c(t)$ is not important. To directly test this, we use a piecewise-estimated $\tilde{c}(t)$ instead to guide the learning and reach the same qualitative results (Supplementary Fig. 5). In this case, the $\tilde{c}(t)$ requires only an approximation of the global error by the animal.

https://doi.org/10.1038/s41467-026-70354-x

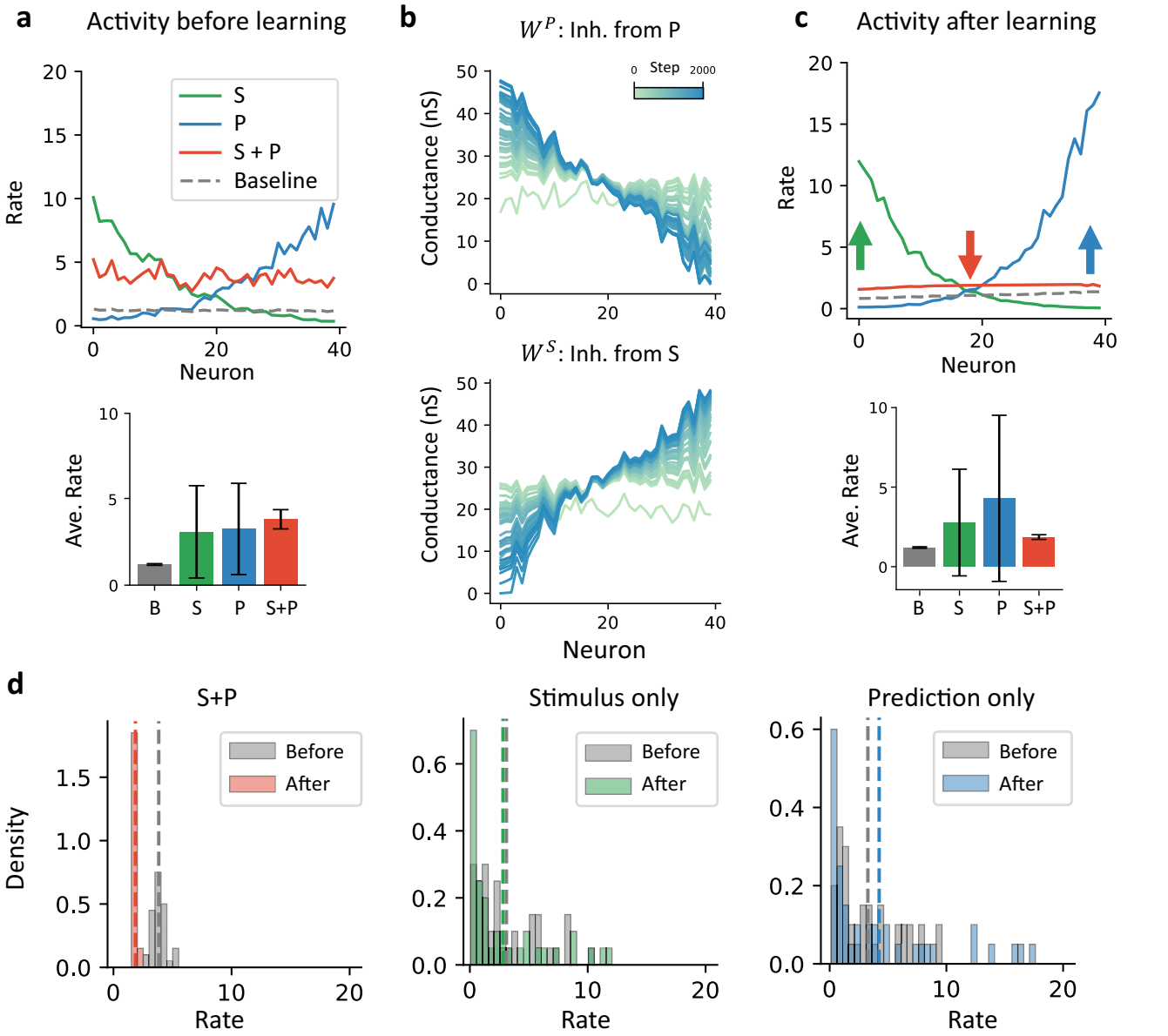

Fig. 3 | Neuron activity before and after learning. a Neuron activity before learning. The top panel shows the activity of individual neurons, sorted by the stimulus input strength. The line color indicates the type of input: Only stimulus, only prediction, stimulus and prediction, or neither (baseline). The bottom panel shows the bar chart of the average rate. The error bar indicates the standard deviation. **b** Connectivity weight that relays prediction inhibition W^P and that relays inhibition from stimulus inhibition W^S. Different colors represent synaptic weight during different training steps. A darker color indicates a later connectivity in the learning. **c** Neuron activity after learning. The activity is minimal in the expected case when both stimulus and prediction input are received, compared to the mismatched cases, when only stimulus or only prediction input is presented. **d** Activity distribution before and after learning across different inputs. Population activity only decreases when both prediction and stimulus are simultaneously presented.

Since the neurons on the left respond selectively to positive prediction error, $[S-P]_+$, after learning, we refer to them as pPE neurons. The same applies to the neurons on the right, which respond to negative prediction error, and are referred to as nPE neurons. Along the x-axis of input strength, the neuron identity gradually transitions from pPE (neuron 1 to neuron 20) to nPE (neuron 21 to neuron 40). In the following sections, the representative pPE and nPE neurons correspond to neuron 1 and neuron 40, respectively.

Prediction input selectively modulates the expected stimulus

In many situations, we selectively suppress expected stimuli in complex environments, while leaving the representation of other stimuli unchanged. Recent experiments have shown that this suppression is highly specific: only the response to the expected sound associated with self-generated motion is reduced, whereas responses to other, non-predicted sounds remain unaffected[10,28]. Does the prediction error computation learned by our model show the same specificity to expected stimuli?

We assess this selectivity by introducing a probed stimulus in addition to the expected stimulus. The excitation of the probed stimulus is generated by shuffling that of the expected stimulus. Importantly, the prediction excitation to each neuron is also shuffled, such that it is uncorrelated with both the excitation of the expected and the probed stimuli (Fig. 4a).

During training, as shown previously, the synaptic weights gradually align the inhibition relayed from the stimulus with the

https://doi.org/10.1038/s41467-026-70354-x

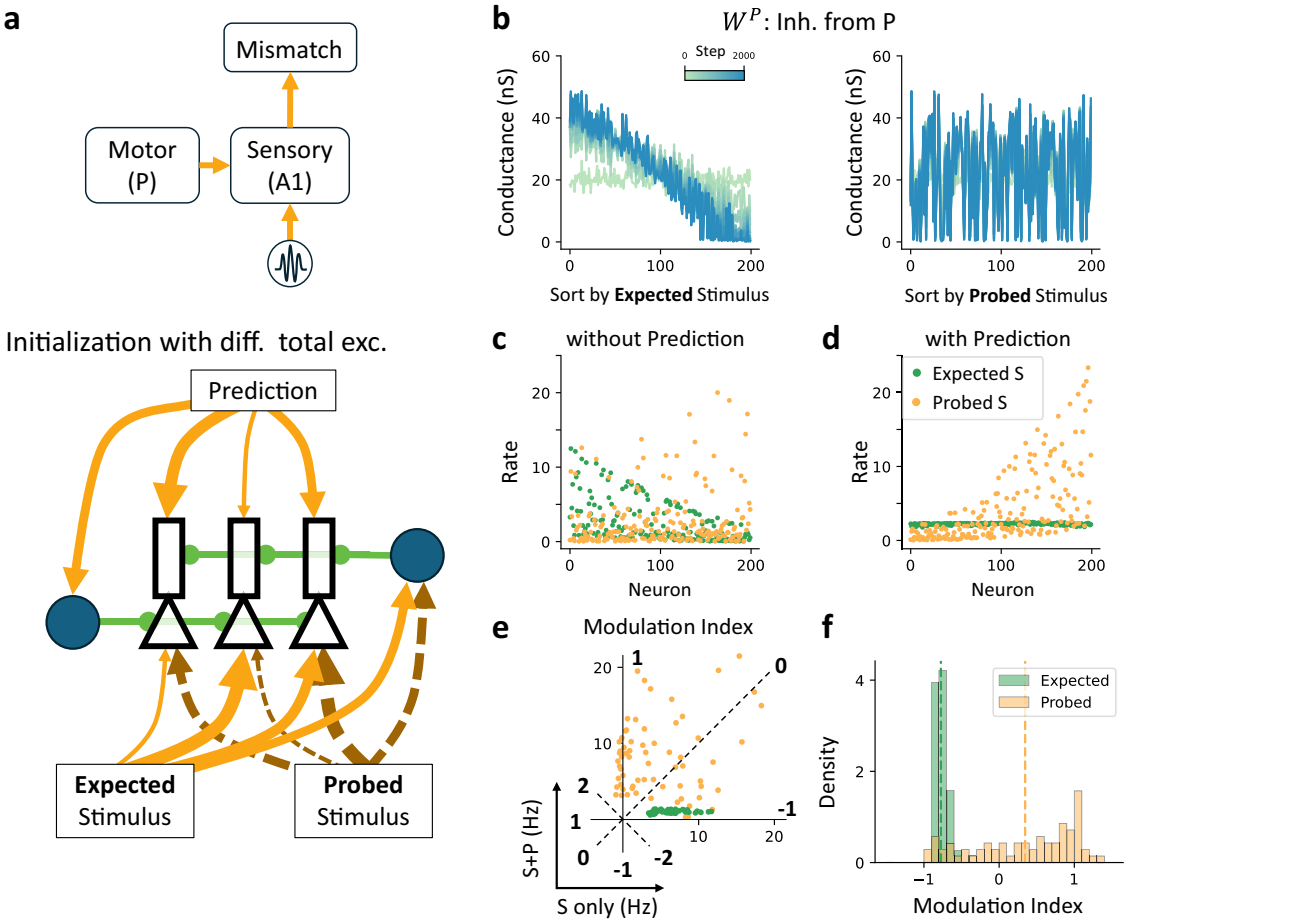

Fig. 4 | Selective suppression in the learned connectivity. a Schematic of stimulus-specific suppression. The top panel shows that A1 compares motor and auditory inputs to generate a stimulus-specific mismatch signal. The bottom panel shows the schematic of the model setup. During learning, an expected stimulus is presented. A probed stimulus is generated by shuffling the original stimulus inputs across neurons. In this case, the total excitation from prediction and either the expected stimulus or the probed stimulus is not the same across neurons. **b** Learned inhibitory connectivity (W^P) aligns with the expected stimulus. W^P is strongly aligned with neurons receiving stronger input from the expected stimulus, while no such alignment is observed when neurons are sorted by probed stimulus input. **c** Firing rates in response to expected or probed stimuli alone. Neuronal activity does not differ significantly between the two conditions when no prediction input is present. **d** Selective suppression emerges when prediction is added. The expected stimulus is selectively inhibited in the presence of its associated prediction, whereas the probed stimulus is not. **e** Quantifying suppression with a modulation index. The scatter plot of responses to stimulus alone (*x*-axis) vs. stimulus with prediction (*y*-axis) shows stronger suppression for the expected stimulus. **f** The distribution of the modulation index shows a peak near -1 for the expected stimulus, consistent with strong learned suppression. **e, f** agree with the selective suppression observed in refs. 10,28.

prediction input (Supplementary Fig. 6a). In contrast, inhibition relayed from the prediction aligns specifically with the expected stimulus, but not with the probed stimulus (Fig. 4b). This is because the error signal driving the third-factor learning rule reflects only the mismatch between the expected stimulus and the prediction, and does not incorporate the probed stimulus.

