Desegregation of neuronal predictive processing

Bin Wang[1], Nicholas J Audette[2], David M Schneider[2], Johnatan Aljadeff[3,*]

[1] Department of Physics, University of California San Diego, La Jolla, CA, 92093, USA

[2] Center for Neural Science, New York University, New York, NY 10003, USA

[3] Department of Neurobiology, University of California San Diego, La Jolla, CA, 92093, USA

Display items: 6 Figures, 8 Supplementary Figures

* To whom correspondence should be addressed; E-mail: aljadeff@ucsd.edu (J.A.).

Abstract

Neural circuits construct internal 'world-models' to guide behavior. The predictive processing framework posits that neural activity signaling sensory predictions and concurrently computing prediction-errors is a signature of those internal models. Here, to understand how the brain generates predictions for complex sensorimotor signals, we investigate the emergence of high-dimensional, multi-modal predictive representations in recurrent networks. We find that robust predictive processing arises in a network with loose excitatory/inhibitory balance. Contrary to previous proposals of functionally specialized cell-types, the network exhibits desegregation of stimulus and prediction-error representations. We confirmed these model predictions by experimentally probing predictive-coding circuits using a rich stimulus-set to violate learned expectations. When constrained by data, our model further reveals and makes concrete testable experimental predictions for the distinct functional roles of excitatory and inhibitory neurons, and of neurons in different layers along a laminar hierarchy, in computing multi-modal predictions. These results together imply that in natural conditions, neural representations of internal models are highly distributed, yet structured to allow flexible readout of behaviorally-relevant information. The generality of our model advances the understanding of computation of internal models across species, by incorporating different types of predictive computations into a unified framework.

Introduction

Predictive coding, the process of computing the expected values of sensory, motor, and other task-related quantities, is thought to be a fundamental operation of the brain [1, 2]. Violation of internally-generated expectations, known as *prediction-errors*, is an important neural signal that can be used to guide learning and synaptic plasticity [3, 4]. Signatures of predictive coding, including neural correlates of prediction-errors, were identified in multiple brain circuits, and across animal species [2,5–7]. Two well-studied examples are motor-auditory [8–13] and visual-auditory predictions [14–16] in the mouse cortex. Previous work has proposed that a *canonical cortical microcircuit* underlies the computation of predictions and prediction-errors [2,8,9,17–19]. While some predictions of this proposed microcircuit were confirmed in restricted scenarios, the hypothesis that the circuit-motif within the mouse cortex is a general mechanism for predictive processing faces a number of challenges.

First, typical experimental paradigms study predictive coding in animals trained to make a single association [12, 16, 20], while natural sensorimotor associations are typically high-dimensional (e.g., speech production [21]), as well as context-dependent [22, 23]. Little is known about how specific neural architectures in the brain learn to implement such high-dimensional computations. Second, multiple brain circuits outside of the mammalian cortex exhibit predictive coding, including subcortical circuits mediating placebo analgesia (*prediction-based* suppression of pain [24]); and motor-visual circuits in cephalopods that predict the animal's appearance to an external observer, and use it to generate high-dimensional camouflage patterns [25]. It is not known whether these neural circuits use similar or altogether different strategies for predictive processing as the mammalian cortex. Third, predictive neural representations emerge on timescales ranging from $\sim$1 minute [26,27], $\sim$1 hour [28,29], to days [16,30]. This suggests that predictive processing is supported by plasticity mechanisms operating on a range of timescales (including short-term plasticity [31]), and that circuit reorganization may not always be required for implementing predictive computations.

The evidence that computing predictions is an integral part of sensory processing has garnered significant attention from the theoretical neuroscience community. Several studies have proposed recurrent network models that may perform these computations [32–39]. These studies typically focus on predicting a small number of stimuli within a single sensory modality.

Moreover, in most cases, these models have been compared with coarse-grained neuroimaging data [7, 40]. Therefore, we lack cellular-level and circuit-level understanding of neural mechanisms underlying multi-modal predictive computations, which limits our ability to test hypotheses related to the circuit computation of predictive coding based on modern large-scale neural recordings.

Another major current gap from both experimental and modeling perspectives is predictive processing in high-dimensions: (*i*) What are the neural representations of predictable and unpredictable sensory variables in natural conditions with rich stimulus ensembles and complex inter-dependencies between stimuli [7,41,42]? (*ii*) What are the circuit mechanisms underlying the computation of those representations, and how are they learned? Specifically, it remains unknown whether circuits that implement predictive coding of high-dimensional stimulus ensembles are functionally segregated [2, 17, 18], and if so, whether this segregation emerges during learning or depends on molecularly distinct cell-types.

We address these questions by developing a mathematical framework to examine the predictive representations in recurrent networks processing naturalistic inputs during and after learning, and by relating this model to cellular- and population-level neural recordings. From a mechanistic perspective, we provide novel predictions into the expected degree of excitation/inhibition balance in the high-dimensional case, and shed light on the role that E/I balance plays in canceling interference between multiple learned stimuli. Moreover, since E/I balance is enforced by mechanisms operating on heterogeneous timescales [43], our model may allow incorporating seemingly unrelated phenomena into a unified framework, e.g., predictive responses that change as a result of short- or long-term plasticity. From a functional perspective, the model suggests that predictive processing of high-dimensional stimuli is robust when the representations of stimuli and of prediction-errors are *desegregated* at the cellular-level, and distributed across excitatory and inhibitory neurons. Finally, we applied our theory to examine the distinct roles played by excitatory and inhibitory neurons in generating internal predictions, and to assess the layer-specific predictive representations.

Our modeling and analysis overcomes key limitations of previous studies of predictive processing, and generates novel predictions that we confirmed here based on experimental data. Therefore, we believe that our work reveals principles of predictive processing across species and brain-regions and provides a quantitative framework for design and analysis of future experiments to decipher neural circuits underlying those computations.

Results

Recurrent networks that learn to generate high-dimensional predictions

We studied the neural representations formed in recurrent neural networks that perform predictive processing of multi-modal sensory and motor inputs. We focused on a typical associative training scenario where animals are presented with pairs of sensory stimuli simultaneously [9, 11, 12] or after a short delay [16]. The stimuli comprising each pair are typically of different sensory modalities (e.g., auditory-visual [16]), or involve a sensory-motor association (e.g., locomotion-auditory [12]). In this scenario, predictive computations are thought to be learned over time through synaptic-weight updates [9, 11, 12, 16, 20, 44]. Our network model consists of N recurrently connected neurons whose firing-rates depend nonlinearly on the input current driving their responses (Fig. 1a). The presentation of stimuli to the network is determined by the variables x and y. The strength of the input to each neuron corresponds to the components of the stimulus-specific feedforward synaptic weight vectors $\boldsymbol{w}$ and $\boldsymbol{v}$. There are P stimulus-pairs, and when P is of the same order as the number of neurons N, the network is said to perform *high-dimensional* predictive processing.

Before training, the feedforward weight vectors corresponding to each stimulus-pair are random and uncorrelated within the pair (i.e., $\boldsymbol{w} \cdot \boldsymbol{v} = 0$). During training, those weights become correlated ($\boldsymbol{w} \cdot \boldsymbol{v} = \mu$, with $\mu > 0$), consistent with measurements of learning-induced functional reorganization of excitatory synaptic connections [45–47]. Weights of recurrent connections are chosen to minimize errors between internally generated predictions and the actual stimuli, while maximizing the overall encoding efficiency (Methods). Under these assumptions, we obtained key statistics of neural activity in the network for different stimulus inputs, at different stages of learning (Methods). The resulting neural activity allows flexibly reading-out the stimulus identity, predicting the 'missing' stimulus (i.e., predicting y based on x), and evaluating the prediction-error (Fig. 1a). We applied the modeling framework developed here (SI §1-2) to investigate the structure of multi-modal predictive neural representations and the circuit mechanisms supporting it.

We first examined neural responses during learning in the *match* ($x = y$), and *mismatch* ($x \neq y$) conditions. We set x and y to be binary variables corresponding to the presence ($x,\ y = 1$) or absence ($x,\ y = 0$) of visual-auditory, visual-motor, or auditory-motor pairings [12, 13, 16, 20]. Our mathematical formalism extends to scenarios where more than two stimuli are predictive of each other, and where the inputs to the network vary continuously (e.g., running- or visual-flow-speed [20, 44]; Methods). Before associative training ($\mu = 0$), most of the neurons in the network have comparable match ($\boldsymbol{r}_{xy}$) and mismatch ($\boldsymbol{r}_x,\ \boldsymbol{r}_y$) responses (Fig. 1b). After training ($\mu = 0.9$), match responses are suppressed while mismatch responses are amplified (Fig. 1b). Correspondingly, the ratio of average mismatch and match firing-rates increases (Fig. 1c), consistent with associative learning experiments [12, 16, 20]. Thus, the presence of stimulus y suppresses the response evoked by stimulus x, and generates a *prediction* (or *expectation*) of x. Amplified mismatch responses are interpreted as *prediction-errors* [2, 7].

During learning, the mismatch responses ($\boldsymbol{r}_x,\ \boldsymbol{r}_y$) become *anti*-correlated (Fig. 1d,e), i.e., the presence of stimulus y more effectively suppresses responses to x alone. This anti-correlation does not appear between $\boldsymbol{r}_x$ and $\boldsymbol{r}_y$ of *another stimulus-pair* (Fig. S1a), suggesting that the predictive signal triggered by stimulus y, is *specific* to its paired stimulus x, consistent with Refs. [12, 48]. The *specific* suppression of responses to predictable stimuli is accompanied by a weaker, *global* gain that depends on the overall magnitude of sensory input (SI §2). Furthermore, match and mismatch neural responses *decorrelate* during learning (Fig. 1d,e), consistent with Ref. [16], suggesting that neural responses can be used to distinguish between presentation of stimulus x in the match or mismatch condition. Notably, Owing to the neural response nonlinearity, the match response is not a sum of the two mismatch responses ($\boldsymbol{r}_{xy} \neq \boldsymbol{r}_x + \boldsymbol{r}_y$, Fig. 1c).

Next we examined neural responses when the network is trained with *two* stimulus-pairs ($P = 2$, Fig. 1f), making a step towards the high-dimensional scenario. [2, 17, 18, 49, 50] proposed that neurons involved in predictive processing are functionally segregated, i.e., neurons that signal prediction-error for one stimulus association tend to signal prediction-error for other associations, and similarly for 'representation' neurons that encode the stimulus itself. This proposal would predict a high degree of correlation between neural responses to two stimulus-pairs (Fig. 1f, right). However, we found no such correlation in our model (Fig. 1f, left). This implies, for example, that a neuron that signals prediction-error for stimulus-pair 1, may have a selective response to stimulus x 'itself' for pair 2, and raises the question of what circuit mechanisms may support this cellular-level desegregation of response types.

Learning and stimulus dimensionality determine the properties of effective predictive processing circuits

We then investigated circuit mechanisms underlying multi-modal high-dimensional predictive processing. We decomposed the input to each neuron into feedforward and recurrent components, which respectively correspond to the actual stimulus signal and to internally generated predictions (Fig. 2a), similarly to analyses of previous experiments [2, 12, 17, 20]. To quantify the relative contribution of each component, we follow the excitatory/inhibitory (E/I) balance literature [33, 51], and define the *balance* level B as the ratio between the total feedforward input and the net input to each neuron, in each condition (Fig. 2a).

During associative learning, internally generated predictions become more accurate, facilitating more robust cancellation of the feedforward stimulus input by recurrent feedback conveying prediction signals. Thus, the overall balance level increases in the match condition but decreases in the mismatch condition (Fig. 2b, left). Notice that the balance level distributions (over neurons and stimuli) are initially similar in the match and mismatch conditions, but become significantly different in late stages of learning (Fig. 2b, right). Indeed, after learning, the mode of the balance level distribution is at $B \approx 0$ in the mismatch condition, which explains the strong prediction-error responses.

To understand the role of balance in predictive processing, we examined its effect on the nonlinear transformation the network performs, from input stimuli to neural activity (Fig. 2c). In our model, the geometry of neural responses facilitates robust readout of prediction-errors. Specifically, while prediction-errors cannot be read-out by a linear decoder from the stimulus input, such a readout is feasible once the input is transformed into the network's high-dimensional response (Fig. 2d). Moreover, while the prediction-error itself is stimulus-specific, the decoder that performs this computation is stimulus-independent after learning—it is simply the average firing-rate (Fig. 2d). In other words, the learned structure of neural responses enables applying the same decoder to all stimulus-pairs without 're-learning'.

Given the essential role of the nonlinear transformation for predictive processing, we next focused on the effect of the overall nonlinear gain parameter b (Methods, [34]). We found that increasing b leads to increases of the average match and mismatch firing-rate responses, together with a wider margin between them (Fig. 2e, top). Therefore, large b facilitates decoding prediction-errors, at the cost of increased overall neural activity. Motivated by this observation, and since b is an intrinsic network quantity that can potentially be adjusted dynamically, we sought to find an optimal value (denoted $b^\star$). Specifically, we constrained the average network response in the mismatch condition to be larger than a certain threshold, while requiring a minimal but nonzero average response in the match condition (Fig. 2e), consistent with reports of weak neural responses to predictable stimuli [12, 20]. The resulting $b^\star$ corresponds to an optimal balance level $B^\star$ supporting efficient encoding *and* robust decoding (Fig. 2e, bottom).

We carried out this optimization procedure for networks trained to perform predictive processing of stimulus ensembles with increasing dimensionality (i.e., increasing $\alpha = P/N$), with the same firing-rate constraints chosen such that the value of $B^\star$ at $\alpha = 0$ matches experimental data. We additionally assumed that an ‘over-trained’ animal learns a single stimulus-pair (i.e., $\alpha = 1/N \approx 0$). Surprisingly, we found that the optimal balance level *decreases* with α (Fig. 2f), independently of the stimulus statistics (Fig. S1b,c). This is because as the number of stimulus-pairs learned by the network increases, so does the interference between internally generated predictions corresponding to different stimulus-pairs (Methods). We therefore expect networks performing predictive processing in natural conditions (large α) to exhibit ‘loose’ balance, which minimizes the overall effect of interference arising from learning to generate a large number of internal predictions.

We used neural activity recorded from animals trained on visual-motor (V-M) [20] and auditory-motor (A-M) associations [12] to constrain our network model. Specifically, we estimated the balance levels in mouse sensory cortex by assuming that after training the neural network *in vivo* reaches the optimal balance level. In the V-M experiment [20], mice were trained to associate their running speed with the speed of visual-flow in virtual reality (Fig. 3a). The voltage of primary visual cortex neurons was intracellurlary recorded in the match and mismatch conditions. Fitting the average voltage change in the two conditions to our model gives the estimated balance level $B^\star_{\text{V-M}} = 162 \pm 61$. A consistent result was obtained in the A-M experiment [12], where mice were trained to press a lever and received closed-loop auditory feedback (Fig. 3b,c). Here the recording was extracellular, so fitting $B^\star$ relied on a slightly modified procedure (Methods).

It is notable that balance level estimates were consistent across animals (Fig. S2); and laboratories (Fig. 3), despite the fact that the experiments studied different brain regions and sensory modalities, using different methods. While these factors may affect the balance level to some degree, our model predicts that the balance level can decrease by up to one order of magnitude when the stimulus dimension increases (Fig. 2e, Fig. S1b,c). This prediction could be confirmed if future experiments reveal a more loose balance in animals habituated to rich sensory environments.

Stimulus and prediction-error representations are desegregated in the model

We next investigated how different functional responses are organized within the network. Previous work postulated that two distinct neural populations exist in predictive processing circuits: (*i*) *internal representation* (*R*) neurons that 'faithfully' represent external sensory stimuli and encode internal predictions, and (*ii*) *prediction-error* (*PE*) neurons, which signal the difference between the actual stimulus inputs and internal predictions. Given that neurons selective to these signals also exist in our network model, we wondered whether they form functionally segregated populations. We adopted classification criteria used in experimental work (Methods, [2, 48]): *R* neurons are those which respond strongly and similarly in match and mismatch conditions, while *PE* neurons are those which respond strongly in the mismatch condition but weakly in the match condition (Fig. 4a).

Based on these criteria, we first computed the fractions of *R* and *PE* neurons when the network learns a single stimulus association ($P = 1$, Fig. 4b). As training progresses, the fraction of *PE* neurons increases significantly, consistent with experiments [16, 52], and with the notion that the network learns to ‘recognize’ the stimulus pairing. This result is independent of the classification criterion (Fig. S3). The fraction of *R* neurons remains unchanged (Fig. 4b), though we note that the trend does depend on the criterion (Fig. S3).

We next asked how neurons responded to more complex stimulus ensembles, specifically for two learned pairs of stimuli. The hypothesis that predictive processing is segregated [2, 18] asserts that if a neuron is a *PE* neuron for stimulus-pair 1, and if it is active during presentation of stimuli from pair 2, it will likely be categorized as a *PE* neuron with respect to those stimuli too. To test this hypothesis, we computed the joint distribution of neural responses in the four relevant conditions (mismatch/match, stimulus-pair $1/2$) and categorized each neuron as *R* or *PE*, separately for each stimulus-pair (Methods). We started with the low-dimensional scenario, where the two stimulus-pairs in question are the only stimuli learned by the network ($P = 2$, $\alpha = P/N \approx 0$). Surprisingly, under the data-constrained parameters, although many neurons belong to the same functional type with respect to the two stimulus-pairs, approximately 25% of neurons are in fact *‘mixed’*: they are classified as having different functional types (Fig. 4c, left).

Furthermore, increasing the dimension of the stimulus the network learns, leads to a twofold increase in the fraction of mixed neurons (Fig. 4c,d). Intuitively, loose balance between high-dimensional feedforward and recurrent inputs leads to a broad balance level distribution across the network (Fig. S4a). That broad distribution, in turn, affords each neuron flexibility to encode different features for different stimulus-pairs. The fraction of mixed neurons shown in Fig. 4d corresponds to two *specific* stimulus-pairs. When we considered instead the entire learned stimulus-set, *most* of the neurons are mixed with respect to at least two pairs (Fig. S4b). Thus, contrary to the previous hypothesis [2], neurons with mixed representations of stimuli and predictions are common in the network model, especially in high-dimensional scenarios.

Experimental evidence for desegregated predictive representations

We then turned to testing this key prediction of our network model, by looking for signatures of mixed representations of predictions and stimuli in experimental data. In our recent work, we recorded primary auditory cortex responses in mice that were trained to associate a simple behavior, pressing a lever, with a simple outcome, a predictable tone [13]. Following extensive training, we made extracellular recordings from auditory cortex while animals were presented with *probe* auditory stimuli that differed from the *expected* stimulus along a variety of different dimensions, and while animals either pressed the lever or heard the tone passively (Fig. 4e).

Here we analyzed this data as follows. For each neuron, we computed the difference (Δ) between the mismatch (passive: sound only) and match (active: lever press + sound) neural responses (Fig. 4e, bottom), similar to our analysis of the neural activity in the model (Fig. 4c). Note that for each of the four probe sounds, 'match' corresponds to a lever press paired with the probe sound, while 'mismatch' corresponds to responses following the probe sound without lever press. We expected Δ values of mixed neurons to lie in the upper left or lower right corners of the plot (similarly to Fig. 4c, blue rectangles). This would correspond to neurons with match and mismatch responses that are similar for the expected sound but differ for the probe sound, or vice versa.