After learning, in the absence of prediction input, the response distributions to the expected and probed stimuli are similar (Fig. 4c). In contrast, when both stimulus and prediction inputs are present, strong inhibition emerges selectively for the expected stimulus, but not the probed stimulus (Fig. 4d, and Supplementary Fig. 6b).

To quantify the prediction-modulated response, we used the same modulation index (MI; see "Methods") as in refs. 10,28 to measure activity changes within the same neuron population (Fig. 4e). The MI for most neurons for the expected stimulus is close to minus one, indicating strong selective inhibition. In contrast, the MI for the probed stimulus shows a broader distribution centered above zero, consistent with weaker or absent modulation. This difference in MI between expected and probed stimuli was also observed experimentally[10,28]. One notable difference is that, in the experiments, the MI for the probed stimulus is closer to zero or slightly negative. This discrepancy may be explained by a global, non-specific inhibitory effect from motor to auditory cortex, potentially mediated by Parvalbumin-expressing (PV) interneurons[10].

It is also important for the sensory cortex to maintain accurate stimulus processing in the absence of prediction input. We find that discriminability, measured using a form of the Fisher information metric (see "Methods'), is comparable between the expected and probed stimuli (Supplementary Fig. 6c).

Model reproduces motor-visual mismatch response after learning

In the previous section, we examined how neural responses depended on whether a stimulus was predicted. Here, we investigate a complementary scenario in which the stimulus is absent when a motor signal is present (Fig. 5a). Recent studies[9,19] used a motor-visual mismatch protocol to explore neural responses when expected sensory input was omitted. In this paradigm, dark-reared mice were passively trained in a virtual environment where their movement on a treadmill controlled the visual flow of a grating pattern (coupled training, CT). In

https://doi.org/10.1038/s41467-026-70354-x

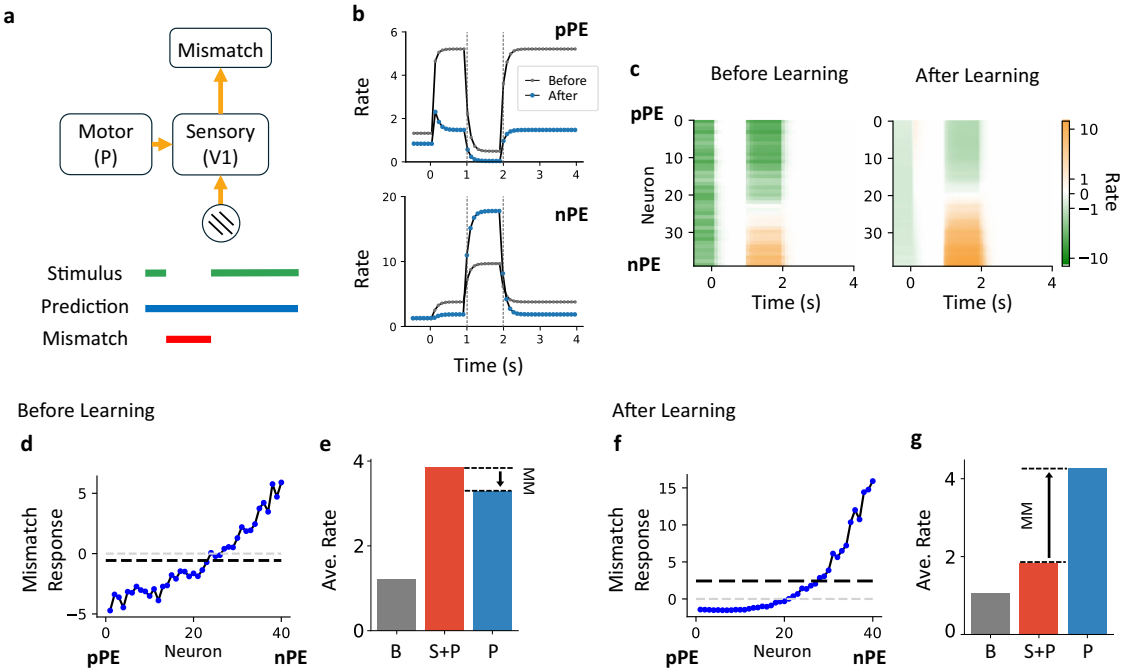

Fig. 5 | Three-factor learning captures responses in a motor-visual mismatch protocol. a Schematic of the mismatch paradigm. V1 compares motor-related and visual inputs to generate a surprise signal. The mismatch is introduced by pausing the flow of visual stimuli during locomotion. **b** Example neuron responses before and after learning. Firing rates over time for one pPE and one nPE neuron are shown before learning (black) and after learning (blue). **c** Learning-induced changes in firing rates. Firing rate changes across all neurons, comparing pre- and post-learning conditions. **d** Mismatch responses before learning across neurons. The black dashed line indicates the population average. **e** Population activity before learning. The averaged response in the expected case (S+P) is higher than the mismatch case (P only), suggesting no mismatch response at the population level. **f**, **g** Mismatch responses after learning. Same as (**d**, **e**). Here, a mismatch response is detected at the population level. The model behavior before learning and after learning reflects the behavior of the NT and CT groups in ref. 9, correspondingly.

parallel, a control group of mice received the same visual input generated from the CT animals, but the visual flow was decoupled from their own movement (non-coupled training, NT). As a result, the CT group may learn that the visual flow was a consequence of their movement, but not the NT group. After training, Mismatch events were introduced by abruptly pausing the visual flow during locomotion for both groups. In these experiments, a larger fraction of neurons in the CT group responded to the mismatch, and the average mismatch response was significantly stronger in the CT group compared to the NT group.

We simulate the same protocol in our model, both before and after learning. During the expected period, both prediction and stimulus inputs are active, whereas during the mismatch period, the stimulus input is paused while the prediction input remains. In our model, both nPE and pPE neurons exhibit reduced activity during the expected condition after learning (Fig. 5b), consistent with previous results (Fig. 3d). However, these two neuron types respond differently to the mismatch signal. pPE neurons show a negative mismatch response, indicating that they are normally excited by the stimulus. In contrast, nPE neurons show a positive mismatch response, suggesting that the stimulus inhibits their activity. At the population level, more neurons are excited than inhibited during mismatch (Fig. 5c). Before learning (Fig. 5d), the strongest excitatory and inhibitory mismatch responses at the single-neuron level are similar in magnitude. On average (Fig. 5e), responses in the expected condition (stimulus with prediction) are comparable to those in the mismatch condition (prediction only), indicating that mismatch signals are not readily distinguishable. After learning (Fig. 5f), however, the strongest excitatory mismatch responses (e.g., neuron 40) are ~ten times larger than the strongest inhibitory responses (e.g., neuron 1). This asymmetry arises from two factors: first, responses during the expected condition are strongly suppressed; second, nPE neurons become selectively inhibited by stimulus-driven inhibition, resulting in stronger disinhibition when stimulus input is absent.

Mismatch response under perturbation

If the learned connectivity is optimal in computing prediction error, then any perturbations of it are suboptimal and should do worse. To test this, we probed the model by perturbing the activity of interneuron populations, mimicking a photostimulation experiment (Fig. 6a).

By inhibiting the I^S population, the pyramidal neuron population becomes disinhibited (Fig. 6b). However, because inhibition from I^S is aligned with prediction-driven excitation, nPE neurons experience stronger disinhibition than pPE neurons. This asymmetry is reflected in the increasing activity across neurons along the neuron index, which is sorted by the stimulus affinity. As a result, nPE neurons no longer compute prediction error accurately: they exhibit elevated responses in the expected condition. Consequently, the change of population response stops reflecting the prediction error. In the opposite scenario, excitation of the I^S population leads to widespread inhibition across the pyramidal neuron population (Fig. 6c). Under this condition, nPE neurons are strongly inhibited in both the expected and mismatch cases. However, accurate mismatch computation requires nPE neurons to show increased activity in the mismatch condition, and this differential response is lost under excessive inhibition. As a result, nPE neurons fail to compute the prediction error accurately in this case as well. When plotting the mismatch response as a function of perturbation strength (Fig. 6d; additional examples shown in Supplementary Fig. 7a, b), we observe that performance is optimal at zero perturbation, indicating that any deviation impairs the network's ability to compute prediction error. This degradation in performance

https://doi.org/10.1038/s41467-026-70354-x

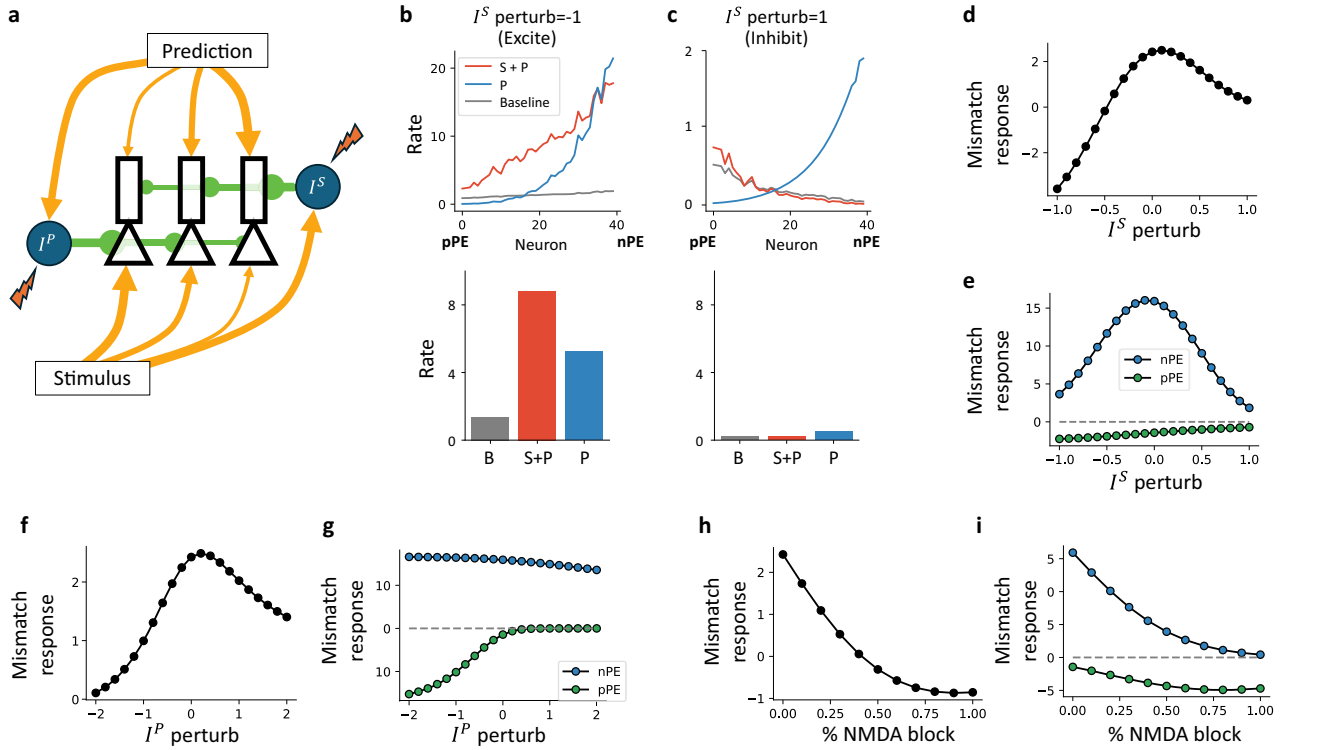

Fig. 6 | Population activity under perturbation. a Schematic of the task. **b** Neuronal activity when inhibiting the I^S population, shown at the level of individual neurons and population average across contexts. **c** Same as **b** but when exciting the I^S population. **d** Average mismatch response of the population across I^S perturbation strengths. **e** Mismatch response of a representative pPE neuron and nPE neuron across I^S perturbation strengths. **f**, **g** Same as **d**, **e**, but for perturbing the I^P population. **h**, **i** Same as **d**, **e** but for blocking NMDA receptors.

is primarily driven by nPE neurons, rather than pPE neurons (Fig. 6e), consistent with the fact that nPE neurons receive stronger inhibition from the I^S population.

In addition to perturbing the I^S population, perturbations of the I^P population (Fig. 6f) yield qualitatively similar outcomes. However, the specific mechanisms underlying performance degradation differ across conditions. When the I^P population is inhibited, pPE neurons become disinhibited in the expected condition (Fig. 6g, and Supplementary Fig. 7c), resulting in spurious responses that distort prediction error signaling. When the I^P population is excited, the example pPE and nPE neurons appear less affected (Fig. 6g), but nPE neurons that receive both prediction and stimulus inhibition (e.g., neurons 20–30) become excessively inhibited and cease contributing to the population-level prediction error signal (Supplementary Fig. 7d).

The impairment of predictive coding by NMDA receptor blockade is widely observed across species[29–31]. To test this in our model, we block NMDA channels on all excitatory connections. As expected, this impairs mismatch computation (Fig. 6h), due to reduced excitatory drive across neurons, leading to a general decrease in response amplitude (Fig. 6i, j; and Supplementary Fig. 7e, f). Despite the differing mechanisms, the same principle holds: any perturbation disrupts prediction error computation.

This concept of optimization can be further tested in a conjugated task, where a mismatch signal is introduced by pausing the prediction input in the model (Fig. 7a). Such a mismatch could arise experimentally through optogenetic suppression of axons projecting from motor areas to the visual cortex, thereby disrupting the prediction signal.

In contrast to the previous scenario, pPE neurons exhibit a strong positive response to the mismatch, while nPE neurons show a modest negative response (Fig. 7c). Across the neuronal population, the mismatch response gradually decreases from positive to negative values, although the average response remains positive (Fig. 7d). When examining mismatch responses as a function of perturbation strength, the population response is again optimized in the unperturbed condition: when neither the I^S nor the I^P populations, nor the NMDA channels, are manipulated (Fig. 7e, g, i). The mechanisms underlying performance degradation under perturbation are similar to those described earlier (Fig. 7f, h, j, and Supplementary Fig. 8). For example, inhibiting the I^P population leads to strong disinhibition of pPE neurons in the expected condition, reducing the difference between the expected and stimulus-only responses (Fig. 7f, and Supplementary Fig. 8a). As a result, the population-level mismatch response is diminished (Supplementary Fig. 8b). This mechanism is analogous to the effect of inhibiting the I^S population in the original mismatch task.

Bimodal correlation distribution emerges after learning

Our model predicts that, after learning, prediction-driven excitation should be balanced by stimulus-driven inhibition, and stimulus-driven excitation should be balanced by prediction-driven inhibition. These relationships between excitation and inhibition can be revealed by examining the correlation between neuronal responses and varying levels of stimulus and prediction input. Here, we focus on how these correlations change as a result of learning.

We first fix the prediction input strength while randomly varying the stimulus input each second (Fig. 8a). In our model, before learning, a subset of neurons exhibit stronger responses when both stimulus and prediction inputs are present, compared to either input alone (Fig. 3a; approximately from neuron 15–25). Indeed, two example neurons (Fig. 8b; neuron 15 and neuron 25) show positive correlations with stimulus input. After learning, however, almost no neurons maintain elevated responses when both inputs co-occur (Fig. 3c). After learning, neuron 25 switches from a positive to a negative correlation

https://doi.org/10.1038/s41467-026-70354-x

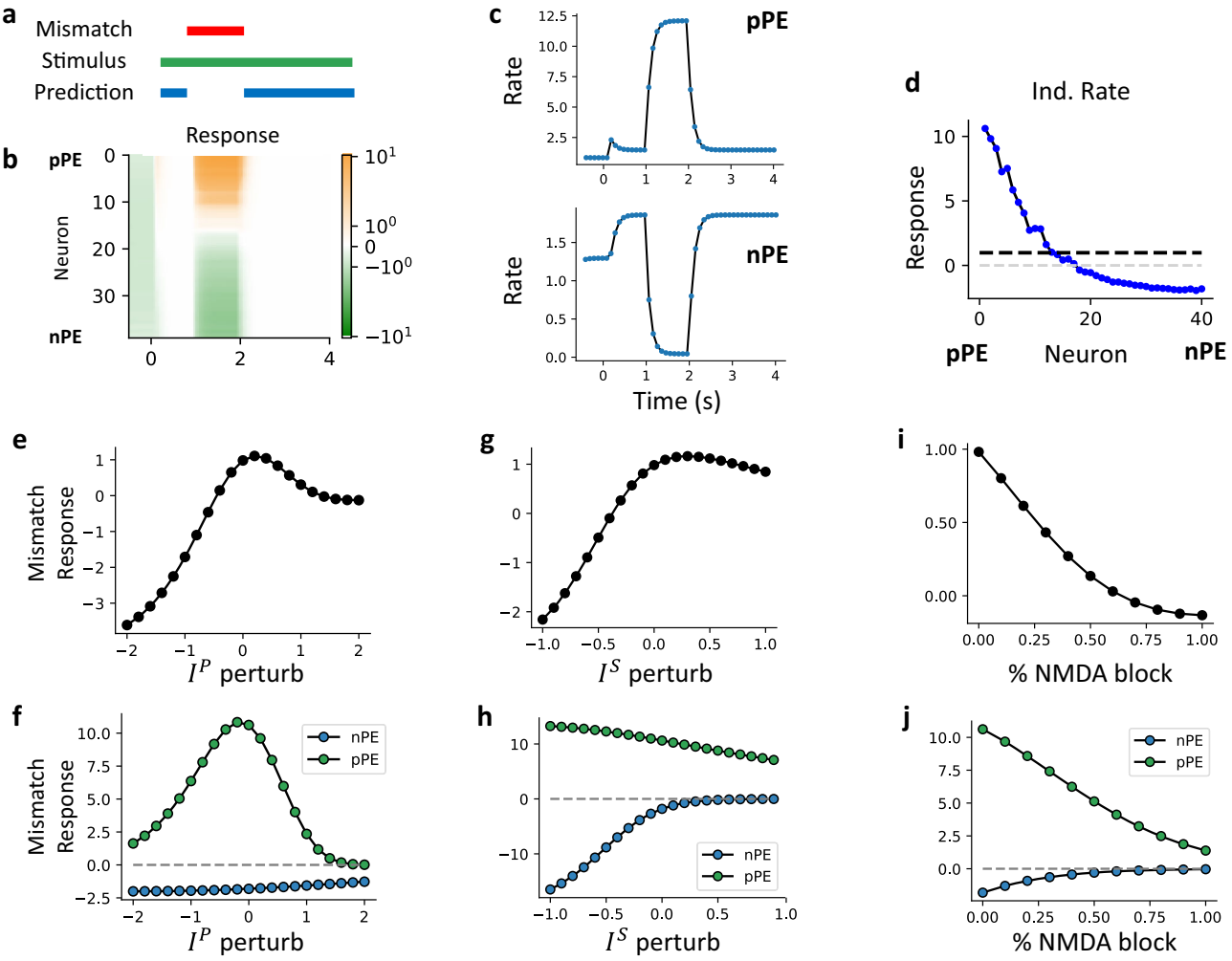

Fig. 7 | Behavior in a conjugated motor-visual mismatch task. a Schematic of the simulation. A mismatch is introduced by pausing the prediction signal in the model. **b** Neuronal response in the task. pPE neurons show a positive mismatch response, while nPE neurons exhibit a negative mismatch response. **c** Example neuron dynamics of a pPE and an nPE neuron. **d** Mismatch responses across individual neurons. The population average is shown as the dashed black line. **e** Population-averaged activity across I^P perturbation strength. **f** Mismatch response of a representative pPE and nPE neuron under I^P perturbation. **g**, **h** Same as **e**, **f** but for perturbing the I^S population. **i**, **j** Same as **e**, **f** but for blocking NMDA receptors.

with stimulus strength (Fig. 8b). We then reverse the setup by fixing the stimulus input while randomly varying the prediction input (Fig. 8c). Before learning, both example neurons again show positive correlations with the prediction strength. After learning, neuron 15 switches to a negative correlation with prediction strength (Fig. 8d).