We quantified the degree of mixing, or desegregation of the predictive representation, by computing the Pearson correlation coefficient of the Δ values corresponding to the expected sound and each probe sound separately (Fig. 4e). We defined this coefficient as the *segregation index*, which is close to 1 if the Δ's are strongly correlated between the two stimulus-pairs (expected, probe). A segregation index close to 0 means that the representations of stimuli and predictions are 'maximally mixed'. We additionally computed representation similarity between the expected and probe sounds, as the correlation between neural responses to those stimuli. Crucially, representation similarity was based on neural responses in a separate experimental window during which sounds were presented passively, not following a lever press [13]. If neurons are segregated into two functional classes, the segregation index should be close to 1 irrespective of the representation similarity. By contrast, we found that the segregation index depends strongly on the representation similarity (Fig. 4f). Specifically, when the expected and probe sounds are similar (Fig. 4e,f, green shades), the segregation index is close to 1, though a random subsampling analysis indicates a statistically significant effect of the representation similarity on the segregation index. When the probe differs from the expected sound more substantially (Fig. 4e,f, orange), the segregation index drops to ~ 0.5. This relation between representation similarity and degree of segregation is consistent with the prediction of our model, with an appropriate level of coding sparsity (Fig. 4f). The significant dependence of the segregation index on the representation similarity, and the fact that the segregation index is substantially smaller than 1, suggest that predictive processing is mixed in the mouse auditory cortex. A similar relationship was found when we used the ‘complementary’ mismatch response to compute the Δ’s, i.e., based on the neural response to a lever press with no sound, rather than a sound with no lever press (Fig. S5).

We note that the analysis presented here is an indirect test of the model prediction that predictive representations are mixed. Indeed, the desegregation in the model involves two learned stimulus-pairs (Fig. 4c), while in the experiment the animal was only trained on the expected sound. Nevertheless, the decreased segregation index we found for probe sounds markedly different from the expected sound provides strong evidence against the notion that predictive processing circuit is functionally segregated into separate neural populations. Our model provides a framework to generate hypothesis that could be tested more directly in future experiments.

Predictive processing in excitatory–inhibitory networks

Thus far we have focused on relating neural responses in the model to measurements of excitatory neurons' activity [12, 13, 16]. Each neuron's projections in our network could be both excitatory (E) and inhibitory (I), so it does not obey Dale's law. Given the growing literature on the role of inhibitory neurons in computing predictions [35, 36], we sought to link our model to experiments more tightly by extending it to a network with separate E and I neurons. We did so by requiring that the activity of E neurons in the E/I network matched exactly that of neurons in the original model. This guarantees that the E neurons possess the predictive coding properties we studied so far, and opens the door to study the functional role of I neurons. The connectivity in the E/I network has four components, corresponding to synapses to and from E and I neurons (Fig. 5). We used non-negative matrix factorization to 'solve' for those components (Methods, [53, 54]). The balance level B defined previously based on feedforward and recurrent inputs (Fig. 2), is equal to the stimulus-specific component of the E/I balance in the E/I networks (SI §4).

The aforementioned mathematical procedure did not yield a *unique* connectivity structure. Rather, we found a one-parameter family of connectivity structures that all meet those constraints. This parameter, denoted λ_{EI}, interpolates between two extremes of structured E/I connectivity (Fig. 5b). In one extreme ($\lambda_{EI} = 0$), inhibition is *'private'*: Each 'parent' E neuron projects to a single 'daughter' I neuron with equal activity. This has been an implicit assumption of previous predictive coding models with lateral inhibition [38, 55]. In the opposite extreme ($\lambda_{EI} = 1$), each I neuron receives a large number of excitatory inputs and signals an *'internal prediction'* of one stimulus learned by the network, similar to previous models with segregated neural populations [35, 36]. We investigated the continuum of inhibitory representations between these extremes using the same approach applied to E neurons (Fig. 1b-e, Fig. 4b). We started with the alignment of inhibitory responses to stimulus x in the match ($\boldsymbol{r}_{xy}$) and mismatch

($\boldsymbol{r}_x$) conditions, at different learning stages (Fig. 5c). Before learning ($\mu = 0$), increasing λ_{EI} leads to a marked decrease in the alignment of inhibitory responses. After learning ($\mu \approx 1$), increasing λ_{EI} leads to a non-monotonic effect on alignment. Intriguingly, for $\lambda_{EI} = 1$, after learning, the alignment of I responses in the two conditions is larger than that of E responses (Fig. 5c, compare green and black for $\mu = 1$).

These properties allowed us to estimate the parameter λ_{EI} based on empirical measurements of regular-spiking (RS, putative excitatory) and fast-spiking (FS, putative inhibitory) neurons. To achieve that, we computed the correlation between auditory cortex match and mismatch responses, separately for RS and FS neurons recorded in Ref. [12], and then compared those correlations to the model before and after learning (Fig. 5d). Specifically, The pairing between movement and a probe sound (not presented during training) was regarded as *before*-learning and the pairing between movement and the expected sound as *after*-learning (Methods). This correlation decreased significantly during learning for RS neurons, consistent with the change in the model's E population responses (Fig. 5c, blue circles). By contrast, correlation of FS population responses did not change significantly during learning, which rules out small values of λ_{EI}. Moreover, the correlation value after learning was similar for RS and FS neurons, which rules out large values of λ_{EI}. Taken together, our analysis suggests that an intermediate value of $\lambda_{EI} \approx 0.6$ best captures the experimental observations, consistent with the suggestion of 'promiscuous' inhibitory connections mediating suppression of expected stimuli [11].

Given this experimentally-constrained value ($\lambda_{EI} = 0.6$), our theory generates testable predictions for inhibitory predictive representations. First, we expect that anti-alignment of mismatch I responses (x-only, y-only) is significantly weaker when compared to anti-alignment of E responses in the same conditions (Fig. 5e, left; Fig. 1d,e). Second, we predict large correlations between inhibitory responses in the match and y-only mismatch conditions (Fig. 5e, middle), when compared with E responses. The asymmetry of $\boldsymbol{r}_x \cdot \boldsymbol{r}_{xy}$ and $\boldsymbol{r}_y \cdot \boldsymbol{r}_{xy}$ overlaps in the model may in the future be related to distinct functional responses of inhibitory neuron subtypes [23, 56]. Third, the fraction of I neurons with *R* responses decreases moderately during learning, compared to E neurons. We note however that the fraction of E neurons with *R* responses shows moderate dependence on the threshold, particularly before learning (Fig. S6), which may make it challenging to detect differences in fractions of neurons with *R* responses between E and I neurons.

Previous work on predictive coding suggested that associative learning enhances top-down inhibitory projections from outside the local circuit [2, 16], which cancels bottom-up excitation and suppresses neural responses in the match condition. We therefore wondered what changes in inhibitory connectivity during learning lead to stimulus-specific suppression of neural activity in our E/I network model. One option is that inhibitory connections that predict the stimulus are strengthened [2]. Alternatively, inhibition could undergo more subtle reorganization such that inhibitory signals are distributed differently before and after learning.

We calculated the distribution of I-to-E synaptic weights before and after learning in the family of E/I network models. When inhibition is private ($\lambda_{EI} = 0$), this distribution broadens during learning (Fig. 5f). Examining the change in synaptic weights conditioned on the functional cell-type of pre- and post-synaptic neurons (*R* or *PE*), suggests that stimulus-specific suppression of E responses arises from potentiated I synapses from neurons ‘faithfully’ representing the stimulus. In other words, when inhibition is private, the predictive signal arises in part due to strengthened projections from inhibitory *R* neurons to excitatory neurons (Fig. S7). By contrast, when inhibitory structure was matched to experimental data ($\lambda_{EI} = 0.6$), learning leads to overall sparsification of I connections (Fig. 5g). Interestingly, here *R*-to-*R* connections can be either potentiated or depressed, unlike the $\lambda_{EI} = 0$ case (compare middle panel of Fig. 5f,g). Moreover, when $\lambda_{EI} = 0.6$, inhibitory connections originating from *PE* neurons that are initially very weak get strongly potentiated.

Together, our results suggest that (*i*) Predictive processing is learned without large increases of the average inhibitory connection strength. This was also seen for other values of λ_{EI} (Fig. S8). (*ii*) The ‘strategy’ for learning predictive processing can differ substantially, and depends on the underlying circuit structure (different values of λ_{EI} in the model). (*iii*) When inhibitory structure is matched to data, the ‘internal model’ is highly distributed and, surprisingly, arises in part from potentiated connections from inhibitory neurons signaling prediction-error. Another signature of this distributed strategy is the *decrease* of total inhibitory input to each excitatory neuron during learning (Fig. S8), which suggests that predictions are primarily computed by recurrent circuitry rather than directly from top-down inputs.

Predictive representations in hierarchical neural networks

Sensory brain regions are known to have a laminar structure, and distinct layer-specific response characteristics in associative learning tasks [17, 20, 57]. In the context of the task involving sensorimotor predictions, it has been suggested that motor-related input originates from motor regions and first enters the primary sensory region via deep layers (L5/6) [2, 58–60]. On the other hand, the bottom-up sensory-related inputs first enter the primary sensory region via L4, which further projects to L2/3 where the bottom-up and top-down inputs are integrated and processed [61, 62]. To investigate the effects of the laminar structure on predictive processing, we extended the recurrent network model which has a single-*module* and no hierarchical structure, to a network model with three recurrently interconnected *modules* (Fig. 6). During associative learning, the network receives paired multimodal inputs. Crucially, the first module (M1) of the network receives inputs from one modality, and the last module (M3) receives inputs from the other modality (Fig. 6a). Differently from previous work [17, 39, 63], each module in our network computes bidirectional predictions, corresponding to inputs from the level above and below it in the hierarchy. For example, M2 computes predictions of activity in M1 and M3. Our hierarchical model can also be applied to cross-modal processing performed by distinct brain regions that exchange predictive signals bidirectionally (e.g., auditory and visual cortices, [16]), beyond laminar organization within a single brain-region.

We first studied the effects of module-specific gain parameters. After learning, the average mismatch responses increase monotonically with b_1 and b_2 (Fig. 6b). We constrained the average mismatch response to be larger than certain threshold value and minimized the match responses for each module. Doing so gave a continuous set of parameter combinations for which the network satisfies those constraints (Fig. 6b, magenta line). We fixed b_2 such that the fraction of prediction error neurons in M2 after learning is similar to the fraction in the single-module model (Fig. 3b), which also fixes b_1 and b_3 (Fig. 6b, star). With these constrained parameters, we assessed how associative learning shapes neural representations across different modules.

In the x-only mismatch condition ($x = 1, y = 0$), the overall mismatch responses increase during learning, with notable module-specific differences (Fig. 6c): neurons in M1 that directly receives the x-stimulus input have remarkably similar responses in the match and mismatch conditions throughout learning. In contrast, neurons in M3 respond predominately to stimulus y but gradually become tuned to stimulus x as learning progresses. Neurons in M2 exhibit the largest mismatch-match response ratio and develop the most significant prediction error responses after learning.

Next we categorized neurons along the hierarchy into functional cell-types. Before learning, neurons activated by the stimulus x independently of y (i.e., x representation neurons) are concentrated in M1–the module receiving the stimulus x input directly. During learning, x representation neurons arise also in M2 and M3, though the overall fraction of these neurons decreases from M1 to M3 (Fig. 6d). PE neurons are initially very rare, and emerge in all modules during learning, with the largest fraction concentrated in M2 (Fig. 6e). These results are consistent with activity of layer-specific primary sensory cortex neurons [12, 20, 58].

We finally evaluated the network responses for two stimulus-pairs. Similar to the single-module network model, mixed representation neurons arise in all modules after learning (Fig. 6f, g). The fraction of mixed representation neurons is maximal in M2, and it increases with the number of learned stimulus-pairs (Fig. 6f,g). We found that the more pronounced desegregation of neural representations is accompanied with a significant decrease in the median balance level in that module (Fig. 6h), suggesting that loose balance is the underlying circuit mechanism supporting the mixed predictive responses at the cellular level. Unlike our findings in M2, the fraction of mixed representation neurons and the median balance level in M1 and M3 do not show strong dependence on the stimulus dimensionality. These results highlight the impact of anatomical structure on shaping network function. Specifically, we found that different modules have different fractions of representation and prediction error neurons, reminiscent of recent experimental findings [18]. However, despite this heterogeneity, representations of stimuli and prediction error are desegregated in all modules after learning.

Discussion

We investigated the neural representations formed in a class of recurrent neural networks that learn to generate high-dimensional predictions in natural conditions. Our mathematical analysis reveals key neural mechanisms supporting high-dimensional predictive coding; generates novel testable hypotheses for functional properties of the corresponding neural circuits; and provides a framework within which experimental data of large-scale neural recordings can be quantitatively analyzed. Additionally, our framework allows incorporating information on cell-types and anatomical structure into the model, which can elucidate their role in predictive computations.

We focused on a *recurrent* network model (Fig. 1) for two reasons. First, cortical circuitry that performs predictive processing is known to be highly recurrent. Plasticity of recurrent connections forms functional neuronal assemblies [64], which were suggested to underlie behaviorally-relevant sensory discrimination [65]. Second, predictions for sensory stimuli typically unfold over time, which can be naturally implemented by intrinsic dynamics of recurrent networks [32, 66]. While we focused on steady-state neural responses for mathematical tractability, our model could be extended in the future to study the temporal properties of high-dimensional predictive coding. Other interesting directions to extend our study are: networks with asymmetric connectivity, which could be done by imposing sparse connectivity [67]; and networks that learn predictions online [68, 69].

Our model suggests that balance between feedforward and recurrent input, or indeed between excitation and inhibition, can lead to robust internal predictions within local circuits. While this has been suggested previously [32–34, 70, 71], an important novel prediction revealed by our analysis is that in realistic conditions there is an optimal, finite balance level, which decreases with stimulus dimension (Fig. 2). Our theory further suggests that a network with infinitely high balance [33] could be especially vulnerable to noise in high-dimensional scenarios.

Based on our results, we hypothesize that the large degree of heterogeneity of empirical E/I balance levels in different experiments [51] may be a signature of the differences in the stimulus ensembles animals were exposed to. Our results in Fig. 2 and Fig. 3 suggest that this hypothesis could be tested systematically by exposing animals to increasingly rich sensory environments. Here too the temporal dynamics of the network may be important, as synaptic delays may affect the optimal degree of balance in circuits performing low-dimensional predictions [34, 72].

The role that balance plays in computing predictions has important implications for the source of predictive signals and the timescale of learning them. (*i*) Previous work has shown that cross-modal predictions are often *stimulus-specific* [12, 16, 48]: signals from one brain region can suppress responses to a particular predictable stimulus in another region (e.g., motor cortex activity suppressing visual cortical responses). It is notable that within our model those computations are performed without fine-tuning long-range projections [2]. Rather, local recurrent connections in the 'receiving region' can extract the predictions from long-range inputs with 'promiscuous' connectivity [11], relying on E/I balance and activity-dependent synaptic plasticity. (*ii*) Prediction-error responses in the same cortical region can arise at very different timescales, from as little as minutes [26] to days of training [12, 16]. We believe that the diversity of the identified E/I balance mechanisms (e.g., firing-rate adaptation, synaptic-scaling, Hebbian plasticity; see review in Ref. [43]), may explain this wide temporal range of predictive processing learning dynamics. Future work may reveal that our model has explanatory power also for the emergence of predictions over faster timescales than the experiments considered here and thus could be applied to predictive processing circuits in subcortical regions and in invertebrates.

An important finding of our work is that predictive representations are *desegregated*: neurons that signal prediction-errors for one stimulus-pair may faithfully represent the presence of stimulus for a second pair. Based on experiments where animals were probed with multiple types of unexpected sounds, we found signatures of this desegregation at the cellular-level in mouse auditory cortex (Fig. 4). Another recent study in mice performing multiple stereotyped motor actions reported mixed representations of the motor variables and reward prediction-errors across the neocortex [73], as suggested by our model for high-dimensional scenarios. Our model differs from previous work (e.g., [17,39,49,63]) by not explicitly assuming that separate neural populations encode prediction and prediction errors. Rather, the network develops mixed neural representations as a direct consequence of minimizing the multimodal prediction errors under energy constraints.

Our findings are related to the expanding literature on *mixed-selectivity* [74–76], where neurons exhibit complex tuning to multiple stimulus features. While even a random network can exhibit mixed-selectivity [75], the neurons' tuning curves there are unstructured, which requires finely-tuned decoders to readout task-relevant variables. Here we report neurons that have mixed-selectivity to internally generated predictions of sensory and motor variables (Figs. 4, 5, 6). Crucially, the learned neural representations in our model are highly structured, and enable the reading out different stimulus features without 're-learning' the decoder (Fig. 2).

Although neurons in our model network and in electrophysiological recordings from auditory cortex have mixed selectivity for stimuli and prediction-errors, the auditory cortex also contains neurons that more specifically encode prediction-errors [13]. Notably, the abundance of neurons with pure or mixed selectivity to stimulus and error could be also layer-specific [12]. This is recapitulated by our hierarchical network model (Fig. 6). Recent work in the mouse visual cortex identified specific genetic markers that are over-expressed in neurons that encode positive versus negative prediction-errors [18]. The differences in methodologies and time courses of analysis make direct comparisons across these studies difficult, and it remains possible that sensory cortex contains a large population of neurons that have shared roles in encoding stimuli and prediction-errors, as well as neurons that more strictly encode one or the other. Indeed, our analysis reveals that those classes of neurons may exist in different modules within a single network.

In summary, predictive processing is a ubiquitous and fundamental computation supporting diverse behaviors across animal species. Here we take a first step towards bridging the gap between theory of predictive processing and circuit-level neural recordings in predictive processing paradigms. Our results reveal the functional roles of specific circuit motifs and mechanisms in performing multimodal high-dimensional predictive processing. In a broader context, our work will advance the understanding of how the brain constructs complex internal-models by shedding light on commonalities and differences between biological predictive coding circuits and artificial systems, particularly those trained using self-supervised algorithms [39, 77].

Methods

Recurrent network model

Our model network consists of N neurons whose firing-rates are described by the time-dependent vector $\boldsymbol{r}(t) = (\, r_1(t), \ldots, r_N(t)\,)$. The network is driven by high-dimensional stimulus input, denoted $\boldsymbol{x}(t) = (\, x^1(t), \ldots, x^P(t)\,)$ and $\boldsymbol{y}(t) = (\, y^1(t), \ldots, y^P(t)\,)$. The vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ correspond to stimuli from two modalities that are paired during training.

The dynamics of the recurrent network are given by

$$\frac{\mathrm{d}h_i(t)}{\mathrm{d}t} = -h_i(t) + b\left(\underbrace{\sum_{j=1}^{N} J_{ij}\phi(h_j(t))}_{-I_i^R} + I_i^F(\boldsymbol{x}(t), \boldsymbol{y}(t))\right). \tag{1}$$

Here $h_i(t)$ is the voltage level of each neuron and is related to its firing-rate via a nonlinear activation function, $r_i(t) = \phi(h_i(t))$. Note that the input each neuron receives in Eq. (1) is decomposed into the recurrent (I_i^R) and feedforward (I_i^F) components. We rescaled the connectivity matrix J_{ij} and the feedforward input $I_i^F(\boldsymbol{x}(t), \boldsymbol{y}(t))$ by a constant b, which can be interpreted as a gain parameter.

The explicit forms of J_{ij} and $I_i^F(\boldsymbol{x}(t), \boldsymbol{y}(t))$ were determined based on a normative approach as follows (derivation details appear in SI §1). We assume that the neurons' dynamics jointly minimize the following objective

$$E(t) = \underbrace{\sum_{k=1}^{P} \left[\left(x^k(t+d) - \hat{x}^k(t)\right)^2 + \left(y^k(t+d) - \hat{y}^k(t)\right)^2 \right]}_{\text{Prediction-errors}} + \underbrace{\frac{2}{b}\sum_{i=1}^{N} F(r_i(t))}_{\text{Encoding efficiency}}, \tag{2}$$

where $\hat{x}(t)$ and $\hat{y}(t)$ are the internal predictions generated by the network at time t and $F(r)$ is a monotonically increasing function whose explicit form depends on ϕ, the nonlinear activation function (SI §1.1). For ReLU nonlinearity $[\phi(z) = \max(z - \theta, 0)]$, $F(r) = (r + \theta)^2/2$. Minimizing Eq. (2) is equivalent to performing Bayesian inference to extract the latent 'cause' of the sensory signals (SI §1.2). Furthermore, our network model generalizes previous models of predictive coding [1, 36, 38, 39, 63], by incorporating the effect of response nonlinearity into a regularization term that controls encoding efficiency. We note that the parameter b in Eq. (2) controls a trade-off between minimizing prediction-errors and maximizing encoding efficiency.