We next vary the stimulus and prediction input strengths simultaneously and compute each neuron's response correlation with stimulus and prediction input separately. To improve statistical reliability, we use a larger model with $N = 200$ neurons and a continuous simulation duration of 200 seconds. In Fig. 8e, each dot represents a single neuron, with its correlation with stimulus input on the x-axis and correlation with prediction input on the y-axis. The color indicates the mismatch response after learning. The dots lie approximately on a unit circle, suggesting that the variance in neural responses can be fully explained by the combination of stimulus and prediction input.

We define the angular deviation θ for each neuron as the difference between the angle of its correlation vector in polar coordinates and $\pi/4$. A value of $\theta = 0$ indicates equal correlation with stimulus and prediction input. We further weight each θ by the radius of the corresponding point in polar coordinates and plot the weighted distribution of θ. Before learning, θ follows a unimodal distribution centered near zero (Fig. 8f). After learning, however, the correlation points cluster in the second and fourth quadrants, consistent with a functional split between nPE and pPE populations (Fig. 8g). This separation is further reflected in the emergence of a bimodal distribution in the weighted distribution of θ (Fig. 8h).

Finally, we test whether a similar bimodal correlation structure emerges in experimental data. Previously in Fig. 5, we showed that our model reproduces neuronal responses observed in both non-coupled training (NT) and coupled training (CT) mice during a motor-visual mismatch protocol[9]. Here, we analyze data from the same study, focusing on neuronal responses in NT and CT mice under the open-loop condition, where movement and visual flow are decoupled.

As in the model, we compute the correlation between neuronal responses and either visual flow (stimulus) or locomotion speed (prediction), and represent these in polar coordinates (Fig. 8i, k). We then plot the weighted distribution of the angular difference θ as defined earlier (Fig. 8j, l). In the NT group, the distribution of θ is unimodal and centered near zero, indicating similar correlation with stimulus and prediction inputs. In contrast, the CT group shows a bimodal distribution, suggesting that the functional separation into

https://doi.org/10.1038/s41467-026-70354-x

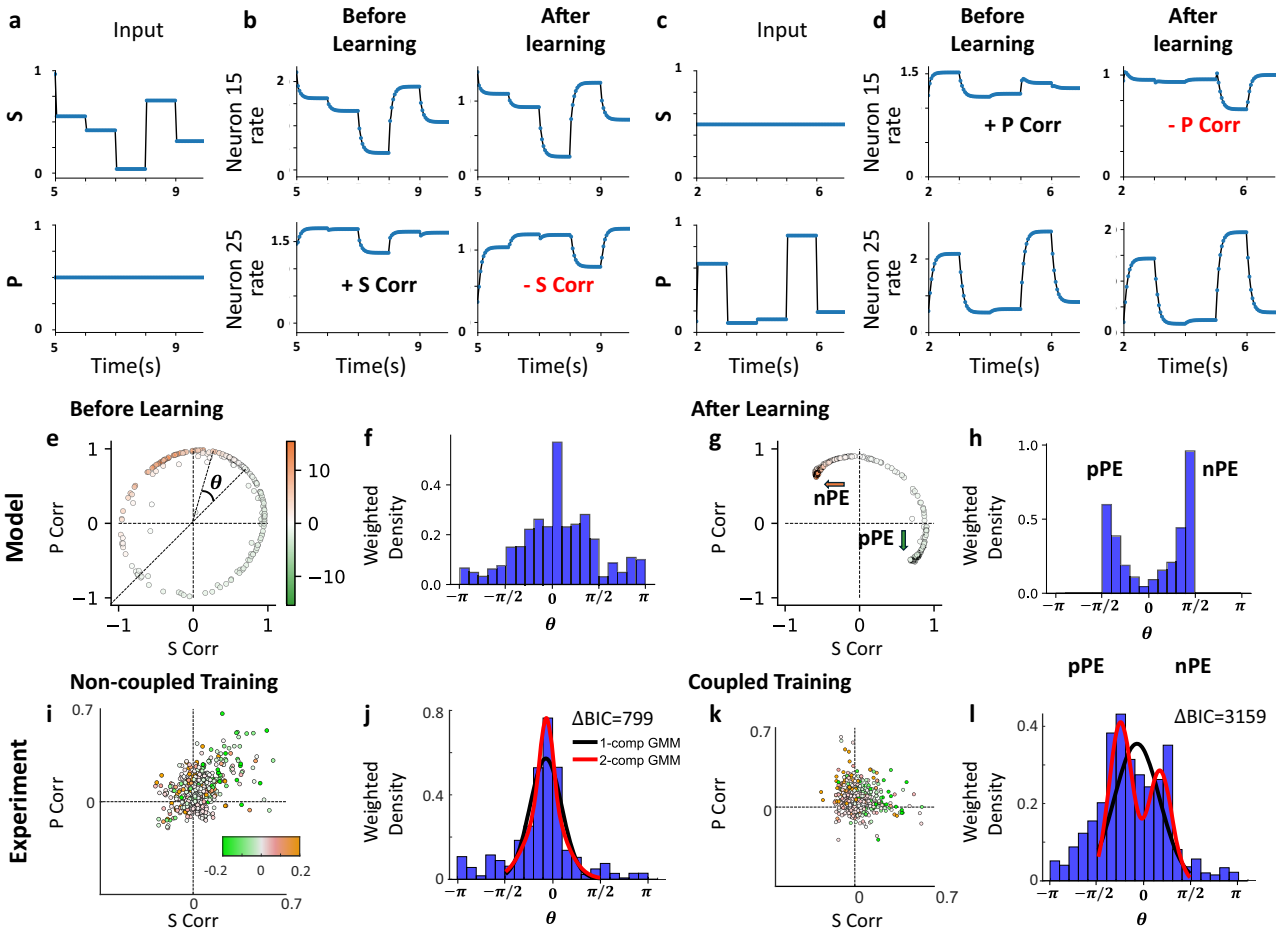

Fig. 8 | Bimodal distribution of correlation emerges after learning. a Stimulus input is varied across trials while prediction input is fixed. **b** Responses of two example neurons (Neuron 15 and 25) to the varied stimulus input, before and after learning. The firing rate of neuron 25 changes from positively to negatively correlated with stimulus strength after learning. **c** Reversed input conditions compared to (**a**) stimulus input is fixed (top), while prediction input is varied (bottom). **d** Responses of the same two neurons to varied prediction input before and after learning. The firing rate of neuron 15 changes from positive to negative correlation with prediction strength. **e** Correlation of firing rates with prediction and stimulus before learning across all neurons in the model. Here, both stimulus and prediction input are varied. The color scale represents the mismatch response of individual neurons after learning. **f** Distribution of the weighted angular difference between stimulus and prediction correlations (see "Methods"). A value of zero indicates identical correlation with both inputs. **g**, **h** Same as **e**, **f** but after learning. A value near $\pi/2$ indicates positive correlation with stimulus and negative correlation with prediction. The color for a neuron is the same in (**e**, **g**). **i**–**l** Reanalysis of experimental data from[9] shows a similar bimodal distribution in the coupled training condition, but not in the non-coupled condition. The color scale represents the z-scored firing rate of individual neurons. Here, the color represents the mismatch response for that group. N_{CT} = 939, N_{NT} = 690. The red and black line in (**j**, **l**) suggests the fitted Gaussian mixture model (GMM) with one or two components (see "methods" and Supplementary Fig. 9 for details). The difference of the Bayesian Information Criterion (ΔBIC) is used to determine which model is more favorable. ΔBIC > 10 suggests strong evidence that the 2-component GMM is more favorable[77]. Although fitting both CT and NT favors two-component GMMs, only the fitting of the CT dataset shows two distinct peaks in the fitted curve, indicating a true bimodal distribution.

pPE and nPE neurons emerges only with coupled sensorimotor experience, consistent with our model predictions. This bimodality is further confirmed by fitting the angular data to Gaussian mixture models (GMM) with one or two components (see "Methods"; Fig. 8j, l, Supplementary Fig. 9).

Discussion

Our work proposes that a three-factor learning rule optimizes local inhibitory connectivity for the computation of prediction errors in sensory areas. Each component of the rule is biologically supported, essential for the learning process, and converges to the same optimal solution as the gradient descent algorithm. After learning, positive prediction error (pPE) and negative prediction error (nPE) neurons emerge simultaneously, and their population activity reflects the prediction error: the difference between prediction and stimulus input. Because error is minimized only for stimuli that have been paired with a prediction signal, the model captures the selective modulation of responses observed in auditory cortex[10,28]. We reproduce the response in a motor-visual mismatch task[9] and further argue that any perturbation to the optimized circuit impairs performance. Finally, the model predicts a bimodal distribution of correlation with stimulus and prediction inputs after learning, a pattern that is confirmed by reanalysis of existing data.

The three-factor learning rule, also known as the neoHebbian learning rule, was first introduced by Gerstner and colleagues[21,32,33]. It extends traditional Hebbian plasticity[34] by incorporating a third factor in addition to pre- and postsynaptic activity. This third factor can represent signals such as reward, punishment, or surprise, and is often mediated by neuromodulators like dopamine[18,22,35,36]. In the three-factor framework, the pre- and postsynaptic activity sets a "tag" at the synapse, marking it as eligible for modification. The third factor can then arrive later (within a certain time window) to modify the actual synaptic weight using that tag. In doing so, three-factor learning bridges the timescales between fast neuronal activity (milliseconds)

and slower behavioral outcomes (seconds)[21,37]. Moreover, any biologically realistic learning mechanism must address the stability-plasticity dilemma[38], in which high plasticity leads to forgetting and high stability impairs new learning. The third factor has been proposed as a solution to this dilemma[39,40], by dynamically gating plasticity in a context-dependent manner.

In motor-sensory mismatch paradigms, animals are typically trained passively to associate specific motor outputs with corresponding sensory stimuli, thereby minimizing confounding top-down influences such as attention[9,10]. Because these tasks are not reward-based, the dopamine-dependent temporal-difference (TD) learning mechanism[41] is unlikely to be engaged. Recent findings show that locus coeruleus axonal activity correlates with the amplitude of unsigned prediction error signals, presumably reflecting noradrenaline release in sensory areas[19,20]. We therefore propose an alternative learning rule, in which the sign of synaptic weight change depends on the magnitude of the unsigned prediction error ($|S - P|$) rather than on a signed reward prediction error (RPE). In addition to noradrenaline, serotonin may serve a complementary modulatory role in guiding synaptic plasticity and local circuit optimization[42].