We further assume that the internal predictions are linear readouts of the network activity

$$\hat{x}^k(t) = \frac{1}{N}\boldsymbol{w}^k \cdot \boldsymbol{r}(t), \qquad \hat{y}^k(t) = \frac{1}{N}\boldsymbol{v}^k \cdot \boldsymbol{r}(t). \tag{3}$$

Here $\boldsymbol{w}^k, \boldsymbol{v}^k \in \mathbb{R}^N$ are the readout weight vectors. These internal predictions are, by definition, predictions of future input, as indicated by the delay d in Eq. (2). However, we will focus on the scenario where the input changes much more slowly than the neurons' firing-rates. Therefore, on the timescale of firing-rate changes [Eq. (1)], we will regard the stimulus inputs to be approximately constant, i.e.,

$$x^k(t+d) \approx x^k(t) \approx x^k, \qquad y^k(t+d) \approx y^k(t) \approx y^k. \tag{4}$$

We assume that the weight vectors $\boldsymbol{w}^k$ and $\boldsymbol{v}^k$ change during learning so as to minimize the objective function $E(t)$ [Eq. (2)]. This optimization process can be viewed as weight-changes governed by a combination of gradient descent on the squared prediction error in Eq. (2), and homeostatic plasticity (SI §1.1). If weights are initialized randomly, learning increases the correlation between the weight vectors (SI §1.1). Specifically, we show that in the large network size limit ($N \to \infty$), the weight vectors have the following statistics,

$$\begin{aligned} \langle w_i^k \rangle = \langle v_i^k \rangle &= 0, \\ \langle (w_i^k)^2 \rangle = \langle (v_i^k)^2 \rangle &= 1, \\ \langle w_i^k v_i^k \rangle &= \mu^k. \end{aligned} \tag{5}$$

Here w_i^k and v_i^k are the components of $\boldsymbol{w}^k$ and $\boldsymbol{v}^k$, which have zero mean and unit variance due to homeostatic plasticity. The correlation between them is μ^k, which increases during learning (i.e., as the objective function E decreases). These weight changes can also arise from local plasticity rules applied to dendritic compartments (SI §1.3). For simplicity, unless noted otherwise, all stimulus-pairs have the same 'age', i.e., $\mu^k = \mu$ does not depend on the index k. We further assume that the weight vectors have multivariate Gaussian distribution. Under these assumptions, we obtained analytical solutions for the dependence of steady-state firing-rate distribution on the stimulus input and the correlation μ in two limits (SI §2): the high-dimensional case where both N and P are large, and their ratio $\alpha = P/N$ is finite; and the low-dimensional case where only N is large, and $\alpha = 0$.

The presence or absence of each stimulus was modeled by setting the corresponding components of $\boldsymbol{x}$ and $\boldsymbol{y}$ to 0 or 1. For example, the mismatch and match conditions for the k-th stimulus-pair correspond to,

$$\begin{aligned}
(x^k, y^k) &= (1, 0) \qquad (x\text{-only mismatch condition}),\\
(x^k, y^k) &= (0, 1) \qquad (y\text{-only mismatch condition}),\\
(x^k, y^k) &= (1, 1) \qquad (\text{match condition})
\end{aligned}$$

We extended our results to apply in scenarios with associations between more than two stimuli (SI §1.3).

Geometry of representations of stimuli, predictions and prediction-errors

Under the above assumptions, the steady-state neural response vector [Eq. (1)] can be expressed as,

$$\boldsymbol{r} \propto \left[\boldsymbol{a}_x(\mu)x + \boldsymbol{a}_y(\mu)y + \sqrt{\alpha}\cdot \text{noise}\right]_+ . \tag{6}$$

This form is revealing, since the stimulus-specific, μ-dependent vectors $\boldsymbol{a}_x(\mu), \boldsymbol{a}_y(\mu)$ correspond to the directions along which the network encodes the stimuli in the x-only and y-only mismatch conditions. Eq. (6) also shows that, owing to the nonlinearity, the readout in the matched condition is not $\boldsymbol{a}_x(\mu) + \boldsymbol{a}_y(\mu)$. The geometry of representing stimuli in the match and mismatch conditions is illustrated in Fig. 1d. Changes to these vectors during training (i.e., μ increases) correspond to the learned structure of neural representations of stimuli and prediction-errors. We further note that the magnitude of the noise in Eq. (6) depends on the stimulus dimensionality α, and thus it captures the interference between learned stimuli.

Definition of balance level

The balance level for neuron i is defined as,

$$B_i = \left| \frac{I_i^F}{I_i^F - I_i^R} \right| . \tag{7}$$

Here, I_i^F and I_i^R are the feedforward and recurrent input currents to neuron i at steady-state [Eq. (1)]. The balance level varies between neurons and between stimuli, because the weights w_i^k and v_i^k are different for different neurons and stimuli (indexed by i and k, respectively). The balance level distribution and its median shown in Fig. 2 were computed analytically (SI §2.3).

Extracting the optimal balance level from experimental data

V-M experiment, Ref. [20]. We calculated the trial-averaged voltage of all the recorded L2/3 neurons as a function of time (Fig. 3a). Voltage level of each neuron was measured with respect to its baseline. We sampled 50 voltage levels from all recorded neurons and all time points in the match and mismatch time windows (Fig. 3a), which were $-0.1 - 0$s (match) and $0 - 0.1$s (mismatch). The time $t = 0$ corresponds to point at which the treadmill was decoupled from visual flow in virtual reality. We then computed the standard deviation over those 50 samples of the voltage level in the match and mismatch conditions. By taking the ratio of these standard deviations, we obtained a dimensionless quantity that has a direct analog in the model: the standard deviation of h_i over neurons in the network in Eq. (1). Specifically, for $P = 1, \theta = 0$, we computed this ratio explicitly (SI §2.1),

$$\frac{\sigma^2_{\text{mismatch}}}{\sigma^2_{\text{match}}} = \frac{1}{2} \frac{\mu^2 + (1-\mu^2)(1+b/2)^2}{\mu^2 + \mu + (1-\mu^2)[1 + b + (1-\mu)b^2/4]}. \tag{8}$$

We use $\mu = 0.97$ as the correlation value after training and fit this formula to the ratio obtained from data by adjusting the value of b. Using the best-fit value $b^\star$, we computed the median of balance level $B^\star$ in the network model (Fig. 3c).

A-M experiment, Ref. [12]. We calculated the trial-averaged firing-rates for all regular spiking neurons ($n = 815$) in the passive (mismatch) and movement (match) condition in two time windows: from $t = -0.1$s to stimulus onset ($t = 0$), and from stimulus onset to $t = 0.06$s (Fig. 3b). For every neuron, we calculated the change in its firing-rate between the two time windows in both conditions. We sampled 400 firing-rate change values from 815 neurons with replacement, and calculated the average firing-rate change in the passive and movement conditions. We computed the equivalent quantity in the model, i.e., average of $\phi(h_i)$ over neurons in the network [Eq. (1)] in the match and mismatch conditions. For ReLU activation function, the ratio is also given by Eq. (8) and can be fit to the ratio obtained from the data by adjusting the parameter b. Again we calculated the median of balance level $B^\star$ based on the best-fit value of $b^\star$. The fitting procedure for both experiments was repeated 100 times, giving the scatter plot of estimated $B^\star$ values (Fig. 3c).

Definition of functional cell types

We denote the steady-state voltage of neuron i in the mismatch conditions as h_i^x (x-only) and h_i^y (y-only), and in the match condition as h_i^{xy}. To classify neurons into functional types, deviations of individual neurons' voltage response relative to the mean were compared to the standard deviation (denoted σ) of the steady-state voltage distribution. We evaluated σ using the voltage distribution in the x-only mismatch condition after learning ($\mu = 0.97$).

A neuron i is a representation (*R*) neuron for the x-stimulus if it is depolarized upon presentation of the stimulus x, i.e., its voltage response in x-only mismatch condition is large, and its voltage responses in the match and mismatch conditions are similar. Mathematically,

$$h_i^x > \frac{\sigma}{2} \quad \text{and} \quad |h_i^x - h_i^{xy}| < \frac{\sigma}{2}. \tag{9}$$

A similar criterion was used to identify *R* neurons for the y-stimulus. A neuron i is a prediction-error (*PE*) neuron if it signals the 'mismatch' between x and y, i.e., its voltage response in the x-only mismatch condition is large, and its voltage response in the match condition is small. Mathematically,

$$h_i^x > \frac{\sigma}{2} \quad \text{and} \quad h_i^x - h_i^{xy} > \frac{\sigma}{2}. \tag{10}$$

Neurons meeting these criteria are referred to as *positive* PE neurons, because their activity increases when x is presented but not expected (based on y). The activity of *negative* PE neurons increases when x is not presented but is expected. In our model, E neurons have a centered (zero mean) distribution of voltages for $\alpha = 0$, therefore the threshold is applied to the voltage itself. For excitatory neurons in the high-dimensional regime ($\alpha > 0$) and inhibitory neurons, since their voltage distribution has a non-zero mean, we used the centered voltage levels (h_i^x, h_i^{xy}) in the above criteria.

Note that neurons in the network may not belong to any of the those three classes (Fig. S3a). We computed the firing-rate statistics of neurons in the network analytically (SI §2, §3), which allowed use to obtain the fraction of *R* and *PE* neurons for different values of μ and α, shown in Fig. 4b,d. We further explored the effects of threshold level on the fraction of different functional types in Fig. S3b.

Estimating functional segregation from responses to multiple stimuli from experimental data

We calculated the trial-averaged firing-rate change of each neuron in the match (active) and mismatch (passive) conditions, separately for each sound stimulus from our experimental data [13]. To calculate the segregation index for each type of probe sound, we restrict the analysis to neurons responsive in the passive condition to that probe sound and the learned (expected) sound. Responsive neurons were defined as those having firing-rate that was one half of the standard deviation above the mean firing-rate for the expected sound in the passive condition. Changing the threshold does not affect the results in Fig. 4e,f. For these neurons, we computed pairs of Δ values, defined as the difference between mismatch and match responses, for the probe and expected stimulus. The Pearson correlation coefficient between those Δ values was defined as the segregation index.

To estimate the similarity of the expected and probe stimuli, we computed individual neurons' trial-averaged firing-rate change following presentation of those stimuli in the passive condition from our experimental data [13] (the same time windows used in the A-M experiment, Fig. 3). For each animal, we considered population firing-rate vectors consisting of all its recorded neurons. Representation similarity was defined as the Pearson correlation of those vectors for pairs of auditory stimuli (expected and probe, Fig. 4f). We note that this similarity in the model is calculated from the activity of all neurons that are active in either the expected or probe stimuli in passive condition.

E/I network model

In the network with separate E and I neurons, the time-dependent voltages of E and I neurons are given by the following set of differential equations,

$$\begin{aligned}\frac{\mathrm{d}h_i^E}{\mathrm{d}t} &= -h_i^E + \sum_{j=1}^{N_E} J_{ij}^{EE}\phi(h_j^E) + \sum_{j=1}^{N_I} J_{ij}^{EI}\phi_I(h_j^I) + I_i^E,\\ \tau_I\frac{\mathrm{d}h_i^I}{\mathrm{d}t} &= -h_i^I + \sum_{j=1}^{N_E} J_{ij}^{IE}\phi(h_j^E) + \sum_{j=1}^{N_I} J_{ij}^{II}\phi_I(h_j^I) + I_i^I.\end{aligned} \tag{11}$$

We assume that the activation function for inhibitory neurons is ReLU with zero threshold, $\phi_I(x) = \max\{x, 0\}$. Matching the E neurons' activity at steady state to the activity of neurons in our original network [Eq. (1)] gives constraints on the connectivity components and the feedforward input (SI §4),

$$\begin{aligned}J^{EE} - J^{EI}(I + J^{II})^{-1}J^{IE} &= J,\\ \boldsymbol{I}^E - J^{EI}\boldsymbol{I}^I &= \boldsymbol{I}^F.\end{aligned} \tag{12}$$

Here J and $\boldsymbol{I}^F$ are the connectivity matrix and feedforward input used in Eq. (1). We further assume that the matrix $I + J^{II}$ is invertible. In general, there are many possible solutions $\{J^{EE}, J^{EI}, J^{IE}, J^{II}, \boldsymbol{I}^E, \boldsymbol{I}^I\}$ satisfying Eq. (12). We therefore identify a family of solutions. This continuum interpolates between the solution with private inhibition, where J^{IE} is equal to the identity matrix; and solutions with an inhibitory internal prediction, where rows of J^{IE} are given by the stimulus weight vectors (SI §4). Moreover, we show that up to a constant, the balance level defined earlier [Eq. (7)] is the same as the stimulus-specific, local component of the E/I balance level in the E/I network (SI §4).

We extended the definition of functional cell-types [Eqs. (9,10)] to I neurons. We note that here the average input to inhibitory neurons is not 0, so we subtracted the mean from the voltage level [h's in Eqs. (9,10)] before applying the criteria on the deviations from the mean.

Analyzing responses of regular spiking and fast spiking neurons

We estimated the connectivity structure parameter λ_{EI} based on recordings of regular spiking and fast spiking neurons [12]. Using the same time windows as Fig. 3b and Fig. 4e,f, we calculated individual neurons' trial-averaged firing-rate change in the passive and movement conditions for the expected sound and the probe sound. Those firing-rate changes recorded in each animal form eight population vectors (regular/fast spiking, expected/probe sound, movement/passive). We calculated the Pearson correlation between population vectors under movement and passive conditions, giving four values for each animal, shown in Fig. 5d. The correlation values for presentation of the expected sound were regarded as 'after learning', while correlation values for presentation of the probe sound that was not associated with the lever press were regarded as 'before learning'.

Hierarchical recurrent network model

In the hierarchical network model, each neuron belongs to one of three modules, indicated by superscripts in the equations governing neural activity,

$$\begin{aligned}
\frac{\mathrm{d}h_i^1}{\mathrm{d}t} &= -h_i^1(t) + b_1\Big(\sum_j J_{ij}^1\phi(h_j^1(t)) + \sum_k W_{ik}^1 x_k + \sum_{k'} V_{ik'}^1\phi(h_{k'}^2(t))\Big) && \textbf{(M1)}\\
\frac{\mathrm{d}h_i^2}{\mathrm{d}t} &= -h_i^2(t) + b_2\Big(\sum_j J_{ij}^2\phi(h_j^2(t)) + \sum_k W_{ik}^2\phi(h_k^1(t)) + \sum_{k'} V_{ik'}^2\phi(h_{k'}^3(t))\Big) && \textbf{(M2)}\\
\frac{\mathrm{d}h_i^3}{\mathrm{d}t} &= -h_i^3(t) + b_3\Big(\sum_j J_{ij}^3\phi(h_j^3(t)) + \sum_k W_{ik}^3\phi(h_k^2(t)) + \sum_{k'} V_{ik'}^3 y_{k'}\Big) && \textbf{(M3)}
\end{aligned} \tag{13}$$

The definitions of feedforward and recurrent connectivity are generalizations of the single module network. Moreover, this model can be extended to a hierarchical network with a arbitrary number of layers. Details are provided in SI §1.3.

Statistical tests

In Figs. 3c, 4f and 5d, we used two-sided, unpaired t-tests. $^{\star} = p < 0.05$ and $^{\star\star\star} = p < 0.0005$.

Acknowledgments

The authors thank S. Azizpour Lindi, D. Bambah-Mukku, E. Mukamel, and I. Nelken for useful conversations, and I. Nelken for comments on a previous version of the manuscript. This work was supported by DARPA grant D21AP10162-00 (J.A.), DOE grant DE-SC0022042 (J.A.), NIH grants 1R01-NS135853 (J.A.), K99-DC020770 (N.J.A.), 1R01-DC018802 (D.M.S). B.W. thanks the UCSD Friends of the International Center for support. D.M.S. is a New York Stem Cell Foundation - Robertson Neuroscience Investigator. J.A. and B.W. thank the Kavli Institute for Theoretical Physics (KITP), Tel Aviv University, and ICERM at Brown University for hospitality during summer and fall 2023. KITP is supported by NSF grant PHY-1748958 and the Gordon and Betty Moore Foundation Grant No. 2919.02. ICERM is supported by NSF grant DMS-1929284.

Declaration of interests

The authors declare no competing interests.

Author contributions

Developed the project and modeling approach, B.W., J.A. Solved, analyzed and simulated model, B.W. with inputs from J.A. Designed and performed experiments, N.J.A., D.M.S. Designed and performed data analysis, B.W. with inputs from J.A., N.J.A., D.M.S. Wrote paper, B.W., J.A. with inputs from N.J.A., D.M.S. Supervised the project, J.A.

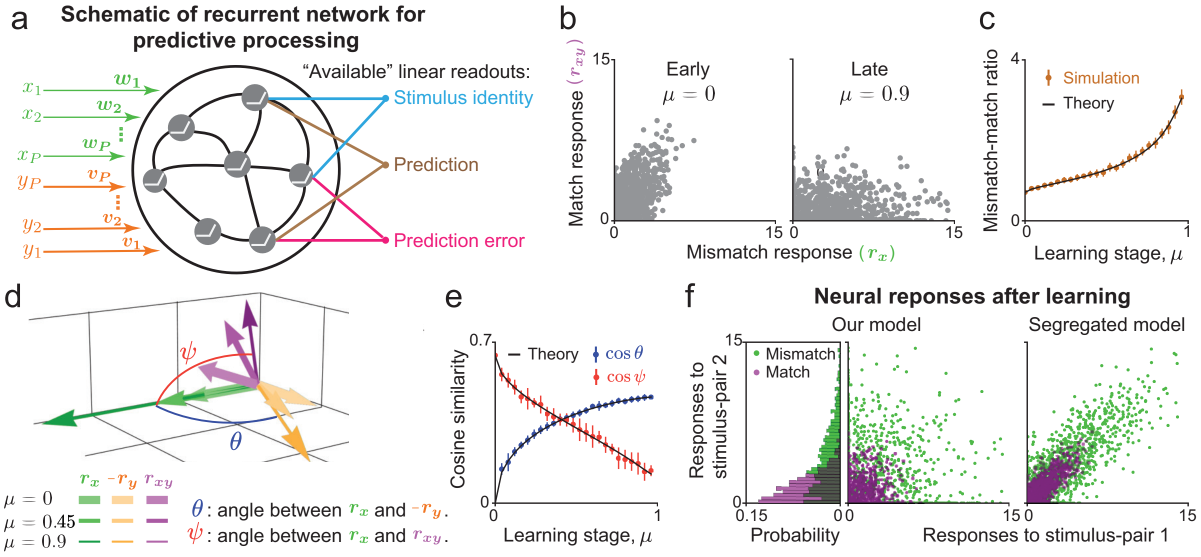


Fig. 1. Emergence of predictive stimulus representations in a recurrent network model during learning. (a) Schematic of a recurrent network model driven by P pairs of stimuli (x and y). Associative training increases the correlations between the feedforward weights carrying the input signals ($\boldsymbol{w}$ and $\boldsymbol{v}$). The recurrent weights jointly optimize prediction-errors and overall encoding efficiency. The neural representation formed under such optimal recurrent connectivity allows reading-out the identity of the presented stimulus; predicting a 'missing' stimulus; and evaluating the prediction-error. (b) Firing-rate responses of individual neurons in the match and mismatch conditions. Initially match and mismatch responses are correlated. After learning, responses are less correlated, and match responses are suppressed while the mismatch responses are amplified. (c) The ratio between average firing-rates in the mismatch and match conditions increases during learning. (d) Reduced three-dimensional neural activity space. Each vector represents the mean-subtracted firing-rate vector of neurons in the network at different conditions and stages of learning. (e) Learning leads to anti-correlation between neural responses to the stimuli x and y when presented separately (blue), and decorrelates the neural responses in the match and mismatch conditions (red), quantified by the angle between the population vectors. (f) Firing-rate responses of individual neurons to two stimulus-pairs in the match and mismatch conditions. In our model (left) there are no correlations between the responses to the two stimuli. Those responses are expected to be strongly correlated in a model in which predictive coding is functionally segregated (right). Error bars indicate standard deviations over 10 instances of the network. See Methods for additional details.

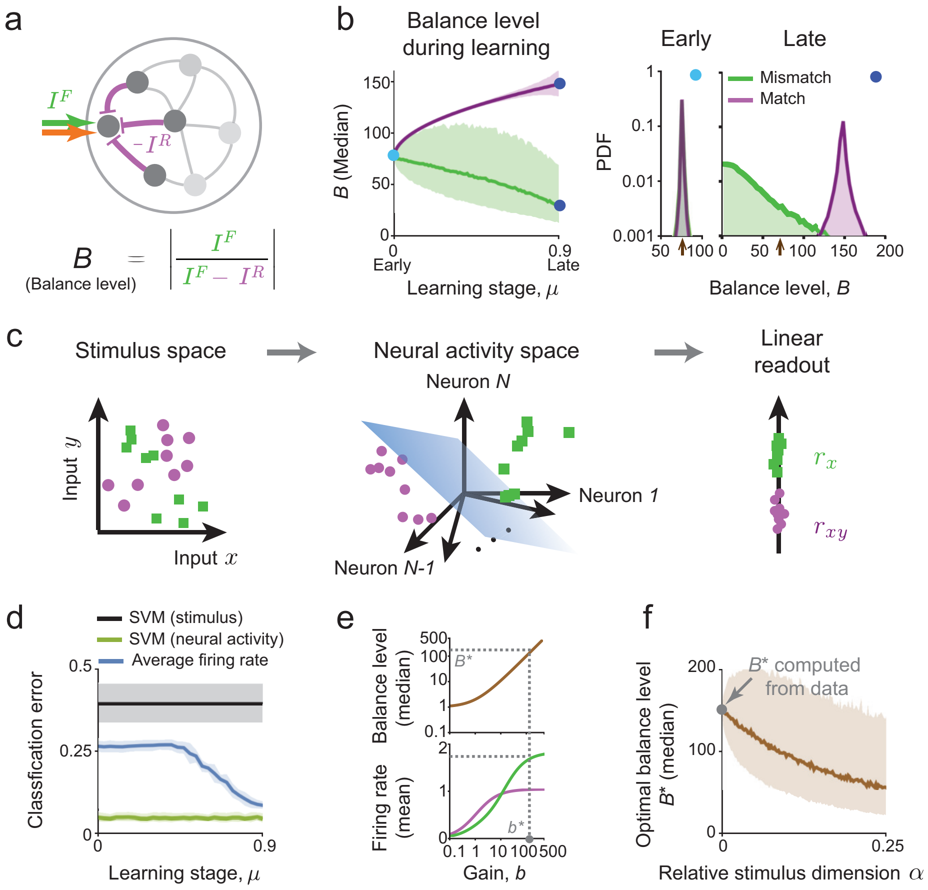


Fig. 2. Balance between feedforward and recurrent inputs is an important mechanism supporting predictive processing. (a) The input to each neuron is decomposed into feedforward and recurrent components, which respectively correspond to the actual stimulus signal and internally generated predictions. Each neuron's balance level B is the ratio between the total feedforward input and the net input (Methods). (b) The median of B in the match and mismatch conditions during learning (left, shaded area indicates inter-quartile range). 'Snapshots' of the distributions of B early and late in learning show that the distributions become separable in match and mismatch conditions (right). The arrows on the x-axis indicate the distribution mode early in learning. (c) Schematic showing the nonlinear transformation from the stimulus space (left) to neural activity space (center), which facilities a linear readout of relevant stimulus features (here, decoding if x is presented in the match/mismatch condition).(d) Error of a support vector machine classifier trained to identify the match/mismatch condition based on the input stimuli (black) and on neural responses (green). After learning, a linear classifier based on the average firing-rate (blue) performs almost as well as the optimal classifier, suggesting that functionally relevant features from all stimulus-pairs can be extracted without re-learning. (e) Illustration of the procedure to determine the optimal $b^\star$. The balance level B increases monotonically with the gain parameter b (top). Increasing b leads to a larger margin between match

and mismatch responses (improved separability) at the cost of higher firing-rates (bottom). The optimal balance level $B^\star$ is determined by constraining the average firing-rate in the mismatch condition and minimizing it in the match condition. (f) Increasing the stimulus dimension leads to decrease in $B^\star$, i.e., a more loose balance (shaded area indicates inter-quartile range). At $\alpha = 0$, we fit $B^\star$ to experimental data (Methods, [20]).

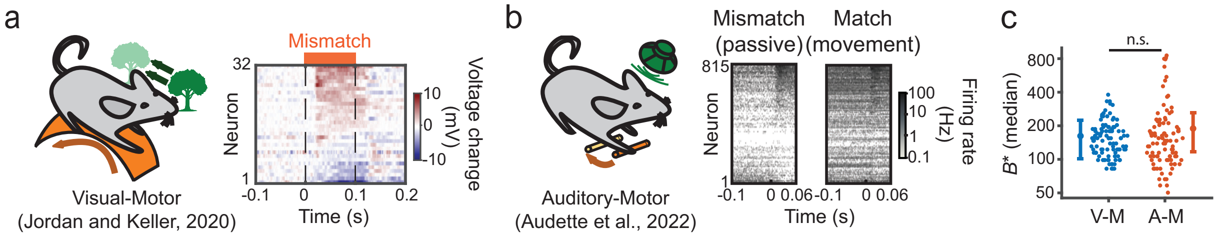


Fig. 3. Estimating the balance level from predictive coding experiments. (a) Schematic of a learned visual-motor association between running and virtual reality visual flow [20]. Voltage levels of different neurons in primary visual cortex reveal tuning to mismatch between running speed and visual flow (prediction-errors). (b) Schematic of a learned audio-motor association between a lever press and a sound [12]. Neurons' firing-rates reveal tuning to auditory stimuli presented without (passive, prediction-errors) and with a lever press (movement). (c) Estimating the median optimal balance level for V-M (blue) and A-M (red) experiments gives similar values. We assume that $\alpha = 0$ based on the fact that the animals underwent extensive training on a single pair of stimuli in both experiments. Error bars are based on repeated subsampling (Methods).

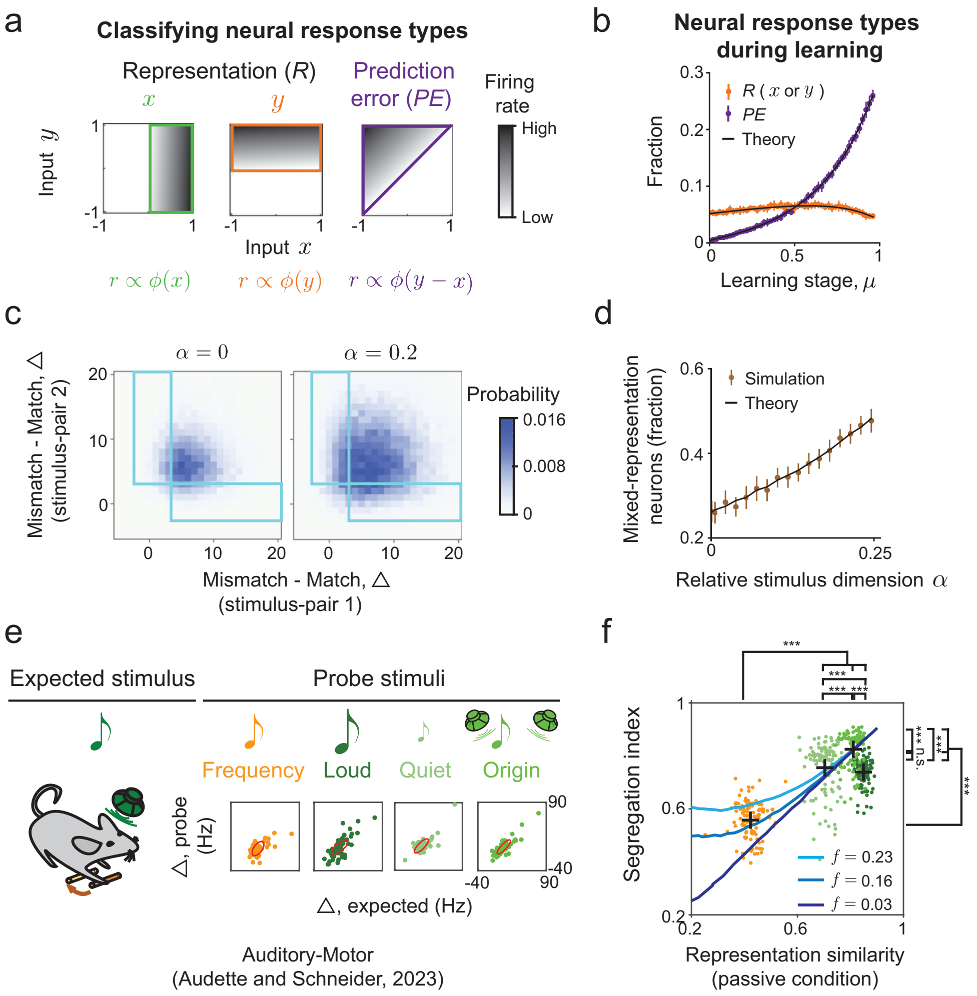


Fig. 4. Desegregated stimulus and error representations in networks performing high-dimensional predictive processing. (a) Schematic of typical tuning profiles of different functional cell-types to the stimuli x and y. (b) Fraction of representation (R) and prediction-error (PE) neurons in the model at different learning stages. Error bars indicate standard deviation over 10 instances of the network. (c) Joint distribution of individual neurons' Δ values, the difference between mismatch and match responses to two specific stimulus-pairs. Only neurons responsive to both stimulus-pairs are included in the distribution (Methods). Mixed representation neurons have significantly different Δ values for the two stimulus-pairs, i.e., they are in the blue rectangular regions. As the stimulus dimension (α) increases, more neurons have a mixed representation of stimuli and prediction-errors. (d) The fraction of mixed representation neurons increases as stimulus dimension increases. Error bars indicate standard deviations over 200 instances of the network. The threshold for defining response types was based on neural

activity statistics at $\alpha = 0$ and was used for all values of α (SI §5). (e) Evaluating the segregation of stimulus and prediction-error representations based on neural recordings during a learned auditory-motor association. Shown are the Δ values of stimulus-responsive neurons for the expected sound and each probe type (colors). Red ellipses indicate the spread of data. The length and direction of major and minor axes correspond to the amplitude and direction of the two leading principal components. (f) Segregation index as a function of representation similarity for different pairs of expected and probe sounds. Colored points correspond to subsamples of the data, and crosses correspond to the average for each probe type (Methods). Experimental data is compared with equivalent quantities from the model, obtained by varying the sparsity of responses in the model (f, see SI §3).

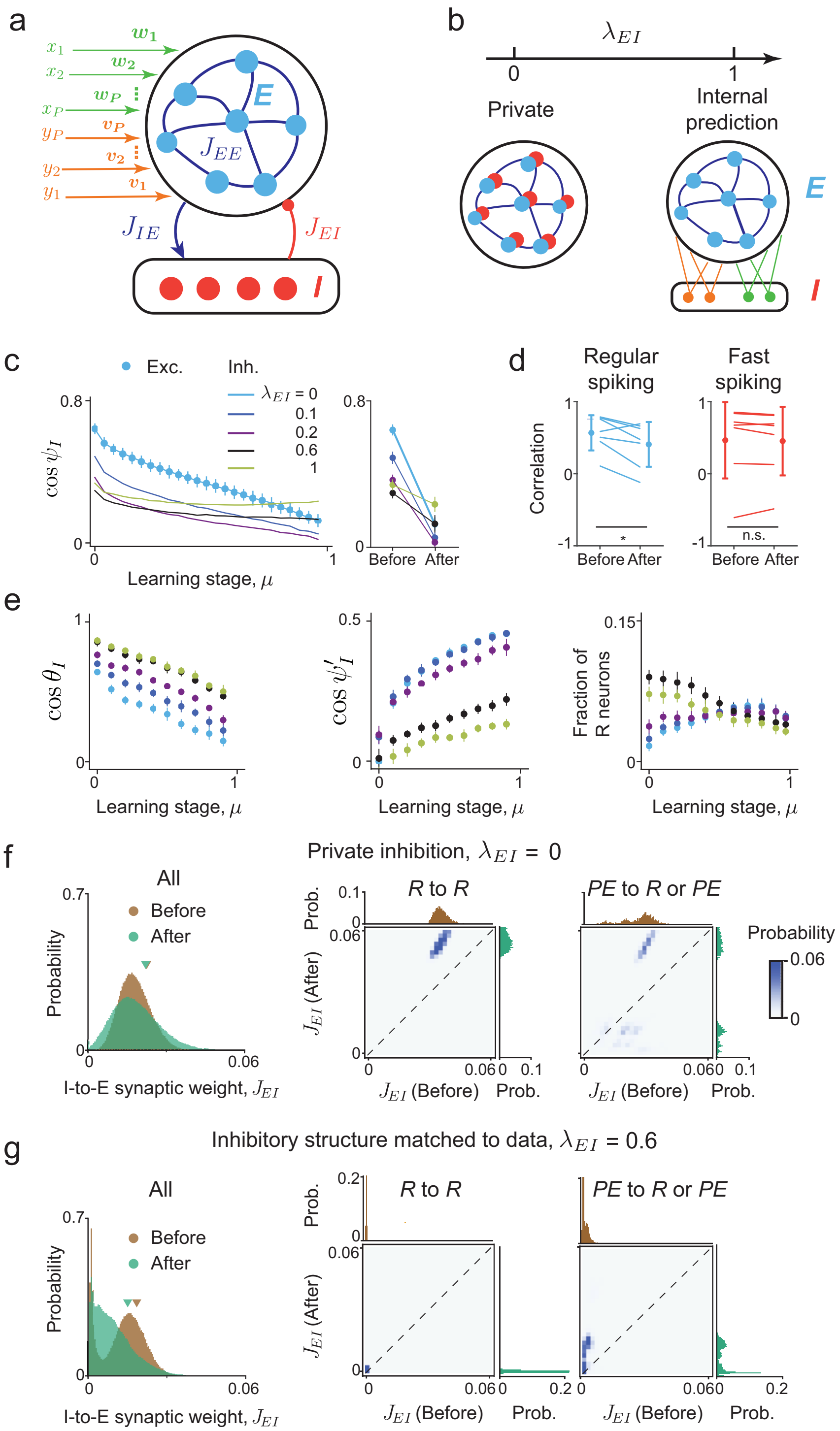


Fig. 5. A data-constrained excitatory/inhibitory model suggests that internally-generated predictions are distributed across the network. (a) Schematic of the E/I network with separate connectivity components. Excitatory neurons receive external inputs, and their activity is

constrained to equal that of neurons in our original model. (b) A family of E/I networks that satisfy the desired constraints, identified based on non-negative matrix factorization. Solutions are parameterized by λ_{EI}, which interpolates between 'private' inhibition and inhibition that signals 'internal predictions'. Varying λ_{EI} gives different patterns of inhibitory responses and connectivity structures. (c) The cosine similarity $(\cos\psi_I)$ between the match and mismatch inhibitory responses to stimulus x ($\boldsymbol{r}_{xy}$, $\boldsymbol{r}_x$), for different values of μ and λ_{EI} (left). Comparing $\cos\psi_I$ before and after learning (right) allowed us to link inhibitory connectivity structure to inhibitory representations. (d) Analogous correlation between population responses, computed separately for regular-spiking (RS) and fast-spiking (FS) neurons from Ref. [12]. Each point represents data from one animal. The mean and standard deviation of the correlations across animals are also shown. RS neurons significantly decorrelate during learning, while FS neurons' correlation does not change. Correlations of RS and FS neurons after learning are similar. (e) The angle θ_I (left) between inhibitory population responses to the paired stimuli in the mismatch conditions ($\boldsymbol{r}_x$, $-\boldsymbol{r}_y$), and the angle ψ_I' (center) between match and mismatch inhibitory population responses to stimulus y ($\boldsymbol{r}_{xy}$, $\boldsymbol{r}_y$). Angles are shown as a function of μ and λ, leading to experimentally testable predictions pertaining to inhibitory representations. Fraction of inhibitory *R* neurons (right) as a function of μ and λ_{EI}. For the experimentally constrained parameter $\lambda_{EI} = 0.6$, this fraction decreases for inhibitory neurons (black), while it does not change significantly for excitatory neurons ($\lambda_{EI} = 0$, blue, Fig. 4b). (f) Synaptic weight distribution of all I-to-E connections before and after learning, when $\lambda_{EI} = 0$ (left), and for pairs of E and I neurons belonging to specific functional classes (*R* to *R*, middle; *PE* to *R* or *PE*, right). Learning broadens the overall synaptic weight distribution and potentiates the inhibitory connections between inhibitory *R* neurons. (g) Same as (f), when inhibitory structure is matched to data ($\lambda_{EI} = 0.6$). Here learning sparsifies and depresses inhibitory connections. Connections between *R* neurons remain very small throughout learning. Surprisingly, connections from inhibitory *PE* neurons are strongly potentiated.