In our model, the three-factor learning rule is applied at local inhibitory synapses onto pyramidal cells, but not at the excitatory synapses onto interneurons. We omit the latter for three reasons. First, mismatch responses in interneurons do not appear to depend on experience, as reported in ref. 9. Second, successful balancing of excitation via relayed inhibition requires that interneurons be stimulus-selective, a condition that may not be biologically feasible given their dense and non-selective connectivity[14] (but see ref. 24). Third, optimizing excitatory synapses onto interneurons would require access to the firing rates of their downstream pyramidal targets, implying a non-local learning rule. Nevertheless, the potential role of plasticity of excitatory synapses onto interneurons may contribute to circuit flexibility in ways not addressed here.

By leveraging the framework of the three-factor learning rule, our model simplifies the assumptions required for the emergence of nPE and pPE neurons. First of all, including nPE neurons in the model is not intuitive, as their activity must be anti-correlated with stimulus strength while residing within sensory regions. Consequently, the classic predictive coding framework[8] and subsequent models[43–49] have primarily focused on a single pPE component, representing bidirectional error changes through both positive and negative firing rates. Only recently did Keller and colleagues demonstrate the emergence of nPE neurons after learning[9], highlighting the importance of incorporating both pPE and nPE populations in predictive coding models[7].

To our knowledge, the work from Hertäg and Sprekeler[50] is the first to propose a mechanism by which nPE and pPE neurons emerge simultaneously within the same circuit. In their model, co-emergence relies on establishing excitation-inhibition (EI) balance at both dendritic and somatic compartments through inhibitory plasticity. However, achieving this outcome depends on several specific assumptions. First, the model requires two distinct PV interneuron subtypes, each selectively receiving either stimulus or prediction input. Second, the learning scheme is not entirely local, raising questions about its biological feasibility. Third, dendritic input can only excite the connected soma but cannot exert inhibition. In contrast, we show that a biologically plausible, fully local three-factor learning rule is sufficient to reproduce the emergence of nPE and pPE neurons without these assumptions. In addition, only the sign of the third factor is critical for the emergence of nPE and pPE neurons in our model, whereas its amplitude plays no essential role, though it may facilitate rapid rule switching[51].

Our simplification can be understood as a shift in the objective of circuit optimization. The studies by refs. 50,52 propose that neural circuits aim to minimize overall firing rates to conserve energy, consistent with the original formulation of predictive coding[8]. In their framework[50], the error signal arises as a by-product of the circuit's limited experience with mismatched inputs, rather than as an explicitly learned feature. In contrast, we argue that the neuronal population is explicitly optimized to compute the prediction error with the correct amplitude in mismatched conditions, rather than merely suppressing activity in expected cases. Indeed, we demonstrate mathematically (see Supplementary Notes) that the optimal local connectivity can only be achieved through learning driven by error in mismatched conditions.

However, our model assumes that the animal can estimate a global error signal in some form, which could raise concerns about circular reasoning: if prediction error drives learning, how can it be computed before an accurate internal model exists? Yet, studies show that passive exposure is sufficient for animals to generate prediction error-like signals[9,10]. This suggests that animals may rely on approximate surprise signals, such as arousal responses to unexpected events[19], rather than precise prediction error estimates. Theoretically, this arousal signal could be derived by tracking variations in sensory input over longer timescales, such as across training sessions, and may serve as a latent predictive signal that guides the learning process[39]. Though such arousal signals may not represent the true global error, we show that an approximation is sufficient to guide local synaptic optimization (Supplementary Fig. 5). Still, the mechanisms by which animals generate these signals remain an important topic for future investigation.

Functionally, we argue that learning in sensory areas serves to reduce cognitive load, the mental resources required to perform a task, during prediction error computation[53,54]. For example, when first learning to ride a bicycle, one must consciously monitor every movement to avoid falling. With experience, however, bicycling becomes effortless, even on uneven terrain. Over time, the skill becomes automatic. Similarly, if visual and motor systems learn to compute prediction error locally, as our model proposes, this may offload computation from higher-order areas, such as the prefrontal cortex, and thereby conserve cognitive and metabolic resources. Supporting this view, an fMRI study of humans learning a visual-motion association task shows reduced activity in prefrontal regions in the later stages of learning compared to the early stages[55]. Future work should investigate whether similar signatures emerge in mice trained on predictive coding tasks.

To dissect the circuitry underlying predictive coding, it is essential to selectively target and perturb each component of the model. Recent studies suggest that Layer 2/3 *Rrad* and *Baz1a* pyramidal neurons are enriched for pPE responses, whereas *Adamts2* neurons preferentially exhibit nPE responses[56,57]. These transcriptomic subtypes may also possess distinct electrophysiological properties[58]. Consistent with this, we demonstrate mathematically that different initial conditions, defined by gene expression, lead to distinct pPE and nPE outcomes after learning (see Supplementary Notes), providing a potential mechanism linking transcriptomic identity to learned functional roles. Local interneurons are essential for converting long-range excitation into local inhibition. Among them, somatostatin-positive (SST) interneurons are the most likely mediators of stimulus-driven inhibition, as supported by studies of motion-visual mismatch responses[7,9,59].

However, the identity of the interneurons that relay prediction-driven inhibition remains unclear. In the auditory cortex, parvalbumin-positive (PV) interneurons have been implicated in mediating motion-related inhibition[60,61]. However, findings from[10] suggest that PV cells may instead implement a more uniform gain control mechanism, rather than conveying specific predictive signals. In contrast, SST interneurons have also been shown to mediate stimulus-selective inhibition[61], raising the possibility that they may also relay prediction-related inhibition. If so, the model would require functionally distinct subpopulations of SST cells to carry inhibition from different sources. Indeed, heterogeneous responses within the SST population (see

Fig. S3 in ref. 9) support this possibility. Another major interneuron subtype, vasoactive intestinal peptide-positive (VIP) cells, primarily inhibit SST interneurons and thereby disinhibit pyramidal neurons[62], making it less likely a candidate for mediating prediction-driven inhibition. However, recent findings from our group identify a VIP subpopulation that co-expresses the neuropeptide cholecystokinin (CCK) and targets pyramidal neurons directly, with comparable synaptic strength to that onto SST cells, and receives input directly from the motor cortex[63]. These VIP/CCK cells may serve as a candidate population for mediating prediction-driven inhibition.

To elucidate the functional role of a specific interneuron subtype, our model predicts distinct signature behaviors of I^S or I^P populations under targeted perturbations. These signatures can help determine the functional identity of a given interneuron subtype. Importantly, they are not detectable at the population level, since any perturbation suppresses the overall mismatch response (Fig. 6, Fig. 7), but must be identified at the single-cell level. In the motion-visual mismatch task, both exciting and inhibiting I^S are predicted to reduce the response of nPE neurons (Fig. 6e). In contrast, excitation of the I^P population increases the response of pPE neurons in the expected condition, thereby reducing the mismatch signal at the population level (Fig. 6g). Additionally, inhibition of the I^P population leads to a functional shrinkage of the nPE population (Supplementary Fig. 7c), though not necessarily a decrease in individual response amplitude. Similar patterns are expected in the conjugated task as well (Fig. 7, Supplementary Fig. 8). Future experiments can test these predictions using high-resolution genetic tools to target specific interneuron subtypes, combined with continuous rather than binary perturbation strategies.

In general, our theory provides a unifying framework for any instance in which functionally relevant signals are compared across cortical areas, consistent with ref. 7. Indeed, similar nPE and pPE response patterns have been observed in the visual cortex during audio-visual mismatch[64] and in the auditory cortex during motor-audio mismatch[65]. Notably, in the audio-visual paradigm, the auditory input, serving as the predictive signal, only suppresses visual responses after learning[64], further supporting the idea that prediction-driven inhibition becomes aligned with stimulus-driven excitation through plasticity. Moreover, the functional segregation of suppressive and enhanced neural responses in the auditory cortex of vocalizing marmosets[66] can also be accounted for by the emergence of pPE and nPE neurons in our model. It is worth noting that self-generated sounds can change over time, requiring continuous plasticity to maintain accurate cancellation over age. These observations suggest that our model offers a unified framework for understanding one aspect of corollary discharge[1,67], wherein motor-related signals are sent to sensory areas to suppress predictable reactions. A companion study[68] addresses a separate question, extending the current framework to cases in which prediction is driven by stimulus history rather than motor-related signals, and examines additional phenomena such as mismatch negativity[69].

Methods

Simplified neural circuit

We begin with a simplified neural circuit with ReLU neurons to illustrate the mechanism of our three-factor learning rule. Here, all the variables are dimensionless and are updated instantaneously without temporal dynamics.

For the i-th ReLU unit, activity is given by:

$$R_i = [\lambda_i^S u^S + \lambda_i^P u^P - W_i^S u^S - W_i^P u^P]_+. \tag{2}$$

For simplicity, we omit the subindex i in the following. Within the above equation, λ^S and λ^P denote the excitatory input strengths from stimulus and prediction, respectively. The stimulus and prediction input are denoted by u_S and u_P, or simply S and P when unambiguous. W^S and W^P represent the inhibitory synaptic weights from interneuron populations I^S and I^P, whose activity is the same as the stimulus and prediction input u_S and u_P. The operator $[x]_+$ denotes a rectifier (ReLU) function, where $[x]_+ = \max(0, x)$. The synaptic weights W^S and W^P are plastic and updated according to our three-factor learning rule.

In the case of homeostasis, $\lambda^S + \lambda^P = 1$. Thus, for i-th neuron, we can write $\lambda_i^S = \lambda_i$ and $\lambda_i^P = 1 - \lambda_i$. In the case of two neurons $N_{neu} = 2$, we set $\lambda_1 = 1, \lambda_2 = 0$. In the case of forty neurons $N_{neu} = 40$, the values of λ_i are linearly spaced from 1 to 0 across the population.

In the case without homeostasis, we shuffled the λ_i^P across neurons such that the total input $\lambda_i^S + \lambda_i^P$ is no longer constrained to equal 1.

Three-factor learning rule

Our learning mechanism can optimize the mismatch calculation, as proven in the Supplementary Notes.

The inhibitory connections are initiated with $W_i^j(t_0)$ for both interneuron populations $j = S, P$ with some jitter, mimicking the average synaptic strength from a population that should be similar across post-synaptic neurons. Again, we omit the subscript i in the following derivations.