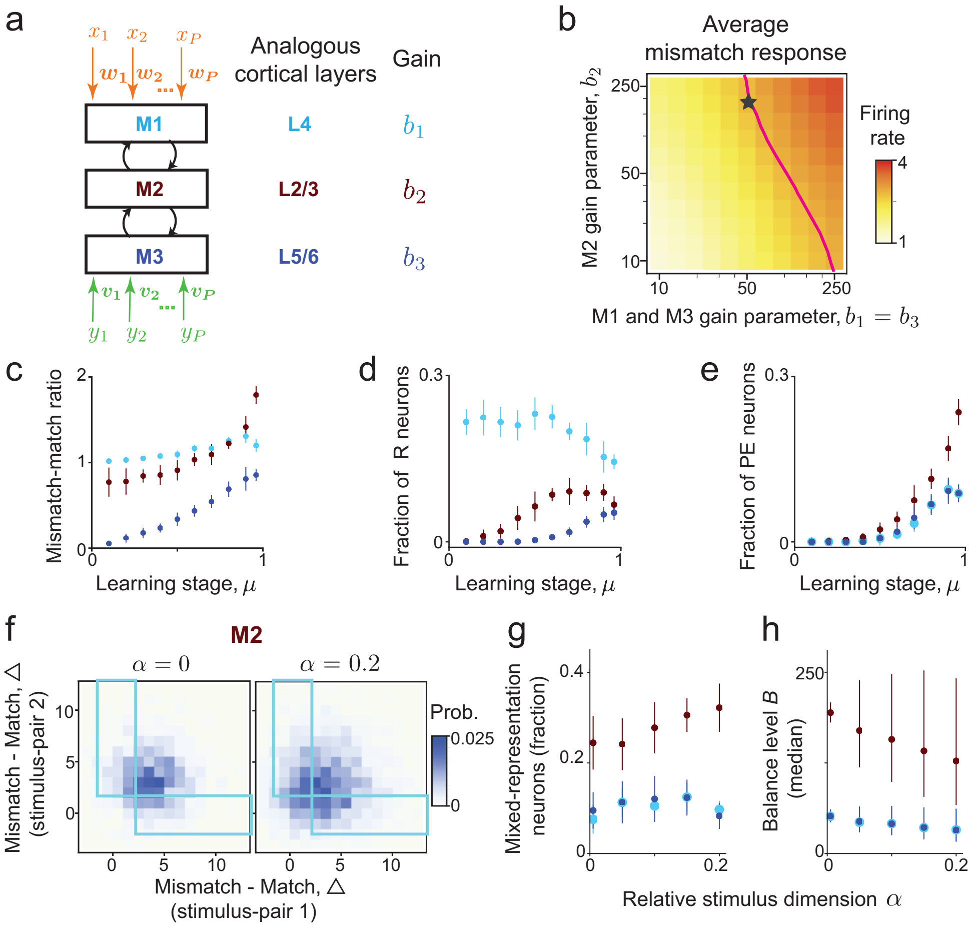


Fig. 6. Representations of stimuli and prediction errors vary across a hierarchical network. (a) Hierarchical network for predictive processing with three modules. M1 and M3 receive stimulus x and y input, respectively. (b) The average x-only mismatch response increases with the module-specific gain parameters $b_{1,2,3}$. Line: mismatch response amplitude used to constrain $b_{1,2,3}$. Star: parameter values further constrained based on the fraction of prediction error neurons in M2, used in panels (c-h). (c) The ratio between the average firing-rates in the x-only mismatch and match conditions increases during learning. The increase is most prominent in M2. (d) The fraction of x representation (R) neurons at different learning stages. Differences between the modules diminish with μ. (e) The fraction of prediction error (PE) neurons at different learning stages. (f) Joint distribution of individual neurons' Δ values, defined the difference between mismatch and match responses to two specific stimulus-pairs in M2. Mixed representation neurons are in the blue rectangular regions. The fraction mixed representation neurons increases with the stimulus dimension α. (g, h) Effects of increasing the stimulus dimension α. (g) The fraction of mixed representation neurons increases with α in M2, and remains constant in M1 and M3. (h) The median balance level decreases with α in M2 and remains approximately constant in M1 and M3. Error bars indicate inter-quartile range.

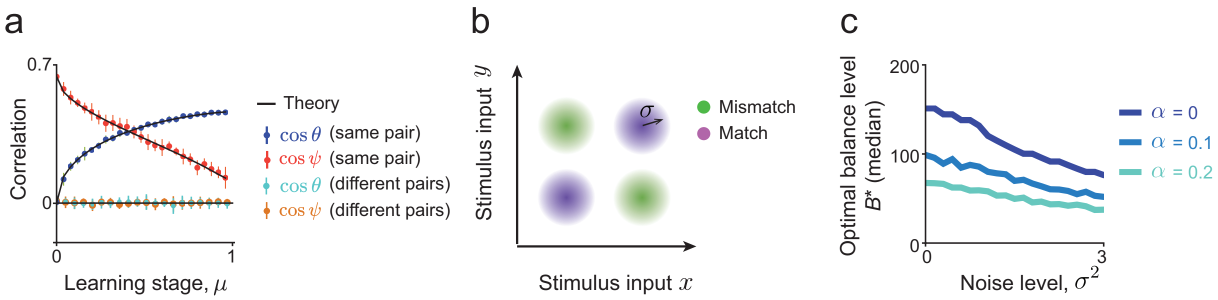


Fig. S1. The geometry of predictive representations in the model. (a) Pearson correlation coefficient between neural responses in different stimulus conditions. As in Fig. 1, the angle θ is measured between the network's responses to the two stimuli in mismatch conditions (i.e., $\boldsymbol{r}_x$ and $-\boldsymbol{r}_y$); while ψ is the angle between responses to the same stimulus in the match and mismatch conditions (i.e., $\boldsymbol{r}_{xy}$ and $\boldsymbol{r}_x$). Neural responses to stimuli from different stimulus-pairs remain uncorrelated, suggesting that the predictive signal learned by the network is stimulus-specific. Here $\alpha = 0$. (b) Schematic of noisy stimulus inputs. Independent isotropic Gaussian noise (with S.D. denoted by σ) is added to the inputs in the match and mismatch conditions, relative to the noiseless stimulus presentation considered in Figs. 1,2. (c) The optimal balance level decreases as stimulus presentation becomes more noisy for all values of α.

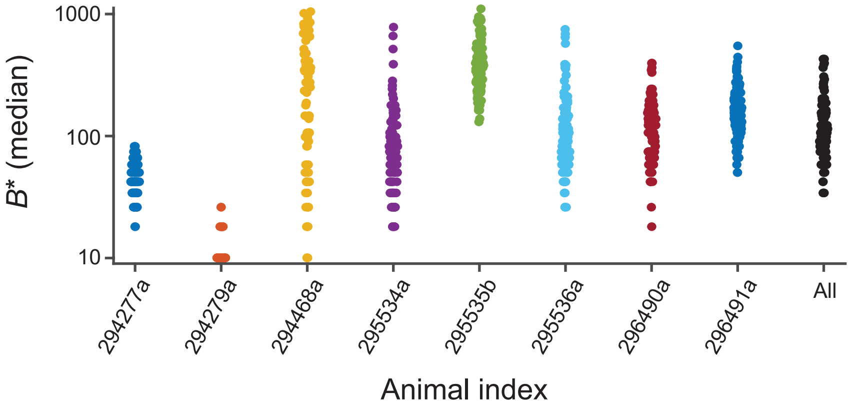


Fig. S2. Estimated balance levels from individual animals. For each animal recorded in [12] ($n = 8$), the balance level was estimated as described in the Methods, sampling the firing-rates separately from each animal. There is marked variability across animals, suggesting that effects of learning multiple stimuli in the future are best studied *within animal* during learning.

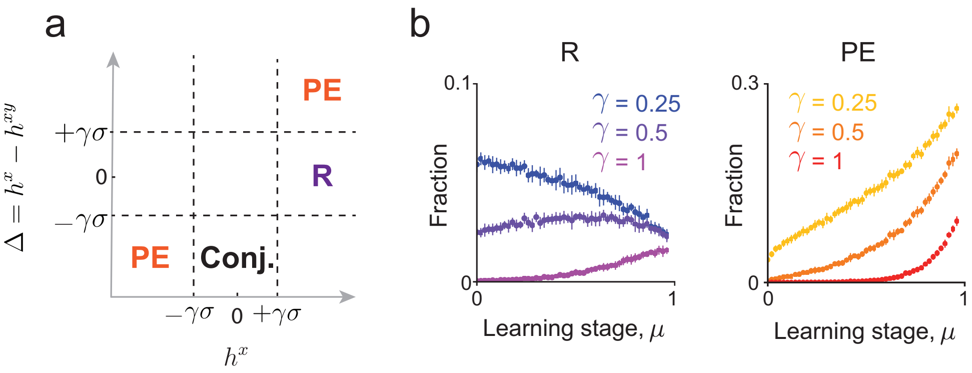


Fig. S3. Abundance of functional cell types as a function of learning stage and classification threshold. (a) Criteria for classifying different functional cell types. The classification is based on setting two thresholds ($\pm\gamma\sigma$) on the voltage response in the x-only mismatch condition (h_i^x), and its difference from the voltage response to match condition ($h_i^x - h_i^{xy}$, see Methods). The regions corresponding to prediction-error (*PE*) and representation (*R*) neurons for stimulus x are shown in the plot. Here we do not distinguish positive or negative *PE* neurons. Similar criteria are applied when replacing x with y. Also shown is the region corresponding to the conjunctive (Conj.) neurons, which have a small response in x-only mismatch condition but a large response in the match condition. (b) Fraction of *R* and *PE* neurons for different threshold values, as a function of the learning stage μ. The fraction of *PE* neurons increases during learning independently of the threshold. Here $\alpha = 0$.

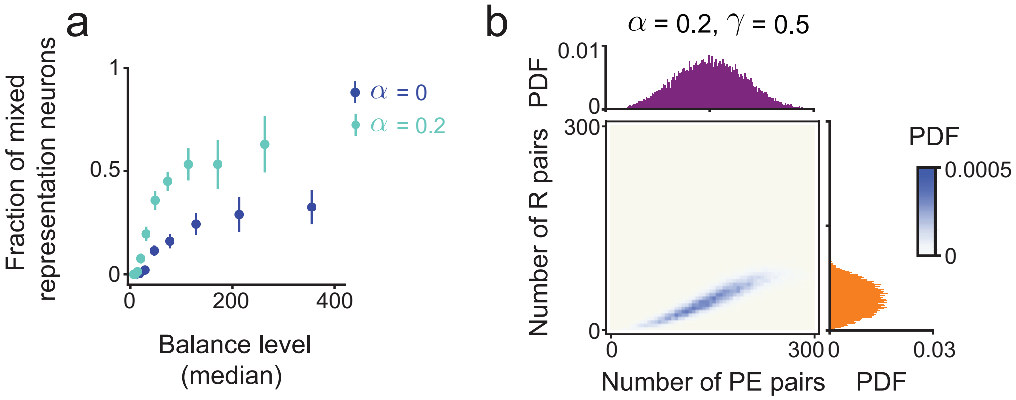


Fig. S4. Fraction of mixed-representation neurons as a function of balance level and stimulus dimensionality α. (a) Here we vary the gain parameter b to generate a range of balance levels (median). As the stimulus dimensionality α increases, the fraction of mixed representation neurons for a fixed balance level also increases. (b) Each neuron in the network is a representation neuron for a certain number of stimulus-pairs ('Number of R pairs') and a prediction-error neuron for other stimulus-pairs ('Number of PE pairs'). Plotted is the joint distribution of these two numbers for neurons in a network when it is trained to associate $P = 400$ stimulus-pairs. The corresponding marginal distributions are also shown. The joint distribution has a positive correlation. This indicates that when all P stimulus-pairs are considered, more neurons have a mixed representation than would be expected if the representation of stimulus and prediction-error was independent across pairs.

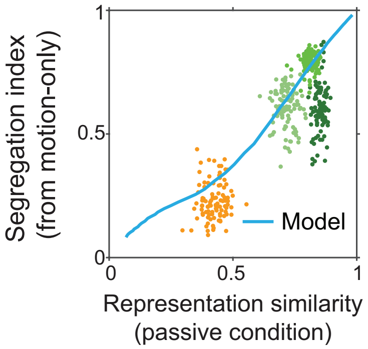


Fig. S5. Segregation index as a function of representation similarity for different pairs of expected and probe sounds. Plotted are the segregation indices as a function of the representation similarity for different probe types (similar to Fig. 4f). Here the segregation indices are computed based on the differences Δ between the motion-only mismatch (passive: movement-only) and match (active: lever press + sound) neural responses. Colored points correspond to subsamples of the data. The results exhibit a similar trend as in Fig. 4f. The model curve shown in this plot is computed using a different sparsity level (by varying the firing threshold θ) compared to the values used in Fig. 4f. Under our main modeling assumptions: connectivity that is symmetric and puts the stimuli x and y on 'equal footing' during learning, synaptic weights with Gaussian statistics, and ReLU nonlinearity, we were not able to find a single value of θ to fit the data with two definitions of mismatch responses. Future work with more realistic network connectivity may give a choice of parameters that is consistent across both ways of comparing neural responses in expected and unexpected stimulus conditions.

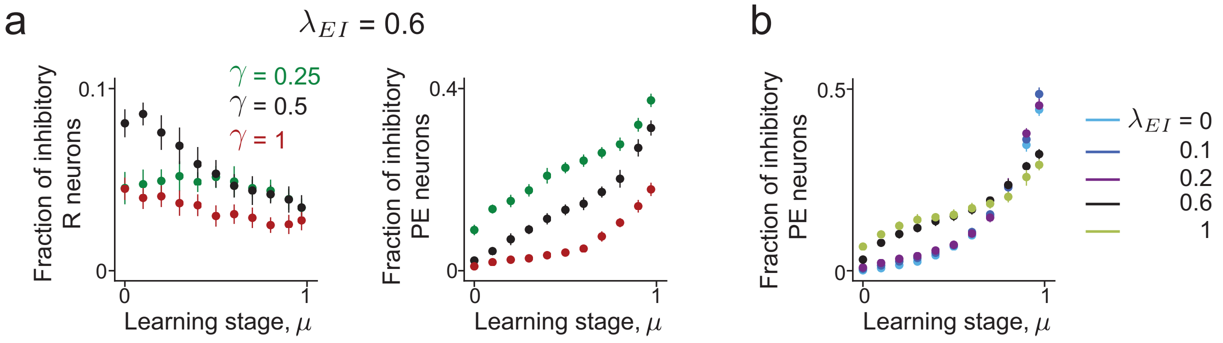
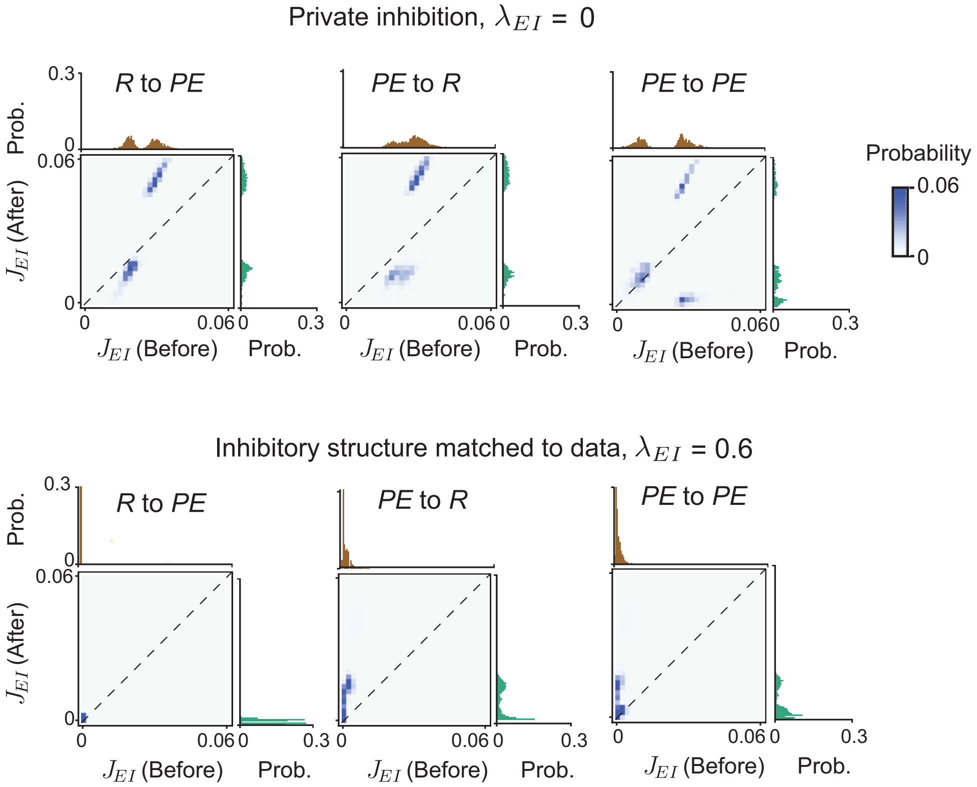


Fig. S6. Abundance of functional cell types among inhibitory neurons. (a) Fraction of inhibitory representation (R) and prediction-error (PE) neurons at different learning stages (different values of μ) when using different voltage thresholds ($\pm\gamma\sigma$). For the connectivity parameter that best matches our data ($\lambda_{EI} = 0.6$), the effect of learning is consistent across different thresholds. (b) Fraction of inhibitory prediction-error neurons at different learning stages for different values of λ. Unlike other network properties that do depend on the architecture of inhibitory connectivity (shown in Fig. 5), this quantity depends weakly on the parameter λ_{EI}. In this plot we set $\alpha = 0$.



862

Fig. S7. Changes to inhibitory to excitatory connections during learning do not depend strongly on the functional cell type of the target. Synaptic weight distribution of I-to-E connections before and after learning, when $\lambda_{EI} = 0$ (top) and $\lambda_{EI} = 0.6$ (bottom), for pairs of E and I neurons belonging to different functional classes: (*R* to *PE*, left; *PE* to *R*, middle; *PE* to *PE*, right). These fine-scale distribution show similar trends as in Fig. 5f,g.

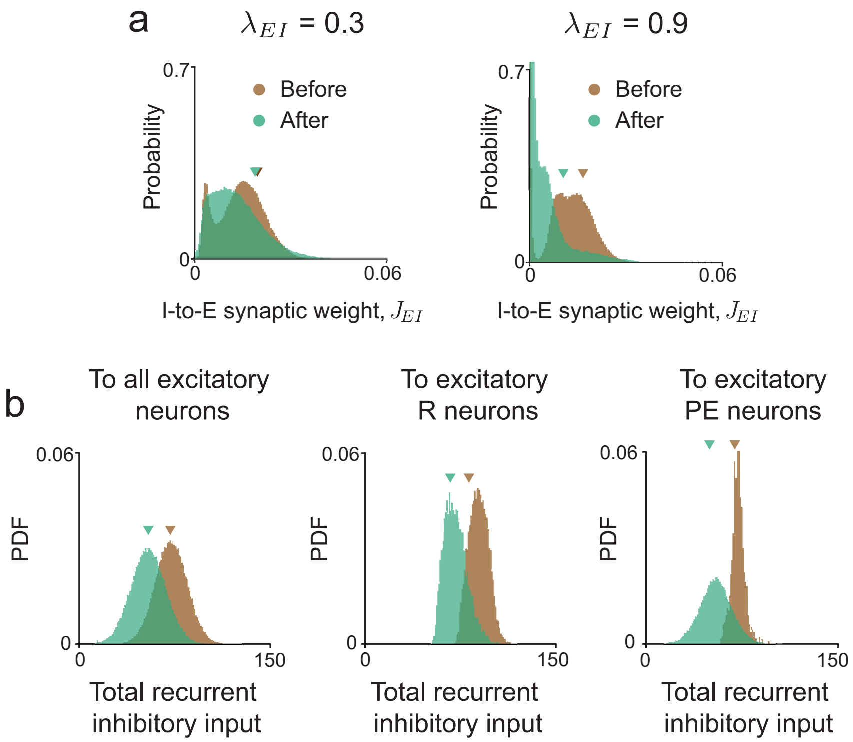


Fig. S8. Learning predictive representations does not rely on overall potentiation of inhibitory connections, across different network architectures. (a) Synaptic weight distribution of all I-to-E connections before and after learning for values of λ_{EI} not shown in Fig. 5. There is no overall increase in the strength of inhibitory synapses after learning, suggesting that across different network architectures, predictive computations that lead to suppressed responses to expected stimuli are distributed. (b) Distribution of the total recurrent inhibitory input received by different populations of excitatory neurons, in the match condition. The overall inhibition received by excitatory neurons in the network decreases after learning.

References

[1] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[2] G. B. Keller and T. D. Mrsic-Flogel, "Predictive processing: a canonical cortical computation," *Neuron*, vol. 100, no. 2, pp. 424–435, 2018.

[3] J. Poort, A. G. Khan, M. Pachitariu, A. Nemri, I. Orsolic, J. Krupic, M. Bauza, M. Sahani, G. B. Keller, T. D. Mrsic-Flogel, *et al.*, "Learning enhances sensory and multiple non-sensory representations in primary visual cortex," *Neuron*, vol. 86, no. 6, pp. 1478–1490, 2015.

[4] F. C. Widmer, S. M. O'Toole, and G. B. Keller, "NMDA receptors in visual cortex are necessary for normal visuomotor integration and skill learning," *Elife*, vol. 11, p. e71476, 2022.

[5] S. J. Eliades and X. Wang, "Neural substrates of vocalization feedback monitoring in primate auditory cortex," *Nature*, vol. 453, no. 7198, pp. 1102–1106, 2008.

[6] G. B. Keller and R. H. Hahnloser, "Neural processing of auditory feedback during vocal practice in a songbird," *Nature*, vol. 457, no. 7226, pp. 187–190, 2009.

[7] K. S. Walsh, D. P. McGovern, A. Clark, and R. G. O'Connell, "Evaluating the neurophysiological evidence for predictive processing as a model of perception," *Annals of the New York Academy of Sciences*, vol. 1464, no. 1, pp. 242–268, 2020.

[8] A. Nelson, D. M. Schneider, J. Takatoh, K. Sakurai, F. Wang, and R. Mooney, "A circuit for motor cortical modulation of auditory cortical activity," *Journal of Neuroscience*, vol. 33, no. 36, pp. 14342–14353, 2013.

[9] D. M. Schneider, A. Nelson, and R. Mooney, "A synaptic and circuit basis for corollary discharge in the auditory cortex," *Nature*, vol. 513, no. 7517, pp. 189–194, 2014.

[10] B. P. Rummell, J. L. Klee, and T. Sigurdsson, "Attenuation of responses to self-generated sounds in auditory cortical neurons," *Journal of Neuroscience*, vol. 36, no. 47, pp. 12010–12026, 2016.

[11] D. M. Schneider, J. Sundararajan, and R. Mooney, "A cortical filter that learns to suppress the acoustic consequences of movement," *Nature*, vol. 561, no. 7723, pp. 391–395, 2018.

[12] N. J. Audette, W. Zhou, A. La Chioma, and D. M. Schneider, "Precise movement-based predictions in the mouse auditory cortex," *Current Biology*, vol. 32, no. 22, pp. 4925–4940, 2022.

[13] N. J. Audette and D. M. Schneider, "Stimulus-specific prediction error neurons in mouse auditory cortex," *Journal of Neuroscience*, vol. 43, no. 43, pp. 7119–7129, 2023.

[14] G. Iurilli, D. Ghezzi, U. Olcese, G. Lassi, C. Nazzaro, R. Tonini, V. Tucci, F. Benfenati, and P. Medini, "Sound-driven synaptic inhibition in primary visual cortex," *Neuron*, vol. 73, no. 4, pp. 814–828, 2012.

[15] L. A. Ibrahim, L. Mesik, X.-y. Ji, Q. Fang, H.-f. Li, Y.-t. Li, B. Zingg, L. I. Zhang, and H. W. Tao, "Cross-modality sharpening of visual cortical processing through layer-1-mediated inhibition and disinhibition," *Neuron*, vol. 89, no. 5, pp. 1031–1045, 2016.

[16] A. R. Garner and G. B. Keller, "A cortical circuit for audio-visual predictions," *Nature Neuroscience*, vol. 25, no. 1, pp. 98–105, 2022.

[17] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston, "Canonical microcircuits for predictive coding," *Neuron*, vol. 76, no. 4, pp. 695–711, 2012.

[18] S. M. O'Toole, H. K. Oyibo, and G. B. Keller, "Molecularly targetable cell types in mouse visual cortex have distinguishable prediction error responses," *Neuron*, 2023.

[19] R. P. Rao, "A sensory–motor theory of the neocortex," *Nature Neuroscience*, pp. 1–15, 2024.

[20] R. Jordan and G. B. Keller, "Opposing influence of top-down and bottom-up input on excitatory layer 2/3 neurons in mouse primary visual cortex," *Neuron*, vol. 108, no. 6, pp. 1194–1206, 2020.

[21] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation in speech production," *Science*, vol. 279, no. 5354, pp. 1213–1216, 1998.

[22] S. J. Blakemore, S. J. Goodbody, and D. M. Wolpert, "Predicting the consequences of our own actions: the role of sensorimotor context estimation," *Journal of Neuroscience*, vol. 18, no. 18, pp. 7511–7518, 1998.

[23] G. Bouvier, Y. Senzai, and M. Scanziani, "Head movements control the activity of primary visual cortex in a luminance-dependent manner," *Neuron*, vol. 108, no. 3, pp. 500–511, 2020.

[24] C. Büchel, S. Geuter, C. Sprenger, and F. Eippert, "Placebo analgesia: a predictive coding perspective," *Neuron*, vol. 81, no. 6, pp. 1223–1239, 2014.

[25] T. Woo, X. Liang, D. A. Evans, O. Fernandez, F. Kretschmer, S. Reiter, and G. Laurent, "The dynamics of pattern matching in camouflaging cuttlefish," *Nature*, pp. 1–7, 2023.

[26] N. Ulanovsky, L. Las, D. Farkas, and I. Nelken, “Multiple time scales of adaptation in auditory cortex neurons,” *Journal of Neuroscience*, vol. 24, no. 46, pp. 10440–10453, 2004.

[27] I. Hershenhoren, N. Taaseh, F. M. Antunes, and I. Nelken, “Intracellular correlates of stimulus-specific adaptation,” *Journal of Neuroscience*, vol. 34, no. 9, pp. 3303–3319, 2014.

[28] A. G. Enikolopov, L. Abbott, and N. B. Sawtell, “Internally generated predictions enhance neural and behavioral detection of sensory stimuli in an electric fish,” *Neuron*, vol. 99, no. 1, pp. 135–146, 2018.

[29] S. Z. Muller, A. N. Zadina, L. Abbott, and N. B. Sawtell, “Continual learning in a multi-layer network of an electric fish,” *Cell*, vol. 179, no. 6, pp. 1382–1392, 2019.

[30] H. Makino and T. Komiyama, “Learning enhances the relative impact of top-down processing in the visual cortex,” *Nature Neuroscience*, vol. 18, no. 8, pp. 1116–1122, 2015.

[31] T. S. Yarden, A. Mizrahi, and I. Nelken, “Context-dependent inhibitory control of stimulus-specific adaptation,” *Journal of Neuroscience*, vol. 42, no. 23, pp. 4629–4651, 2022.

[32] M. Boerlin, C. K. Machens, and S. Denève, “Predictive coding of dynamical variables in balanced spiking networks,” *PLoS Computational Biology*, vol. 9, no. 11, p. e1003258, 2013.

[33] S. Denève and C. K. Machens, “Efficient codes and balanced networks,” *Nature Neuroscience*, vol. 19, no. 3, pp. 375–382, 2016.

[34] J. Kadmon, J. Timcheck, and S. Ganguli, “Predictive coding in balanced neural networks with noise, chaos and delays,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 16677–16688, 2020.

[35] L. Hertäg and H. Sprekeler, “Learning prediction error neurons in a canonical interneuron circuit,” *Elife*, vol. 9, p. e57541, 2020.

[36] L. Hertäg and C. Clopath, “Prediction-error neurons in circuits with multiple neuron types: Formation, refinement, and functional implications,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 13, p. e2115699119, 2022.

[37] F. A. Mikulasch, L. Rudelt, and V. Priesemann, “Local dendritic balance enables learning of efficient representations in networks of spiking neurons,” *Proceedings of the National Academy of Sciences*, vol. 118, no. 50, p. e2021925118, 2021.

[38] F. A. Mikulasch, L. Rudelt, M. Wibral, and V. Priesemann, “Where is the error? hierarchical predictive coding through dendritic error computation,” *Trends in Neurosciences*, vol. 46, no. 1, pp. 45–59, 2023.

[39] Y. Song, B. Millidge, T. Salvatori, T. Lukasiewicz, Z. Xu, and R. Bogacz, “Inferring neural activity before plasticity as a foundation for learning beyond backpropagation,” *Nature Neuroscience*, pp. 1–11, 2024.

[40] R. Hodson, M. Mehta, and R. Smith, “The empirical status of predictive coding and active inference,” *Neuroscience & Biobehavioral Reviews*, p. 105473, 2023.

[41] E. J. Dennis, A. El Hady, A. Michaiel, A. Clemens, D. R. G. Tervo, J. Voigts, and S. R. Datta, “Systems neuroscience of natural behaviors in rodents,” *Journal of Neuroscience*, vol. 41, no. 5, pp. 911–919, 2021.

[42] A. Wallach and N. B. Sawtell, “An internal model for canceling self-generated sensory input in freely behaving electric fish,” *Neuron*, 2023.

[43] T. Keck, T. Toyoizumi, L. Chen, B. Doiron, D. E. Feldman, K. Fox, W. Gerstner, P. G. Haydon, M. Hübener, H.-K. Lee, *et al.*, “Integrating Hebbian and homeostatic plasticity: the current state of the field and future research directions,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1715, p. 20160158, 2017.

[44] G. B. Keller, T. Bonhoeffer, and M. Hübener, “Sensorimotor mismatch signals in primary visual cortex of the behaving mouse,” *Neuron*, vol. 74, no. 5, pp. 809–815, 2012.

[45] H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, and T. D. Mrsic-Flogel, “Functional specificity of local synaptic connections in neocortical networks,” *Nature*, vol. 473, no. 7345, pp. 87–91, 2011.

[46] L. Cossell, M. F. Iacaruso, D. R. Muir, R. Houlton, E. N. Sader, H. Ko, S. B. Hofer, and T. D. Mrsic-Flogel, “Functional organization of excitatory synaptic strength in primary visual cortex,” *Nature*, vol. 518, no. 7539, pp. 399–403, 2015.

[47] S. El-Boustani, J. P. Ip, V. Breton-Provencher, G. W. Knott, H. Okuno, H. Bito, and M. Sur, “Locally coordinated synaptic plasticity of visual cortex neurons in vivo,” *Science*, vol. 360, no. 6395, pp. 1349–1354, 2018.

[48] P. Zmarz and G. B. Keller, “Mismatch receptive fields in mouse visual cortex,” *Neuron*, vol. 92, no. 4, pp. 766–772, 2016.

[49] K. Friston and S. Kiebel, “Predictive coding under the free-energy principle,” *Philosophical transactions of the Royal Society B: Biological sciences*, vol. 364, no. 1521, pp. 1211–1221, 2009.

[50] K. Friston, "The free-energy principle: a unified brain theory?," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[51] Y. Ahmadian and K. D. Miller, "What is the dynamical regime of cerebral cortex?," *Neuron*, vol. 109, no. 21, pp. 3373–3391, 2021.

[52] M. Leinweber, D. R. Ward, J. M. Sobczak, A. Attinger, and G. B. Keller, "A sensorimotor circuit in mouse cortex for visual flow predictions," *Neuron*, vol. 95, no. 6, pp. 1420–1432, 2017.

[53] N. Gillis, *Nonnegative Matrix Factorization*. SIAM, 2020.

[54] T. Haga and T. Fukai, "Extended temporal association memory by modulations of inhibitory circuits," *Physical Review Letters*, vol. 123, no. 7, p. 078101, 2019.

[55] M. V. Srinivasan, S. B. Laughlin, and A. Dubs, "Predictive coding: a fresh view of inhibition in the retina," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 216, no. 1205, pp. 427–459, 1982.

[56] S. Furutachi, A. D. Franklin, T. D. Mrsic-Flogel, and S. B. Hofer, "Cooperative thalamocortical circuit mechanism for sensory prediction errors," *bioRxiv*, pp. 2023–07, 2023.

[57] J. Bolz and C. D. Gilbert, "Generation of end-inhibition in the visual cortex via interlaminar connections," *Nature*, vol. 320, no. 6060, pp. 362–365, 1986.

[58] A. Ayaz, A. Stäuble, M. Hamada, M.-A. Wulf, A. B. Saleem, and F. Helmchen, "Layer-specific integration of locomotion and sensory information in mouse barrel cortex," *Nature communications*, vol. 10, no. 1, p. 2585, 2019.

[59] D. M. Schneider, "Reflections of action in sensory cortex," *Current Opinion in Neurobiology*, vol. 64, pp. 53–59, 2020.

[60] K. K. Clayton, R. S. Williamson, K. E. Hancock, G.-i. Tasaka, A. Mizrahi, T. A. Hackett, and D. B. Polley, "Auditory corticothalamic neurons are recruited by motor preparatory inputs," *Current Biology*, vol. 31, no. 2, pp. 310–321, 2021.

[61] R. J. Douglas and K. A. Martin, "Neuronal circuits of the neocortex," *Annual Reviews of Neuroscience*, vol. 27, no. 1, pp. 419–451, 2004.

[62] K. D. Harris and G. M. Shepherd, "The neocortical circuit: themes and variations," *Nature Neuroscience*, vol. 18, no. 2, pp. 170–181, 2015.

[63] M. W. Spratling, "Predictive coding as a model of biased competition in visual attention," *Vision Research*, vol. 48, no. 12, pp. 1391–1408, 2008.

[64] H. Ko, L. Cossell, C. Baragli, J. Antolik, C. Clopath, S. B. Hofer, and T. D. Mrsic-Flogel, "The emergence of functional microcircuits in visual cortex," *Nature*, vol. 496, no. 7443, pp. 96–100, 2013.

[65] B. Bathellier, L. Ushakova, and S. Rumpel, "Discrete neocortical dynamics predict behavioral categorization of sounds," *Neuron*, vol. 76, no. 2, pp. 435–449, 2012.

[66] O. Barak, "Recurrent neural networks as versatile tools of neuroscience research," *Current Opinion in Neurobiology*, vol. 46, pp. 1–6, 2017.

[67] U. Pereira-Obilinovic, J. Aljadeff, and N. Brunel, "Forgetting leads to chaos in attractor networks," *Physical Review X*, vol. 13, no. 1, p. 011009, 2023.

[68] B. Wang and J. Aljadeff, "Multiplicative shot-noise: A new route to stability of plastic networks," *Physical Review Letters*, vol. 129, no. 6, p. 068101, 2022.

[69] A. Ororbia, A. Mali, C. L. Giles, and D. Kifer, "Lifelong neural predictive coding: Learning cumulatively online without forgetting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5867–5881, 2022.

[70] V. Zhu and R. Rosenbaum, "Evaluating the extent to which homeostatic plasticity learns to compute prediction errors in unstructured neuronal networks," *Journal of Computational Neuroscience*, vol. 50, no. 3, pp. 357–373, 2022.

[71] R. Engelken, A. Ingrosso, R. Khajeh, S. Goedeke, and L. Abbott, "Input correlations impede suppression of chaos and learning in balanced firing-rate networks," *PLoS Computational Biology*, vol. 18, no. 12, p. e1010590, 2022.

[72] J. S. Li, A. A. Sarma, T. J. Sejnowski, and J. C. Doyle, "Internal feedback in the cortical perception–action loop enables fast and accurate behavior," *Proceedings of the National Academy of Sciences*, vol. 120, no. 39, p. e2300445120, 2023.

[73] A. Finkelstein, K. Daie, M. Rózsa, R. Darshan, and K. Svoboda, "Connectivity underlying motor cortex activity during naturalistic goal-directed behavior," *bioRxiv*, pp. 2023–11, 2023.

[74] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, "The importance of mixed selectivity in complex cognitive tasks," *Nature*, vol. 497, no. 7451, pp. 585–590, 2013.

[75] S. Fusi, E. K. Miller, and M. Rigotti, "Why neurons mix: high dimensionality for higher cognition," *Current Opinion in Neurobiology*, vol. 37, pp. 66–74, 2016.

[76] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex," *Nature*, vol. 503, no. 7474, pp. 78–84, 2013.

[77] Y. LeCun, “A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27,” *Open Review*, vol. 62, 2022.

.

Supplementary Information for
"Desegregation of neural predictive processing"

Bin Wang[1], Nicholas J Audette[2], David M Schneider[2], Johnatan Aljadeff[3,*]

[1] Department of Physics, University of California San Diego, La Jolla, CA, 92093, USA

[2] Center for Neural Science, New York University, New York, NY 10003, USA

[3] Department of Neurobiology, University of California San Diego, La Jolla, CA, 92093, USA

(Dated: August 6, 2024)

1. A NORMATIVE FRAMEWORK FOR HIGH-DIMENSIONAL PREDICTIVE PROCESSING

1.1. The recurrent network model

We consider a network of N recurrently connected neurons, where the firing-rates of the neurons are denoted by the vector $\boldsymbol{r}(t) = (\, r_1(t), \ldots, r_N(t)\,)$. The firing-rate of each neuron is related to its voltage level h_i via a nonlinear activation function, $r_i(t) = \phi(h_i(t))$. We denote the learned paired inputs to the network as $\boldsymbol{x}(t) = (\, x^1(t), \ldots, x^P(t)\,)$ and $\boldsymbol{y}(t) = (\, y^1(t), \ldots, y^{P'}(t)\,)$. Notice that the dimensions of the paired inputs are not necessarily the same in this section.

In the predictive coding framework, the network continuously generates an internal prediction of the inputs. We assume that internal predictions (denoted $\hat{x}^k(t)$, $\hat{y}^k(t)$) are linear read-outs from the network activity, i.e.,

$$\begin{aligned} \hat{x}^k(t) &= \frac{1}{N}\boldsymbol{w}^k \cdot \boldsymbol{r}(t), \quad k = 1, \ldots, P, \\ \hat{y}^{k'}(t) &= \frac{1}{N}\boldsymbol{v}^{k'} \cdot \boldsymbol{r}(t), \quad k' = 1, \ldots, P'. \end{aligned} \tag{S1}$$

Here $\boldsymbol{w}^k, \boldsymbol{v}^{k'}$ are the N-dimensional readout weight vectors.

Our aim is to derive a network model where the prediction-errors are minimized subject to some regularization term on encoding efficiency. Mathematically, we define the following objective function,

$$E(t) = \sum_{k=1}^{P} \left(x^k(t) - \hat{x}^k(t)\right)^2 + \sum_{k=1}^{P'} \left(y^k(t) - \hat{y}^k(t)\right)^2 + \frac{2}{bN}\sum_{i=1}^{N} F(r_i(t)). \tag{S2}$$

The first two terms of $E(t)$ correspond to the prediction-errors. The regularization term, and the function $F(z)$ in particular, depend on the nonlinear activation function ϕ. We consider those nonlinear activation functions where the firing-rate is $\phi_+(h-\theta)$ above a threshold θ, and 0 below the threshold. Mathematically,

$$\phi(h) = \begin{cases} \phi_+(h-\theta) & \text{if } h \geq \theta, \\ 0 & \text{if } h < \theta. \end{cases} \tag{S3}$$

Here ϕ_+ is a monotonically increasing smooth function which vanishes at 0, such that ϕ is continuous. This class of functions includes a number of activation functions used in previous work, e.g., rectified linear activation (ReLU, $\phi_+(h) = h$) and rectified *nonlinear* units, $\phi_+(h) = h^p$ ($p > 0$), that coincide with ReLU for $p = 1$.

For this choice of ϕ, we show below that the recurrent network dynamics

$$\tau \frac{\mathrm{d}h_i(t)}{\mathrm{d}t} = -h_i(t) + \sum_{j=1}^{N} J_{ij}\phi(h_j(t)) + \sum_{k=1}^{P} b\, w_i^k x^k + \sum_{k'=1}^{P'} b\, v_i^{k'} y^{k'}, \tag{S4}$$

with the connectivity matrix and choice of regularization,

$$\begin{aligned} J_{ij} &= -\frac{b}{N}\left(\sum_{k=1}^{P} w_i^k w_j^k + \sum_{k'=1}^{P'} v_i^{k'} v_j^{k'}\right), \\ F(r) &= \int_0^r \phi_+^{-1}(z)dz + \theta r = \frac{p}{p+1} r^{1+\frac{1}{p}} + \theta r, \end{aligned} \tag{S5}$$

minimizes the objective [Eq. (S2)]. Note that adding a nonzero firing threshold ($\theta > 0$) in the regularization function enforces sparse neural responses, penalizing large firing-rates. For the ReLU nonlinearity ($p = 1$), we have $F(r) = r^2/2 + \theta r = (r+\theta)^2/2 - \theta^2/2$.