These inhibitory synapses are updated with a three-factor learning rule as follows:

$$\Delta W^j = \alpha c(t)(R - R_0) r^j, \tag{3}$$

where $j = S, P$. Here, R denotes the activity of the postsynaptic pyramidal neuron, and r^j represents the steady-state activity of the presynaptic interneuron. R_0 is a small target firing rate shared by all neurons, and α is the learning rate.

The third factor, $c(t)$, is defined as:

$$c(t) = c_0 - |u^S - u^P|, \tag{4}$$

such that when $u^S = u^P$, $c(t) = 1$, and when $|u^S - u^P| = 1$, $c(t) = -1$. The threshold c_0 of 0.5 is chosen for simplicity. Also, our choice of c_0 implies that a rough estimation of the prediction error is sufficient to serve as the third factor, since any input with $c(t) > 0$ is considered as an expected case.

We further show that an estimated piece-wise third factor is sufficient to guide the local learning. In this case,

$$\tilde{c}(t) = 1, \quad \text{if} \quad |u^S - u^P| < 0.2; \tag{5}$$

$$\tilde{c}(t) = -1, \quad \text{if} \quad |u^S - u^P| > 0.8; \tag{6}$$

$$\tilde{c}(t) = 0, \quad \text{others}. \tag{7}$$

During training, $\epsilon = u^S - u^P$ is drawn from a truncated normal distribution (Fig. 2d) with mean zero and standard deviation $\sigma_\epsilon = 0.5$, reflecting the assumption that training samples are dominated by expected cases where $u^S \approx u^P$.

The mismatch modulation saturates when the mismatch signal becomes sufficiently large, and is defined as:

$$c(t)^{\text{eff}} = c(t)\left[1 - \frac{\bar{R}}{\kappa R_0}\right]_+, \quad \text{if} \quad c(t) < 0, \tag{8}$$

where κ is a scaling factor that sets the saturation level of the mismatch utility, and $\bar{R}$ is the average activity of the neuron population.

In the case of realistic neuron models, synaptic weight updates are further capped at high conductance values:

$$\Delta W_j^{\text{eff}} = \Delta W_j \cdot \frac{\eta W^0 - W_j}{\eta - 1}, \tag{9}$$

where $\eta > 1$ is the saturation factor for synaptic weights, and W^0 is the initial averaged synaptic strength.

Generalization in the 3-dimension

We next tested whether the proposed learning rule generalizes to a scenario in which the model need to use a three-dimensional (3D) prediction signal to cancel a 3D stimulus input. The simulation results are shown in Supplementary Fig. 2.

In this case, the i-th ReLU unit, activity is given by:

$$R_i = \left[\sum_{j=1}^{3}(\lambda_i^{S_j} u^{S_j} + \lambda_i^{P_j} u^{P_j} - W_i^{S_j} u^{S_j} - W_i^{P_j} u^{P_j})\right]_+. \tag{10}$$

For each timestep, the maximum of the pair $\nu_j = \max(u^{S_j}, u^{P_j})$ is randomly drawn from a linear distribution from 0 to 1. Then, the error for the pair $\epsilon^j = u^{S_j} - u^{P_j}$ is drawn from a truncated Gaussian distribution with mean zero and standard deviation $\sigma_\epsilon = 0.4$ and further scaled by the maximum input ν_j. The value of u^{S_j} and u^{P_j} is calculated based on the ν_j and the error ϵ_j. Each stimulus-prediction pair is generated independently. Please note that ϵ^j is only used in generating the training samples, while the model has no access to the error signal in each dimension.

The third factor remains as a scaler, which is calculated based on the L2-norm of the difference between the 3D-simulus and the 3D-prediction at each time step:

$$\epsilon = ||(u^{S_1}, u^{S_2}, u^{S_3}) - (u^{P_1}, u^{P_2}, u^{P_3})||_2, \tag{11}$$

$$c(t) = c_0 - \epsilon, \tag{12}$$

with $c_0 = 0.3$.

Pyramidal neuron and interneuron models

In this work, we use a simplified hybrid dendritic model. For each pyramidal neuron, the dendritic membrane potential V^d and somatic firing rate R^s are updated at each time step as follows:

$$C^d \frac{dV^d}{dt} = I^d + I^{d,\text{leak}} - g^c(V^d - V^s) + I^{\text{bAP}}, \tag{13}$$

$$\tau^s \frac{dR^s}{dt} = -R^s + f(I^s - g^c(V^s - V^d)), \tag{14}$$

where I^s and I^d are the total synaptic currents into the soma and dendrite compartments, respectively. The dendritic leak current is given by $I^{d,\text{leak}} = -g^{d,\text{leak}}(V^d - V^{d,\text{rest}})$, where $V^{d,\text{rest}}$ is the resting dendritic potential and $g^{d,\text{leak}}$ is the leak conductance. The soma and dendrite are coupled via conductance g^c, resulting in a coupling current $I^{sd} = g^c(V^d - V^s)$.

The dendritic compartment also receives input from back-propagating action potentials (bAPs), which we model as a rate-averaged current following[70]: $I^{\text{bAP}} = -g^c(V^d - V^{\text{bAP}})t^{\text{bAP}}R^s$, where V^{bAP} is the averaged membrane potential during a spike, t^{bAP} is the duration of the bAP, and R^s is the firing rate at the soma.

As observed in ref. 70, the membrane potential fluctuations at the soma are considerably smaller (10 mV) than those at the dendrite (50 mV). To simplify the computation of the coupling current between somatic and dendritic compartments, we approximate the real somatic potential in simulations by the time-averaged somatic potential between action potentials, defined as $V^{s,\text{eff}} \equiv V^{s0} = \langle V^s \rangle_t$.

The activation function f that converts somatic input current to firing rate is taken from[71]:

$$f(I) = \frac{\Delta V}{\tau(V_{\text{th}} - V_{\text{reset}})\left(1 - \exp\left(-\frac{a\Delta V}{\sigma_V}\right)\right)}, \tag{15}$$

where $\Delta V = I/g_L + V_l - V_{th}$.

The model parameters are selected based on prior experimental work[70,72]. The coupling conductance g_c is set to a small value for neurons with extensive dendritic trees, such as Layer 5 pyramidal cells, where our single-neuron model reproduces shunting dendritic inhibition (Supplementary Fig. 3a). In contrast, g_c is set higher for neurons with more compact dendritic arbors, such as Layer 2/3 pyramidal neurons, where dendrites are closer to the soma and dendritic inhibition exhibits primarily subtractive effects (Supplementary Fig. 3b).

We use a point-neuron model to describe the average activity of each interneuron population:

$$\tau \frac{dr^j}{dt} = -r^j + f(I), \tag{16}$$

where $j = S, P$ denotes the stimulus- or prediction-driven interneuron population, and $f(I)$ is the same activation function defined previously.

Synapse model

We use a conductance-based synaptic model in our simulations, following[73]. The AMPA and GABA synapses are described by:

$$I^{\text{syn}} = -g^{\text{syn}} s^{\text{syn}}(V - V^{\text{syn}}), \tag{17}$$

$$\frac{ds^{\text{syn}}}{dt} = -\frac{s^{\text{syn}}}{\tau^{\text{syn}}} + R^{\text{pre}}, \tag{18}$$

where syn = A, G denotes AMPA and GABA synapses, respectively. Here, g^{syn} is the maximal synaptic conductance, s^{syn} is the gating variable representing the fraction of open channels, V^{syn} is the synaptic reversal potential (determined by whether the synapse is excitatory or inhibitory), τ^{syn} is the synaptic time constant, and R^{pre} is the presynaptic firing rate.

The NMDA synapse model includes both a voltage-dependent magnesium block, $f^{\text{Mg}}(V)$, and a saturating gating variable, s^N:

$$I^N = -g^N s^N (V - V^E) f^{\text{Mg}}(V), \tag{19}$$

$$f^{\text{Mg}}(V) = \frac{1}{1 + \exp(-0.062V)/3.57}, \tag{20}$$

$$\frac{ds^N}{dt} = -\frac{s^N}{\tau^N} + (1 - s^N)\gamma R^{\text{pre}}. \tag{21}$$

Here, g^N is the maximal NMDA conductance, V^E is the excitatory reversal potential, τ^N is the NMDA time constant, γ is a saturation scaling factor, and R^{pre} is the presynaptic firing rate. The function $f^{\text{Mg}}(V)$ captures the voltage-dependent magnesium block characteristic of NMDA receptors.

The stimulus input to both pyramidal neurons and the first interneuron population I^s is modeled as an excitatory conductance to the soma, denoted by g_E^{soma}. When illustrating the behavior of individual pyramidal neurons in Supplementary Fig. 3, the dendritic compartment additionally receives excitatory and inhibitory input conductances, denoted by $g_{I,E}^{\text{dend}}$. Because our model employs

conductance-based synapses with distinct reversal potentials for excitation and inhibition, the total input current is not simply a linear difference between excitatory and inhibitory conductances. This formulation allows us to capture more complex and dynamic behaviors in the circuit.

Input and network connectivity

We include only the necessary connections required for the model. A schematic of the network architecture is shown in Fig. 1e. The maximum synaptic conductance from interneuron population j to pyramidal neuron i is denoted by W_i^j, where $j = S, P$ corresponds to the stimulus- and prediction-driven interneuron populations, respectively.

In the simulations, synapses originating from the same presynaptic interneuron population share a common gating variable, denoted by s_j. The total inhibitory conductance received by a postsynaptic pyramidal neuron i from both interneuron populations is given by:

$$g_i = \sum_{j\in S,P} W_i^j s_j, \tag{22}$$

where s_j is the gating variable for population j.

Our model captures pyramidal neuron activity at both the single-cell and population levels. The network consists of N_{neu} pyramidal neurons, with two distinct interneuron populations relaying stimulus- and prediction-driven inhibition, respectively.

In the case where the sum of stimulus and prediction inputs is the same across neurons, the excitatory synaptic inputs from the stimulus and prediction are anti-correlated. The stimulus input strength is linearly spaced from a maximum value $g^S_{\max}$ to zero. Thus, for the i-th neuron, the stimulus input strength is defined as:

$$g_i^S = g^S_{max}\frac{N_{neu} - i}{N_{neu}} \tag{23}$$

With input strength $u_S \in [0, 1]$, the total excitatory current from the stimulus is:

$$I_i^S = -u_S g_i^S (V^{s0} - V^E). \tag{24}$$

Similarly, the stimulus input to the I^S interneuron population is defined by the synaptic strength parameter g^{I^S}. In the perturbation experiments, the perturbation level is normalized by this baseline input strength. A perturbation value of I^S perturb $= -1$ indicates that the I^S population receives no excitation when the stimulus is presented, whereas I^S perturb $= 1$ indicates that the I^S population receives double the baseline excitation.