We assume that the timescale of changes to the inputs is much slower than the timescale of changes to neuronal activity, such that we can ignore potential time-dependencies of $\boldsymbol{x}$ and $\boldsymbol{y}$. Under this assumption, the objective [Eq. (S2)] can be written as a function of the neural activity and readout weights,

$$E(\boldsymbol{r}; \{\boldsymbol{w}^k, \boldsymbol{v}^k\}) = \sum_{k=1}^{P}\left[\left(x^k - \frac{1}{N}\boldsymbol{w}^k \cdot \boldsymbol{r}\right)^2 + \left(y^k - \frac{1}{N}\boldsymbol{v}^k \cdot \boldsymbol{r}\right)^2\right] + \frac{2}{bN}\sum_{i=1}^{N} F(r_i). \tag{S6}$$

The neural activity $\boldsymbol{r}(t)$ governed by the dynamical equations [Eq. (S4)] with the connectivity matrix [Eq. (S5)] minimizes the objective function [Eq. (S2)]. This can be shown by directly evaluating the time derivative of $E(t)$:

$$\begin{aligned} \frac{\mathrm{d}E(t)}{\mathrm{d}t} &= \sum_{i=1}^{N} \frac{\partial E}{\partial r_i}\frac{\partial r_i}{\partial h_i}\frac{\mathrm{d}h_i}{\mathrm{d}t} \\ &= -\sum_{i=1}^{N}\left[2\sum_{k=1}^{P}(x^k - \hat{x}^k)\frac{w_i^k}{N} + 2\sum_{k'=1}^{P'}(y^{k'} - \hat{y}^{k'})\frac{v_i^{k'}}{N} - \frac{2}{bN}\phi_+^{-1}(r_i) - \frac{2\theta}{bN}\right]\phi'(h_i)\frac{\mathrm{d}h_i}{\mathrm{d}t} \\ &= -\frac{2}{bN}\sum_{i=1}^{N}\phi'(h_i)\frac{\mathrm{d}h_i}{\mathrm{d}t} \\ &\quad \times \left[\sum_{k=1}^{P} b w_i^k x^k + \sum_{k=1}^{P'} b v_i^{k'} y^{k'} - \frac{b}{N}\sum_{j=1}^{N}\left(\sum_{k=1}^{P} w_i^k w_j^k + \sum_{k'=1}^{P'} v_i^{k'} v_j^{k'}\right)\phi(h_j) - \phi_+^{-1}(r_i) - \theta\right] \\ \circledast &= -\frac{2}{bN}\sum_{i=1}^{N}\phi'(h_i)\frac{\mathrm{d}h_i}{\mathrm{d}t} \end{aligned}$$

$$\times\left[\sum_{k=1}^{P} bw_i^k x^k + \sum_{k=1}^{P'} bv_i^{k'} y^{k'} - \frac{b}{N}\sum_{j=1}^{N}\left(\sum_{k=1}^{P} w_i^k w_j^k + \sum_{k'=1}^{P'} v_i^{k'} v_j^{k'}\right)\phi(h_j) - h_i\right]$$

$$= -\frac{2}{bN\tau}\sum_{i=1}^{N}\left(\frac{\mathrm{d}h_i}{\mathrm{d}t}\right)^2 \phi'(h_i) \qquad \text{(S7)}$$

In the line indicated by ⊛ we used the identity $\phi_+^{-1}(r)\phi'(h) = (h-\theta)\phi'(h)$. Each term in the sum that appears in the last line of Eq. (S7) is positive, so the time derivative of $E(t)$ is negative. The existence of Lyapunov function for Eq. (S4) indicates that the network will reach a (stable) fixed point which satisfies for each neuron i,

$$h_i^\star = \sum_{j=1}^{N} J_{ij}\phi(h_j^\star) + \sum_{k=1}^{P} b\, w_i^k x^k + \sum_{k'=1}^{P'} b\, v_i^{k'} y^{k'}. \qquad \text{(S8)}$$

Moreover, since $E(\boldsymbol{r})$ is a strictly convex function of the firing-rate vector $\boldsymbol{r}$, the optimal fixed-point solution $\boldsymbol{r}^\star$ is unique. From Eq. (S8), $\boldsymbol{h}^\star$ is also unique. Furthermore, that fixed point is a global minimum of E, which can be shown by evaluating the first-order derivatives of Eq. (S2) at the fixed point. Taken together, our results show that the network is guaranteed to reach a stable fixed-point for any input combination (indicated by x^k and y^k), which is the minimum of Eq. (S2).

In the following sections, we will assume that there are P distinct pairs of stimuli indexed by k, (x^k, y^k). The corresponding feedforward weight vectors $\boldsymbol{w}^k, \boldsymbol{v}^k$ are assumed to be random, with mean 0. Associative training induces correlations between each component of the feedforward weights, via, for example, Hebbian-type plasticity. More precisely, for $i, j = 1, \ldots, N$ and $k, k' = 1, \ldots, P$,

$$\langle w_i^k\rangle = \langle v_i^k\rangle = 0, \qquad \langle w_i^k w_j^{k'}\rangle = \langle v_i^k v_j^{k'}\rangle = \delta_{kk'}\delta_{ij}, \qquad \langle w_i^k v_j^{k'}\rangle = \delta_{kk'}\delta_{ij}\mu^k. \qquad \text{(S9)}$$

Here $\langle\cdots\rangle$ denotes the expectation over the probability distribution of synaptic weights. To study how neural representations change during learning we vary μ^k systematically. Note that we have rescaled μ^k by N^{-1} relative to the notation used in the main text.

Our choice of synaptic weight statistics [Eq. (S9)] arises from an optimization procedure that minimizes the objective function [Eq. (S2)]. Indeed, performing gradient descent on E within a short time window Δt induces the following weight changes,

$$\begin{aligned}\Delta w_i^k &= -\eta\tfrac{\partial E(\boldsymbol{r};\{\boldsymbol{w}^k,\boldsymbol{v}^k\})}{\partial w_i^k}\Delta t = \tfrac{\eta}{N}\left(x^k - \tfrac{1}{N}\boldsymbol{w}^k\cdot\boldsymbol{r}\right)\phi(h_i)\Delta t \equiv \frac{\eta}{N}\delta x^k r_i \Delta t,\\ \Delta v_i^k &= -\eta\tfrac{\partial E(\boldsymbol{r};\{\boldsymbol{w}^k,\boldsymbol{v}^k\})}{\partial v_i^k}\Delta t = \tfrac{\eta}{N}\left(y^k - \tfrac{1}{N}\boldsymbol{v}^k\cdot\boldsymbol{r}\right)\phi(h_i)\Delta t \equiv \frac{\eta}{N}\delta y^k r_i \Delta t.\end{aligned} \qquad \text{(S10)}$$

We assume that the learning rate is small $\eta \ll 1$, such that the neural dynamics [Eq. (S4)] remain at the steady state $\boldsymbol{r}^\star$. We will show below (SI §2.1) that during associative learning ($x^k = y^k = 1$), the variables representing prediction errors are non-negative ($\delta x^k, \delta y^k \geq 0$), which implies that the weights could grow unbounded during learning.

To prevent this potential blow-up, we introduce a normalization mechanism that regularizes the weights. After each 'learning-step' [Eq. S10], the weights change according to a 'homeostatic-step',

$$w_i^k(t) \to \frac{w_i^k(t) - m_w^k}{\sigma_w^k}, \qquad m_w^k = \frac{1}{N}\sum_{i=1}^{N} w_i^k(t), \qquad (\sigma_w^k)^2 = \frac{1}{N}\sum_{i=1}^{N}(w_i^k(t) - m_w^k)^2. \tag{S11}$$

Here m_w^k and σ_w^k are the means and the standard deviations of the weight vector $\boldsymbol{w}$ computed over the N neurons. Similar updates are applied to the weights $\boldsymbol{v}$. We show that under these update rules, $\mu^k(t)$, the correlation between $\boldsymbol{w}^k$ and $\boldsymbol{v}^k$ at time t during the learning process, increases monotonically.

We first note that the homeostatic step [Eq. (S11)] ensures that weight vectors have zero mean and unit variance. Upon presentation of the stimulus-pair k, the steady-state input to neuron i is independent of inputs to other neurons. Additionally, in the $N \to \infty$ limit, $\delta x^{k'}$ and $\delta y^{k'}$ are nonzero only if $k' = k$. These properties are shown explicitly using a replica calculation below (SI §2.1). It is therefore sufficient to verify that applying the learning-step [Eq. (S10)] does not lead to a decrease in the correlation. This can be done by a direct calculation of the correlation in Eq. (S11). Notice that in the $N \to \infty$ limit,

$$\begin{aligned}
m_w^k &\to \left\langle \Delta w_i^k \right\rangle, \\
(\sigma_w^k)^{-1} &\to \left\langle (w_i^k(t) + \Delta w_i^k - m_w^k)^2 \right\rangle^{-1/2} \\
&\to (1 + 2\left\langle w_i^k \Delta w_i^k \right\rangle + O(\Delta t^2))^{-1/2} \\
&= 1 - \left\langle w_i^k \Delta w_i^k \right\rangle + O(\Delta t^2) \\
&= 1 - \eta \hat{x}^k \delta x^k \Delta t + O(\Delta t^2).
\end{aligned} \tag{S12}$$

Therefore the weight w_i^k after the learning and homeostatic steps is,

$$\begin{aligned}
w_i^k(t + \Delta t) &= \frac{w_i^k(t) + \Delta w_i^k - m_w^k}{\sigma_w^k} \\
&= (w_i^k(t) + \Delta w_i^k - \left\langle \Delta w_i^k \right\rangle)(1 - \eta \hat{x}^k \delta x^k \Delta t) + O(\Delta t^2) \\
&= w_i^k(t) + \Delta w_i^k - \left\langle \Delta w_i^k \right\rangle - \eta w_i^k(t) \hat{x}^k \delta x^k \Delta t + O(\Delta t^2).
\end{aligned} \tag{S13}$$

Using this approximation and a similar expression for $v_i^k(t+\Delta t)$, the correlation is now,

$$\begin{aligned}\mu^k(t+\Delta t) &= \left\langle w_i^k(t+\Delta t)v_i^k(t+\Delta t)\right\rangle \\ &= \mu^k(t) + \left\langle w_i^k \Delta v_i^k(t) + v_i^k \Delta w_i^k(t)\right\rangle - \eta\mu^k(t)(\hat{x}^k\delta x^k + \hat{y}^k\delta y^k)\Delta t + O(\Delta t^2) \\ &= \mu^k(t) + \eta[\hat{y}^k(\delta x^k - \delta y^k\mu^k(t)) + \hat{x}^k(\delta y^k - \delta x^k\mu^k(t))]\Delta t + O(\Delta t^2). \qquad \text{(S14)}\end{aligned}$$

In the match condition ($x^k = y^k = 1$) we have from symmetry that $\hat{x}^k = \hat{y}^k$ and $\delta x^k = \delta y^k$. We will show in SI §2.1 using a replica calculation that $\hat{x}^k, \delta x^k \geq 0$, which together imply that the bracket is positive when $\mu^k(t) \leq 1$. Thus the correlation between the weight vectors increases during associative learning. This justifies our choice of weight statistics [Eq. (S9)] as a description for the network during associative learning.

1.2. A Bayesian inference perspective of the network model

The predictive coding framework is often used to account for inference of latent causes of sensorimotor inputs to the brain, based on prediction and prediction-error signals [1, 17, 49, 55]. In this section we show that our model can similarly be viewed as a network performing Bayesian inference. Specifically, the network's neural dynamics [Eq. (S4)] implement the inference (or state estimation) of latent variables driving inputs. Moreover, the slow synaptic weight changes during learning [Eq. (S9)] can be viewed as a mechanism for improving the accuracy of the inference performed by the network.

We consider a scenario where sensory inputs in the environment are generated by a probabilistic generative model, $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{r})$, where $\boldsymbol{x}, \boldsymbol{y}$ are the (possibly time-dependent) sensory inputs and $\boldsymbol{r}$ represents the latent variables that determine the statistics of the sensory inputs. We denote the prior distribution over the latent variables as $p_0(\boldsymbol{r})$. Then given the sensory inputs $\boldsymbol{x}, \boldsymbol{y}$, the latent variables $\boldsymbol{r}$ can be inferred by maximizing the posterior distribution via Bayes' rule,

$$p(\boldsymbol{r}|\boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{r})p_0(\boldsymbol{r})}{p(\boldsymbol{x}, \boldsymbol{y})}, \qquad \text{(S15)}$$

where $p(\boldsymbol{x}, \boldsymbol{y}) = \int p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{r})p_0(\boldsymbol{r})d\boldsymbol{r}$ is the marginal distribution of the sensory inputs, independent of the latent variables.

Suppose that the generative distribution is a multivariate Gaussian and that its mean is a linear readout of the latent variables,

$$\ln p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{r}) = -\frac{1}{\sigma_1^2}\sum_{k=1}^{P}\left[\left(x^k - \frac{1}{N}\boldsymbol{w}^k\cdot\boldsymbol{r}\right)^2 + \left(y^k - \frac{1}{N}\boldsymbol{v}^k\cdot\boldsymbol{r}\right)^2\right] + \text{const.} \tag{S16}$$

Further suppose that the prior distribution has the form,

$$\ln p_0(\boldsymbol{r}) = -\frac{1}{\sigma_0^2 N}\sum_{i=1}^{N} F(r_i) + \text{const.} \tag{S17}$$

Then, recalling Eq. (S7), we see that the neural dynamics [Eq. (S4)] maximize the log posterior distribution,

$$\ln p(\boldsymbol{r}|\boldsymbol{x},\boldsymbol{y}) = \ln p(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{r}) + \ln p_0(\boldsymbol{r}) + \text{const} = -\frac{1}{\sigma_1^2}E(\boldsymbol{r}) + \text{const.} \tag{S18}$$

Here $E(\boldsymbol{r})$ is the objective function in the previous section with $b = \sigma_0^2/\sigma_1^2$. Thus, our model's gain parameter b is related to the prediction accuracy σ_1. The latent variables $\boldsymbol{r}$ here correspond to the firing rates of the neurons in the network.

In the more general case where sensory inputs are not generated exactly according to Eq. (S16), prediction accuracy can be improved by adjusting the readout weights $\boldsymbol{w}^k, \boldsymbol{v}^k$ to maximize the log posterior distribution [Eq. (S18)] based on the learning rule [Eq. (S11)]. This weight optimization procedure is equivalent to using a variational approach for maximizing the Bayesian model evidence, as introduced in previous predictive coding literature [17, 49, 78]. We also note that the nonlinear response function ϕ appears in the regularization $F(\boldsymbol{r})$ [Eq. (S5)] is linked to the 'encoding' of prior information on the latent variables, $p_0(\boldsymbol{r})$.

1.3. Extensions of the network model

1.3.1. Associations between more than two modalities

Our network model can be generalized to apply to scenarios in which the animal is trained to associate multiple ($M \geq 3$) sensorimotor inputs. Here the network generates internal predictions for each input, that can be linearly read-out,

$$\hat{x}_l^k(t) = \frac{1}{N}\boldsymbol{w}_l^k\cdot\boldsymbol{r}(t), \qquad k = 1,\ldots,P, \qquad l = 1,\ldots,M, \tag{S19}$$

where $\boldsymbol{w}_l^k$ are the readout weights for each input in each stimulus modality. The objective function [Eq. (S1)] is now,

$$E_M(t) = \sum_{l=1}^{M} \sum_{k=1}^{P} \left(x_l^k(t) - \hat{x}_l^k(t)\right)^2 + \frac{2}{b} \sum_{i=1}^{N} F(r_i(t)). \tag{S20}$$

The network dynamics and recurrent connectivity matrix are,

$$\begin{aligned} \tau \frac{\mathrm{d}h_i(t)}{\mathrm{d}t} &= -h_i(t) + \sum_{j=1}^{N} J_{ij}^M \phi(h_j(t)) + \sum_{l=1}^{M} \sum_{k=1}^{P} b\, w_{l,i}^k x_l^k, \\ J_{ij}^M &= \frac{b}{N} \sum_{l=1}^{M} \sum_{k=1}^{P} w_{l,i}^k w_{l,j}^k. \end{aligned} \tag{S21}$$

Using similar derivations as above, one can show that (i) $E_M(t)$ is a Lyapunov function for the network dynamics, and (ii) the network will reach a unique stable fixed point for any combination of the inputs x_l^k. Assuming that the feedforward weights corresponding to associated stimuli become increasingly correlated during learning (similarly to the $M = 2$ case), will make this model useful for studying predictive representations when training animals on more complex stimulus combinations.

1.3.2. Neurons with dendritic compartments

The network model with point neurons [Eq. (S4)] and the associated learning rules [Eq. (S11)] can be extended to a model with dendritic compartments. Crucially, this extension allows the learning rule to be realized by local plasticity rules.

Following the approach introduced in Refs. [37, 38], we first notice that Eqs. (S4-S5) can be rewritten by decomposing the connectivity to synaptic weights onto specific dendrites, giving,

$$\begin{aligned} J_{ij}^k &= -\frac{b}{N} w_i^k w_j^k, \qquad J_{ij}^{k+P} = -\frac{b}{N} v_i^k v_j^k, \\ \tau \frac{\mathrm{d}h_i(t)}{\mathrm{d}t} &= -h_i(t) + \sum_{k=1}^{P} \left[\sum_{j=1}^{N} J_{ij}^k \phi(h_j(t)) + w_i^k x^k \right] + \sum_{k=1}^{P'} \left[\sum_{j=1}^{N} J_{ij}^{k+P} \phi(h_j(t)) + v_i^k y^k \right]. \end{aligned} \tag{S22}$$

Here we think of $h_i(t)$ as the *somatic* membrane potential of neuron i. Next we introduce $P + P'$ dendritic compartments corresponding to neuron i. The voltages u_i^k for $k = 1, \ldots, P$

and for $k = P+1, \ldots, P' + P$ are respectively governed by the equations,

$$\begin{aligned}\tau_u \frac{\mathrm{d}u_i^k(t)}{\mathrm{d}t} &= -u_i^k(t) + \sum_{j=1}^{N} J_{ij}^k \phi(h_j(t)) + bw_i^k x^k,\\ \tau_u \frac{\mathrm{d}u_i^{k+P}(t)}{\mathrm{d}t} &= -u_i^{k+P}(t) + \sum_{j=1}^{N} J_{ij}^{k+P} \phi(h_j(t)) + bv_i^k y^k.\end{aligned} \tag{S23}$$

The somatic voltage level is then driven by the dendrites,

$$\tau \frac{\mathrm{d}h_i(t)}{\mathrm{d}t} = -h_i(t) + \sum_{k=1}^{P} u_i^k(t) + \sum_{k=1}^{P'} u_i^{k+P}(t). \tag{S24}$$

Under the assumption that dendrite voltage changes faster than somatic voltage, $\tau_u \ll \tau$, this recovers our original model with point neurons [Eq. (S4)].

The learning rule of the dendrite-specific feedforward weights is given by,

$$\Delta w_i^k = \frac{\eta}{N} \frac{x^k u_i^k r_i}{I_i^k}, \qquad \Delta v_i^k = \frac{\eta}{N} \frac{y^k u_i^{k+P} r_i}{I_i^{k+P}}, \tag{S25}$$

where we have denoted $I_i^k = bw_i^k x^k$ and $I_i^{k+P} = bv_i^k y^k$. Note that the quantities on the right hand side are 'local' to the feedforward synapses w_i^k and v_i^k. At steady-state, $u_i^k = bw_i^k \delta x^k$ and $u_i^{k+P} = bv_i^k \delta y^k$. Together with the definition of I_i^k, this learning rule is the same as Eq. (S10). To avoid unbounded growth of the weights in this setting, we assume a similar homeostatic mechanism which recovers the previous learning rule for the feedforward weights [Eq. (S11)].