The top-down input is modeled as a presynaptic predictor neuron firing at a rate of $5(1 + u_P)$ Hz, where $u_P \in [0, 1]$ represents the prediction strength. This yields a predictor firing rate in the range [5, 10] Hz. The synaptic weights from the predictor neuron to the dendritic compartments of pyramidal neurons are linearly spaced from zero to a maximum value $g^{Ed}_{\max}$. For the i-th neuron, the top-down connectivity strength is given by $g_i^{Ed} = g^{Ed}_{max}\frac{i}{N_{neu}}$. Furthermore, the excitatory synaptic input is composed of a mixture of NMDA and AMPA receptors, with a fraction κ mediated by NMDA synapses and the remaining $(1 - \kappa)$ by AMPA synapses.

The input conductance is computed by substituting the predictor neuron's firing rate into Equations (18) and (21). Specifically, the dynamics of the AMPA and NMDA gating variables become:

$$\frac{ds^A}{dt} = -\frac{s^A}{\tau^A} + 5(1 + u_P), \tag{25}$$

and

$$\frac{ds^N}{dt} = -\frac{s^N}{\tau^N} + (1 - s^N)\gamma \cdot 5(1 + u_P). \tag{26}$$

The input current to pyramidal neuron i depends on both the gating variables s_i^{syn} and the dendritic membrane potential V_i^d:

$$I_i^{Ed} = -g_i^A s^A (V_i^d - V^E) - g_i^N s^N (V_i^d - V^E) f^{\text{Mg}}(V_i^d), \tag{27}$$

where $g_i^A = \kappa g_i^{Ed}$ and $g_i^N = (1 - \kappa) g_i^{Ed}$.

Similarly, the prediction input to the I^P population is defined by the input strength g^{I^P}. In the perturbation experiments, the degree of perturbation is normalized by the input strength to the I^S population, g^{I^S}, so that equivalent perturbation magnitudes are comparable across the I^S and I^P populations.

We further include background stimulus conductance to all neurons in the model, ensuring that each cell maintains a similar low, but non-zero, baseline firing rate.

In the realistic model, we shuffle the synaptic weights from the predictor neuron to pyramidal cells, such that the input strengths from the stimulus and prediction are uncorrelated. Naturally, the sum of stimulus and prediction inputs is different across neurons.

We do not include lateral connectivity between pyramidal neurons in this model. While lateral inhibition may aid in differentiating responses between nPE and pPE neurons, it cannot contribute to canceling stimulus or prediction inputs in the expected condition, where activity is low across the population. For simplicity, we omit these connections from the model.

All parameters used in this study are listed in the Supplementary Table. All code will be available on GitHub upon publication acceptance.

Data analysis

In the motion-modulated auditory response analysis (Fig. 3), we quantify the selective modulation effect of the prediction signal using the same modulation index (MI) as defined in ref. 10. If the prediction increases the response to the stimulus, the MI is positive; if it suppresses the response, the MI is negative. Specifically, for each cell, let r_s denote the response to the unmodulated stimulus and r_m the response to the modulated stimulus. We compute the angular deviation from the diagonal (indifference) line as: $\theta = \arctan(r_m/r_s) - \pi/4$ The resulting $\theta \in [-\pi, \pi]$ is linearly rescaled to the range $[-2, 2]$ for visualization clarity. An MI of -1 indicates that the response to the modulated stimulus is zero. Only cells with responses greater than $2 \times (1 + 1.28)$, corresponding to a statistically significant response ($p = 0.1$) compared to the baseline rate $r = 2$, are included in the analysis.

The discriminability between stimulus pairs is computed to assess whether the sensory area can continue to process stimuli accurately. Discriminability is quantified using a form of the Fisher information metric[74], specifically the squared d-prime, summed across neurons:

$$F = \sum_{i=1}^{N_{\text{neu}}} \frac{(r_i^A - r_i^B)^2}{r_i^A + r_i^B}, \tag{28}$$

where A and B denote two different stimulus conditions. With this formulation, the Fisher information can be interpreted as the discriminability of an optimal linear decoder[75].

In the motion-vision mismatch test (Fig. 4), we examine the emergence of a bimodal distribution in the correlation diagram following learning. To generate this distribution, we randomize the strengths of the stimulus and prediction inputs and run 200 trials

continuously in the model. For each neuron, we calculate the correlation between its response and the input strengths u_S and u_P. We then compute an angular difference, θ, in polar coordinates, where $\theta = 0$ indicates equal correlation with stimulus and prediction. This angular metric is further weighted by the radial distance in the polar plane to emphasize neurons with larger response variance. A bimodal distribution of θ after learning reflects the emergence of both pPE and nPE neuronal populations. Because neuronal responses in the model are solely determined by the stimulus and prediction input strengths, the response variance is almost entirely explained by these two variables. Consequently, the data points lie approximately on a ring manifold in the correlation space.

We further apply the same analysis to data collected from C57BL/6J mice of either sex ($n = 6$ per group) under both coupled training (CT) and non-coupled training (UT) conditions, originally published in ref. 9. The total number of cells under examination is $N_{CT} = 939$ for the CT group, and $N_{NT} = 690$ for the NT group. We demonstrate that the bimodal distribution emerges only in the coupled training condition by fitting the angular data to Gaussian Mixture Models (GMMs) with one or two components[76].

To perform this analysis on weighted data, we first construct a pseudo-dataset. For each cell in polar coordinates with angle θ_i and radius r_i, we replicate θ_i by $100 \times r_i$ times to reflect the weighting. Only data with $\theta_i \in [-\pi/2, \pi/2]$ are included, in order to avoid spurious peaks near $\theta = -\pi$. All replicated values are collected to form the pseudo-dataset. The resulting pseudo-dataset is then fitted to Gaussian Mixture Models using the MATLAB function *fitgmdist*, with either one or two components and default parameter settings.

To evaluate whether one-component or two-component fitting is superior, we use the Bayesian Information Criterion (BIC)[77]. Typically, a difference of $\Delta BIC > 10$ is considered strong evidence in favor of the model with the lower BIC. In our analysis, fitting two-component GMMs to both pseudo CT and NT datasets yields $\Delta BIC > 10$, favoring the two-component model in both cases (Supplementary Fig. 9). However, when we overlay the fitted GMMs on the pseudo-datasets, only the CT datasets show a true bimodal distribution. In contrast, the uncoupled condition shows a single dominant peak, suggesting that the means of the two Gaussian components are nearly identical (Supplementary Fig. 9a, blue dashed line).

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work, the authors used ChatGPT-4o in order to assist with proofreading and improving the clarity of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability

We have re-analyzed previously published data from[9], which is available in the original paper. All other simulated data presented in this study were generated using code developed by the authors. The results can be reproduced using the code provided below.

Code availability

We analyze the model-generated data using Python 3.9 and the experimental data using MATLAB R2024b. All analysis code is available on GitHub at (https://github.com/johnhongyumeng/DuetPredictiveCoding_Plasticity). and on Zenodo at[78].

References

1. Crapse, T. B. & Sommer, M. A. Corollary discharge across the animal kingdom. *Nat. Rev. Neurosci.* **9**, 587–600 (2008).
2. Sommer, M. A. & Wurtz, R. H. A pathway in primate brain for internal monitoring of movements. *science* **296**, 1480–1482 (2002).
3. Lappe, M., Bremmer, F. & van den Berg, A. V. Perception of self-motion from visual flow. *Trends Cogn. Sci.* **3**, 329–336 (1999).
4. Bolt, N. K. & Loehr, J. D. The auditory p2 differentiates self-from partner-produced sounds during joint action: contributions of self-specific attenuation and temporal orienting of attention. *Neuropsychologia* **182**, 108526 (2023).
5. Frith, C. The neural basis of hallucinations and delusions. *Comptes Rendus. Biol.* **328**, 169–175 (2005).
6. Corlett, P. R., Taylor, J. R., Wang, X.-J., Fletcher, P. C. & Krystal, J. H. Toward a neurobiology of delusions. *Prog. Neurobiol.* **92**, 345–369 (2010).
7. Keller, G. B. & Mrsic-Flogel, T. D. Predictive processing: a canonical cortical computation. *Neuron* **100**, 424–435 (2018).
8. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
9. Attinger, A., Wang, B. & Keller, G. B. Visuomotor coupling shapes the functional development of mouse visual cortex. *Cell* **169**, 1291–1302 (2017).
10. Audette, N. J., Zhou, W., La Chioma, A. & Schneider, D. M. Precise movement-based predictions in the mouse auditory cortex. *Curr. Biol.* **32**, 4925–4940 (2022).
11. De Kock, C., Bruno, R. M., Spors, H. & Sakmann, B. Layer-and cell-type-specific suprathreshold stimulus representation in rat primary somatosensory cortex. *J. Physiol.* **581**, 139–154 (2007).
12. Niell, C. M. & Stryker, M. P. Highly selective receptive fields in mouse visual cortex. *J. Neurosci.* **28**, 7520–7536 (2008).
13. Sakata, S. & Harris, K. D. Laminar structure of spontaneous and sensory-evoked population activity in auditory cortex. *Neuron* **64**, 404–418 (2009).
14. Karnani, M. M., Agetsuma, M. & Yuste, R. A blanket of inhibition: functional inferences from dense inhibitory connectivity. *Curr. Opin. Neurobiol.* **26**, 96–102 (2014).
15. Vogels, T. P., Sprekeler, H., Zenke, F., Clopath, C. & Gerstner, W. Inhibitory plasticity balances excitation and inhibition in sensory pathways and memory networks. *Science* **334**, 1569–1573 (2011).
16. Kullmann, D. M., Moreau, A. W., Bakiri, Y. & Nicholson, E. Plasticity of inhibition. *Neuron* **75**, 951–962 (2012).
17. Chiu, C. Q., Barberis, A. & Higley, M. J. Preserving the balance: diverse forms of long-term GABAergic synaptic plasticity. *Nat. Rev. Neurosci.* **20**, 272–281 (2019).
18. Brzosko, Z., Mierau, S. B. & Paulsen, O. Neuromodulation of spike-timing-dependent plasticity: past, present, and future. *Neuron* **103**, 563–581 (2019).
19. Jordan, R. & Keller, G. B. The locus coeruleus broadcasts prediction errors across the cortex to promote sensorimotor plasticity. *Elife* **12**, RP85111 (2023).
20. Jordan, R. The locus coeruleus as a global model failure system. *Trends Neurosci.* **47**, 92–105 (2024).
21. Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D. & Brea, J. Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules. *Front. neural circuits* **12**, 53 (2018).
22. Seol, G. H. et al. Neuromodulators control the polarity of spike-timing-dependent synaptic plasticity. *Neuron* **55**, 919–929 (2007).
23. Najafi, F. et al. Excitatory and inhibitory subnetworks are equally selective during decision-making and emerge simultaneously during learning. *Neuron* **105**, 165–179 (2020).
24. Znamenskiy, P. et al. Functional specificity of recurrent inhibition in visual cortex. *Neuron* **112**, 991–1000 (2024).
25. McFarlan, A. R. et al. The plasticitome of cortical interneurons. *Nat. Rev. Neurosci.* **24**, 80–97 (2023).
26. Grier, B. D., Parkins, S., Omar, J. & Lee, H.-K. Selective plasticity of fast and slow excitatory synapses on somatostatin interneurons in adult visual cortex. *Nat. Commun.* **14**, 7165 (2023).

27. Gouwens, N. W. et al. Integrated morphoelectric and transcriptomic classification of cortical GABAergic cells. *Cell* **183**, 935–953 (2020).
28. Audette, N. J. & Schneider, D. M. Stimulus-specific prediction error neurons in mouse auditory cortex. *J. Neurosci.* **43**, 7119–7129 (2023).
29. Krystal, J. H. et al. NMDA receptor antagonist effects, cortical glutamatergic function, and schizophrenia: toward a paradigm shift in medication development. *Psychopharmacology* **169**, 215–233 (2003).
30. Farley, B. J., Quirk, M. C., Doherty, J. J. & Christian, E. P. Stimulus-specific adaptation in auditory cortex is an nmda-independent process distinct from the sensory novelty encoded by the mismatch negativity. *J. Neurosci.* **30**, 16475–16484 (2010).
31. Taaseh, N., Yaron, A. & Nelken, I. Stimulus-specific adaptation and deviance detection in the rat auditory cortex. *PloS one* **6**, e23369 (2011).
32. Frémaux, N. & Gerstner, W. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Front. neural circuits* **9**, 85 (2016).
33. Kuśmierz, Ł, Isomura, T. & Toyoizumi, T. Learning with three factors: modulating hebbian plasticity with errors. *Curr. Opin. Neurobiol.* **46**, 170–177 (2017).
34. Hebb, D. O. et al. *The Organization Of Behavior: A Neuropsychological Theory*. (Psychology Press, 2005).
35. Pawlak, V., Wickens, J. R., Kirkwood, A. & Kerr, J. N. Timing is not everything: neuromodulation opens the STDP gate. *Front. Synaptic Neurosci.* **2**, 146 (2010).
36. Bissière, S., Humeau, Y. & Lüthi, A. Dopamine gates ltp induction in lateral amygdala by suppressing feedforward inhibition. *Nat. Neurosci.* **6**, 587–592 (2003).
37. Zenke, F. & Gerstner, W. Hebbian plasticity requires compensatory processes on multiple timescales. *Philos. Trans. R. Soc. B: Biol. Sci.* **372**, 20160259 (2017).
38. Mermillod, M., Bugaiska, A. & Bonin, P. The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. **4**, 504 (2013).
39. Halvagal, M. S. & Zenke, F. The combination of Hebbian and predictive plasticity learns invariant object representations in deep sensory networks. *Nat. Neurosci.* **26**, 1906–1915 (2023).
40. Bellec, G. et al. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nat. Commun.* **11**, 3625 (2020).
41. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
42. Yogesh, B. & Keller, G. B. Activity in serotonergic axons in visuomotor areas of cortex is modulated by the recent history of visuomotor coupling. *Peer Community J.* **5**, (2025).
43. Friston, K. & Kiebel, S. Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. B: Biol. Sci.* **364**, 1211–1221 (2009).
44. Boerlin, M., Machens, C. K. & Denève, S. Predictive coding of dynamical variables in balanced spiking networks. *PLoS Comput. Biol.* **9**, e1003258 (2013).
45. Spratling, M. W. A review of predictive coding algorithms. *Brain Cognition* **112**, 92–97 (2017).
46. Whittington, J. C. & Bogacz, R. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Comput.* **29**, 1229–1262 (2017).
47. Song, Y., Lukasiewicz, T., Xu, Z. & Bogacz, R. Can the brain do backpropagation?—exact implementation of backpropagation in predictive coding networks. *Adv. neural Inf. Process. Syst.* **33**, 22566–22579 (2020).
48. Millidge, B., Seth, A. & Buckley, C. L. Predictive coding: a theoretical and experimental review. *arXiv preprint arXiv:2107.12979* (2021).
49. Mikulasch, F. A., Rudelt, L., Wibral, M. & Priesemann, V. Where is the error? hierarchical predictive coding through dendritic error computation. *Trends Neurosci.* **46**, 45–59 (2023).
50. Hertäg, L. & Clopath, C. Prediction-error neurons in circuits with multiple neuron types: Formation, refinement, and functional implications. *Proc. Natl. Acad. Sci. USA.* **119**, e2115699119 (2022).
51. Barry, M. L. & Gerstner, W. Fast adaptation to rule switching using neuronal surprise. *PLoS Comput. Biol.* **20**, e1011839 (2024).
52. Hertäg, L. & Sprekeler, H. Learning prediction error neurons in a canonical interneuron circuit. *Elife* **9**, e57541 (2020).
53. Sweller, J. Cognitive load during problem solving: effects on learning. *Cogn. Sci.* **12**, 257–285 (1988).
54. Sweller, J. et al. Cognitive load theory. In *Psychology of learning and motivation*. 55, 37–76 (Elsevier, 2011).
55. Bassett, D. S., Yang, M., Wymbs, N. F. & Grafton, S. T. Learning-induced autonomy of sensorimotor systems. *Nat. Neurosci.* **18**, 744–751 (2015).
56. O'Toole, S. M., Oyibo, H. K. & Keller, G. B. Molecularly targetable cell types in mouse visual cortex have distinguishable prediction error responses. *Neuron* **111**, 2918–2928 (2023).
57. Condylis, C. et al. Dense functional and molecular readout of a circuit hub in sensory cortex. *Science* **375**, eabl5981 (2022).
58. Jordan, R. & Keller, G. B. Opposing influence of top-down and bottom-up input on excitatory layer 2/3 neurons in mouse primary visual cortex. *Neuron* **108**, 1194–1206 (2020).
59. Widmer, F. C., O'Toole, S. M. & Keller, G. B. Nmda receptors in visual cortex are necessary for normal visuomotor integration and skill learning. *Elife* **11**, e71476 (2022).
60. Schneider, D. M., Nelson, A. & Mooney, R. A synaptic and circuit basis for corollary discharge in the auditory cortex. *Nature* **513**, 189–194 (2014).
61. Schneider, D. M., Sundararajan, J. & Mooney, R. A cortical filter that learns to suppress the acoustic consequences of movement. *Nature* **561**, 391–395 (2018).
62. Pfeffer, C. K., Xue, M., He, M., Huang, Z. J. & Scanziani, M. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nat. Neurosci.* **16**, 1068–1076 (2013).
63. Dellal, S. et al. Inhibitory and disinhibitory vip in-mediated circuits in neocortex. Preprint at *bioRxiv* https://doi.org/10.1101/2025.02.26.640383 (2025).
64. Garner, A. R. & Keller, G. B. A cortical circuit for audio-visual predictions. *Nat. Neurosci.* **25**, 98–105 (2022).
65. Solyga, M. & Keller, G. B. Multimodal mismatch responses in mouse auditory cortex. *eLife* **13**, RP95398 (2025).
66. Eliades, S. J. & Wang, X. Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature* **453**, 1102–1106 (2008).
67. Sommer, M. A. & Wurtz, R. H. Brain circuits for the internal monitoring of movements. *Annu. Rev. Neurosci.* **31**, 317–338 (2008).
68. Meng, J. H., Ross, J. M., Hamm, J. P. & Wang, X.-J. Duet model unifies diverse neuroscience experimental findings on predictive coding. *bioRxiv* https://doi.org/10.1101/2025.07.12.664417 (2025).
69. Squires, K. C., Wickens, C., Squires, N. K. & Donchin, E. The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science* **193**, 1142–1146 (1976).
70. Larkum, M. E., Senn, W. & Lüscher, H.-R. Top-down dendritic input increases the gain of layer 5 pyramidal neurons. *Cereb. cortex* **14**, 1059–1070 (2004).
71. Abbott, L. F. & Chance, F. S. Drivers and modulators from push-pull and balanced synaptic input. *Prog. Brain Res.* **149**, 147–155 (2005).
72. Oswald, A.-M. M. & Reyes, A. D. Maturation of intrinsic and synaptic properties of layer 2/3 pyramidal neurons in mouse auditory cortex. *J. Neurophysiol.* **99**, 2998–3008 (2008).

73. Wong, K.-F. & Wang, X.-J. A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.* **26**, 1314–1328 (2006).
74. Cover, T. M. et al. *Elements of Information Theory* (John Wiley & Sons, 1999).
75. Meng, J. H. & Riecke, H. Structural spine plasticity: Learning and forgetting of odor-specific subnetworks in the olfactory bulb. *PLoS Comput. Biol.* **18**, e1010338 (2022).
76. Bishop, C. M. & Nasrabadi, N. M. *Pattern Recognition And Machine Learning*. 4 (Springer, 2006).
77. Schwarz, G. et al. Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464 (1978).
78. Meng, J. H. & Wang, X.-J. Global error signal guides local optimization in mismatch calculation (this paper). *DuetPredictiveCoding_Plasticity: dPC_plastictiy_v1.0* (2026).

Acknowledgements

We thank Aldo Battista, Yue Liu, Tianshu Li, and other members of Xiao-Jing Wang lab for the discussion. We thank Cristina Savin and David Schneider for their suggestions and feedback on an early version of the manuscript. We thank Georg Keller for technical assistance regarding access to the publicly available dataset. This work is supported by ONR grant N00014-23-1-2040 and NIH grant R01MH062349 (to XJW)

Author contributions

J.H.M. and X.J.W. conceptualized the project. J.H.M. performed simulations and data analyses. J.H.M. and X.J.W. wrote the manuscript. X.J.W. supervised the study.

Competing interests

The author declares no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-70354-x.

Correspondence and requests for materials should be addressed to Xiao-Jing Wang.

Peer review information *Nature Communications* thanks Paul Miller, Joel Zylberberg, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.