The recurrent weights are subject to the learning rules,

$$\begin{aligned}\frac{\mathrm{d}J_{ij}^k}{\mathrm{d}t} &= -\frac{\eta}{N} u_i^k (r_j - \langle r_j \rangle) - \left[\frac{\eta_1^k}{I_i^k}(r_i - \langle r_i \rangle) + \eta_2^k\right] J_{ij}^k,\\ \frac{\mathrm{d}J_{ij}^{k+P}}{\mathrm{d}t} &= -\frac{\eta}{N} u_i^{k+P} (r_j - \langle r_j \rangle) - \left[\frac{\eta_1^{k+P}}{I_i^{k+P}}(r_i - \langle r_i \rangle) + \eta_2^{k+P}\right] J_{ij}^{k+P},\end{aligned} \tag{S26}$$

where $\eta_1^k = \langle u_i^k \rangle$, $\eta_2^k = \langle u_i^k \rangle \hat{x}^k$ and $\eta_2^{k+P} = \langle u_i^{k+P} \rangle \hat{y}^k$ are activity-dependent learning rates. The dendrite-specific synaptic weights [Eq. (S22)] are solutions to these learning dynamics.

We note that the increase in correlation between $\boldsymbol{w}$ and $\boldsymbol{v}$ during learning is reflected in this plasticity rule by the dependence of both J^k and J^{k+P} on the firing rates $\boldsymbol{r}$. Since those rates depend on inputs from both modalities, both sets of dendrite-specific synaptic weights change based on the interplay between the multimodal input.

1.3.3. Hierarchical network architecture

In the recurrent network model studied thus far, a single module integrates inputs from multiple sensorimotor modalities. Here we generalize this model to a network consisting of multiple (L) modules arranged in a layered structure. Each module has N neurons with firing rates denoted as $\boldsymbol{r}^l$, $l = 1, \ldots, L$. We assume that the paired stimulus inputs enter the network via the first and the last module respectively (Fig. 6a). For convenience, we denote the inputs as $\boldsymbol{x} \equiv \boldsymbol{r}^0$ and $\boldsymbol{y} \equiv \boldsymbol{r}^{L+1}$.

Each module generates predictions of the activity of 'adjacent' (earlier and later) modules, i.e., neurons in module l generate predictions for neural responses in modules $l-1$ and $l+1$. Those predictions are assumed to be linear readouts of the firing rates,

$$\hat{\boldsymbol{r}}^{l-1} = W^{l\top}\boldsymbol{r}^l, \qquad \hat{\boldsymbol{r}}^{l+1} = V^{l\top}\boldsymbol{r}^l. \tag{S27}$$

Here W^l, V^l are the readout matrices. The objective function for this hierarchical network is a sum of the objective function applied to each module-separately with the corresponding prediction errors and firing-rate regularization,

$$\begin{aligned}
\mathsf{E}(\{\boldsymbol{r}^l\}; \{W^l, V^l\}) &= \sum_{l=1}^{L} \left[\frac{1}{\sigma_l^2}(\boldsymbol{r}^{l-1} - \hat{\boldsymbol{r}}^{l-1})^2 + \frac{1}{\sigma_l^2}(\boldsymbol{r}^{l+1} - \hat{\boldsymbol{r}}^{l+1})^2 + \frac{F(\boldsymbol{r}^l)}{b_l} \right], \\
&= \sum_{l=1}^{L} \left[\frac{1}{\sigma_l^2}(\boldsymbol{r}^{l-1} - W^{l\top}\boldsymbol{r}^l)^2 + \frac{1}{\sigma_l^2}(\boldsymbol{r}^{l+1} - V^{l\top}\boldsymbol{r}^l)^2 + \frac{F(\boldsymbol{r}^l)}{b_l} \right], \\
&= \sum_{l=1}^{L} E(\boldsymbol{r}^l; W^l, V^l).
\end{aligned} \tag{S28}$$

Here σ_l measures the module-specific precision of predictions and b_l is the module-specific regularization. The assumption that the neurons in module l minimize the module-specific loss $E(\boldsymbol{r}^l; W^l, V^l)$ implies that the neural dynamics within each module and the recurrent synaptic weights have identical form to those in the single-module case,

$$\begin{aligned}
\sigma_l^2 \frac{\mathrm{d}h_i^l}{\mathrm{d}t} &= -\sigma_l^2 h_i^l(t) + \sum_{j=1}^{N} J_{ij}^l \phi(h_j^l(t)) + b_l \sum_{k=1}^{N} W_{ik}^l \phi(h_k^{l-1}(t)) + b_l \sum_{k'=1}^{N} V_{ik}^l \phi(h_k^{l+1}(t)), \\
J_{ij}^l &= -\frac{b_l}{N} \sum_{k=1}^{N} \left(W_{ik}^l W_{jk}^l + V_{ik}^l V_{jk}^l \right).
\end{aligned} \tag{S29}$$

Similarly to the network with a single module, we assume that associative learning induces correlations between the corresponding weight vectors for each stimulus-pair. In the hierarchical network, the feedforward weight matrices in the first and last modules $W^{1\top}, V^1$ have dimensions $N \times P$ rather than the $N \times N$ dimensions of matrices in intermediate modules. We assume that the intermediate feedforward weight matrices have rank P (the stimulus dimension). Furthermore, because the process of training the network to associate stimuli x^k with y^k is symmetric under the substitutions $x \leftrightarrow y$, $W \leftrightarrow V$, we assume that W, V are symmetric matrices. With these assumptions, we the weight matrices are,

$$\begin{aligned} W^1 &= \sum_{k=1}^{P} \hat{\boldsymbol{e}}_k (\boldsymbol{w}_k^1)^\top, \qquad W^l = \frac{1}{N}\sum_{k=1}^{P} \boldsymbol{w}_k^l (\boldsymbol{w}_k^l)^\top, \qquad l = 2, \dots, L, \\ V^L &= \sum_{k=1}^{P} \hat{\boldsymbol{e}}_k (\boldsymbol{v}_k^L)^\top, \qquad V^l = \frac{1}{N}\sum_{k=1}^{P} \boldsymbol{v}_k^l (\boldsymbol{v}_k^l)^\top, \qquad l = 1, \dots, L-1. \end{aligned} \tag{S30}$$

During associative learning, these weight vectors become correlated and their statistics are,

$$\begin{aligned} \langle w_{ki}^l \rangle = \langle v_{ki}^l \rangle = 0, \qquad \langle w_{ki}^l w_{k'j}^l \rangle = \langle v_{ki}^l v_{k'j}^l \rangle = \delta_{kk'}\delta_{ij}, \qquad \langle w_{ki}^l v_{k'j}^l \rangle = \delta_{kk'}\delta_{ij}\mu^k, \\ \langle w_{ki}^l w_{k'j}^{l'} \rangle = \langle v_{ki}^l v_{k'j}^{l'} \rangle = \langle w_{ki}^l v_{k'j}^{l'} \rangle = \delta_{kk'}\delta_{ij}\mu^k, \qquad l' \neq l. \end{aligned} \tag{S31}$$

Here the first line specifies the weight statistics within module l, and the second line specifies the statistics across modules. The recurrent connectivity within each module simplifies to a form which is identical to that of the single module network,

$$\begin{aligned} J^l &= -\frac{b_l}{N}\sum_{k,k'=1}^{P} \left[\boldsymbol{w}_{k'}^l \frac{\boldsymbol{w}_{k'}^{l\top}\boldsymbol{w}_k^l}{N} (\boldsymbol{w}_k^l)^\top + \boldsymbol{v}_{k'}^l \frac{\boldsymbol{v}_{k'}^{l\top}\boldsymbol{v}_k^l}{N} (\boldsymbol{v}_k^l)^\top \right] \\ &\overset{N\to\infty}{=} -\frac{b_l}{N}\sum_{k=1}^{P} \left[\boldsymbol{w}_k^l (\boldsymbol{w}_k^l)^\top + \boldsymbol{v}_k^l (\boldsymbol{v}_k^l)^\top \right]. \end{aligned} \tag{S32}$$

2. PREDICTIVE REPRESENTATIONS IN RECURRENT NETWORKS

When the stimulus inputs do not depend on time, the objective function E [Eq. (S2)] can be viewed as a function of the firing-rates and synaptic weights,

$$\begin{aligned} E(\boldsymbol{r}; \{\boldsymbol{w}^k, \boldsymbol{v}^k\}) &= \sum_{k=1}^{P} \left[\left(x^k - \frac{1}{N}\boldsymbol{w}^k \cdot \boldsymbol{r} \right)^2 + \left(y^k - \frac{1}{N}\boldsymbol{v}^k \cdot \boldsymbol{r} \right)^2 \right] + \frac{2}{bN}\sum_{i=1}^{N} F(r_i) \\ &= \frac{2}{bN}\left[\sum_{k=1}^{P} b \left(-x^k \boldsymbol{w}^k \cdot \boldsymbol{r} - y^k \boldsymbol{v}^k \cdot \boldsymbol{r} + \frac{(\boldsymbol{w}^k \cdot \boldsymbol{r})^2 + (\boldsymbol{v}^k \cdot \boldsymbol{r})^2}{2N} \right) + \sum_{i=1}^{N} F(r_i) \right] \\ &\quad + \sum_{i=1}^{P} \left[(x^k)^2 + (y^k)^2 \right]. \end{aligned}$$

$$\equiv \frac{2}{bN} E_0(\boldsymbol{r}; \{\boldsymbol{w}^k, \boldsymbol{v}^k\}) + \sum_{i=1}^{P} \left[(x^k)^2 + (y^k)^2\right]. \tag{S33}$$

The steady state firing-rates can be expressed as minimization over E_0, since the second term in Eq. (S33) does not depend on $\boldsymbol{r}$,

$$\boldsymbol{r}^\star = \underset{\boldsymbol{r} \in \mathbb{R}_+^n}{\operatorname{argmin}}\, E_0(\boldsymbol{r}; \{\boldsymbol{w}^k, \boldsymbol{v}^k\}). \tag{S34}$$

Next we will use the replica method [79, 80] to calculate the firing-rate distribution of neurons in the network,

$$p(r) = \frac{1}{N} \sum_{i=1}^{N} \delta(r - r_i). \tag{S35}$$

In general, firing-rates in the network depend on the specific realization of random weights $\boldsymbol{w}^k, \boldsymbol{v}^{k'}$. We find however that in the $N \to \infty$ limit, the firing-rate distribution is self-averaging and depends only on the distribution of synaptic weights. By choosing which of the x^k and y^k's are nonzero, we can study the network response in different stimulus conditions. For convenience, we assume that at any given time, only a finite number of stimulus-pairs are presented, or equivalently, there are only $K = O(1)$ pairs (x^k, y^k) for $k = 1, \ldots, K$, where at least one stimulus is nonzero. We set the decay timescale to $\tau = 1$.

2.1. Replica calculation of the firing-rate statistics

We consider the partition function

$$Z = \int_{\mathbb{R}_+^N} e^{-\beta E_0(\boldsymbol{r}; \{\boldsymbol{w}^k, \boldsymbol{v}^k\})}\, \mathrm{d}\boldsymbol{r}. \tag{S36}$$

We suppress the domain of integration over firing-rates for readability in the following calculations. In the limit $\beta \to \infty$, the dominant contribution to Z comes from the fixed point solution which minimizes $E_0(\boldsymbol{r}; \{\boldsymbol{w}^k, \boldsymbol{v}^k\})$ in Eq. (S34). The logarithm of the partition function concentrates around its expectation, so we use the replica trick,

$$\lim_{N \to \infty} \frac{\ln Z}{N} = \lim_{N \to \infty} \left\langle \frac{\ln Z}{N} \right\rangle = \lim_{n \to 0} \lim_{N \to \infty} \frac{\ln \langle Z^n \rangle}{nN}. \tag{S37}$$

We make the standard assumption that the order of the limits can be exchanged in the last equality. We first calculate $\langle Z^n \rangle$. For readability, we use g for the gain parameter (instead of b) in Subsection 2.1, and $a, b = 1, \ldots, n$ for the replica indices. Without loss of generality, we assume that the presented stimuli (i.e., indices k such that x_k or y_k is nonzero) are the first K pairs, $k = 1, \dots, K$.

$$\begin{aligned}
\langle Z^n \rangle = &\int \prod_a \mathrm{d}\boldsymbol{r}^a \left\langle \exp\left\{ -\beta \sum_{i,a} F(r_i^\alpha) - \frac{g\beta}{2N} \sum_a \sum_{k=1}^{P} \left[(\boldsymbol{w}^k \cdot \boldsymbol{r}^a)^2 + (\boldsymbol{v}^k \cdot \boldsymbol{r}^a)^2 \right] \right\} \right. \\
&\quad \left. \times \exp\left[g\beta \sum_a \sum_{s=1}^{K} (x^s \boldsymbol{w}^s \cdot \boldsymbol{r}^a + y^s \boldsymbol{v}^s \cdot \boldsymbol{r}^a) \right] \right\rangle , \\
= &\int \prod_{a,i} \mathrm{d}r_i^a \left\langle \exp\left\{ -\beta \sum_{i,a} F(r_i^a) - \frac{g\beta}{2N} \sum_a \sum_{k=K+1}^{P} \left[(\boldsymbol{w}^k \cdot \boldsymbol{r}^a)^2 + (\boldsymbol{v}^k \cdot \boldsymbol{r}^a)^2 \right] \right\} \right\rangle \\
&\quad \times \left\langle \exp\left\{ g\beta \sum_a \sum_{k=1}^{K} \left[\left(x^k \boldsymbol{w}^k \cdot \boldsymbol{r}^a + y^k \boldsymbol{v}^k \cdot \boldsymbol{r}^a \right) - \frac{1}{2N} \left((\boldsymbol{w}^k \cdot \boldsymbol{r}^a)^2 + (\boldsymbol{v}^k \cdot \boldsymbol{r}^a)^2 \right) \right] \right\} \right\rangle \\
= &\int \prod_{a,i} \mathrm{d}r_i^a \exp\left[-\beta \sum_{a,i} F(r_i^a) \right] \left\langle \exp\left\{ -\frac{g\beta}{2N} \sum_a \sum_{k=K+1}^{P} \left[(\boldsymbol{w}^k \cdot \boldsymbol{r}^a)^2 + (\boldsymbol{v}^k \cdot \boldsymbol{r}^a)^2 \right] \right\} \right\rangle \\
&\quad \times \left\langle \exp\left\{ g\beta \sum_a \sum_{k=1}^{K} \left[-\frac{1}{2N} (\boldsymbol{w}^k \cdot \boldsymbol{r}^a)^2 - \frac{1}{2N} (\boldsymbol{v}^k \cdot \boldsymbol{r}^a)^2 + x^k \boldsymbol{w}^k \cdot \boldsymbol{r}^a + y^k \boldsymbol{v}^k \cdot \boldsymbol{r}^a \right] \right\} \right\rangle .
\end{aligned} \tag{S38}$$

Notice that we have split the summation over all P stimulus-pairs and averaging over the corresponding synaptic weights into the presented pairs ($k = 1, \dots, K$) and the rest ($k = K + 1, \dots, P$). We first perform calculations for the $P - K$ 'absent' stimulus-pairs. Using the integral representation of Gaussian function, we get,

$$\begin{aligned}
e^{-\frac{g\beta}{2N}(\boldsymbol{w}^k \cdot \boldsymbol{r}^a)^2} &= \int \frac{\mathrm{d}t^{k,a}}{\sqrt{2\pi}} \sqrt{g\beta} e^{-g\beta\left[\frac{(t^{k,a})^2}{2} + it^{k,a} \frac{\boldsymbol{w}^k \cdot \boldsymbol{r}}{\sqrt{N}} \right]}, \\
e^{-\frac{g\beta}{2N}(\boldsymbol{v}^k \cdot \boldsymbol{r}^a)^2} &= \int \frac{\mathrm{d}s^{k,\alpha}}{\sqrt{2\pi}} \sqrt{g\beta} e^{-g\beta\left[\frac{(s^{k,a})^2}{2} + is^{k,a} \frac{\boldsymbol{v}^k \cdot \boldsymbol{r}}{\sqrt{N}} \right]}.
\end{aligned} \tag{S39}$$

Using these, the term corresponding to the $P - K$ absent stimulus-pairs becomes,

$$\begin{aligned}
&\left\langle \exp\left\{ -\frac{g\beta}{2N} \sum_a \sum_{k=K+1}^{P} \left[(\boldsymbol{w}^k \cdot \boldsymbol{r}^a)^2 + (\boldsymbol{v}^k \cdot \boldsymbol{r}^a)^2 \right] \right\} \right\rangle \\
&= \left\langle \prod_a \prod_{k=K+1}^{P} \frac{g\beta}{2\pi} \int \mathrm{d}t^{k,a} \mathrm{d}s^{k,a} e^{-g\beta\left[\frac{(t^{k,a})^2 + (s^{k,a})^2}{2} + \frac{i}{\sqrt{N}} (t^{k,a} \boldsymbol{w}^k \cdot \boldsymbol{r}^a + s^{k,a} \boldsymbol{v}^k \cdot \boldsymbol{r}^a) \right]} \right\rangle \\
&= \prod_{k=K+1}^{P} \left(\frac{g\beta}{2\pi} \right)^n \int \prod_a \mathrm{d}t^{k,a} \mathrm{d}s^{k,a} e^{-\frac{g\beta}{2} \sum_a \left[(t^{k,a})^2 + (s^{k,a})^2 \right]} \left\langle e^{-\frac{ig\beta}{\sqrt{N}} \sum_a (t^{k,a} \boldsymbol{w}^k \cdot \boldsymbol{r}^a + s^{k,a} \boldsymbol{v}^k \cdot \boldsymbol{r}^a)} \right\rangle \\
&= \prod_{k=K+1}^{P} \left[\left(\frac{g\beta}{2\pi} \right)^n \int \prod_\alpha \mathrm{d}t^a \mathrm{d}s^a e^{-\frac{g\beta}{2} \sum_a \left[(t^a)^2 + (s^a)^2 \right]} \left\langle e^{-\frac{ig\beta}{\sqrt{N}} \left[(\sum_a t^a \boldsymbol{r}^a) \cdot \boldsymbol{w} + (\sum_a s^a \boldsymbol{r}) \cdot \boldsymbol{v} \right]} \right\rangle \right].
\end{aligned} \tag{S40}$$

In the last line we have suppressed the superscript k. Recall that for each k, angle brackets denote the average over a pair of synaptic weight vectors, each of which has components sampled from the same distribution with mean 0 and correlation μ^k [Eq. (S9)]. We work out the last factor of the integrand,

$$\begin{aligned}\left\langle e^{-\frac{ig\beta}{\sqrt{N}}\left[(\sum_a t^a \boldsymbol{r}^a)\cdot\boldsymbol{w}+(\sum_a s^a \boldsymbol{r}^a)\cdot\boldsymbol{v}\right]}\right\rangle &= \left\langle \prod_{j=1}^{N} e^{-\frac{ig\beta}{\sqrt{N}}\left[(\sum_a t^a r_j^a)w_j+(\sum_a s^a r_j^a)v_j\right]}\right\rangle \\ &= \prod_{j=1}^{N} f\left(-\frac{g\beta}{\sqrt{N}}\sum_a t^a r_j^a\,,\,-\frac{g\beta}{\sqrt{N}}\sum_a s^a r_j^a\right), \qquad \text{(S41)}\end{aligned}$$

where $f(x,y)$ is the joint characteristic function of the random vectors $\boldsymbol{w}^k, \boldsymbol{v}^k$ with correlation μ^k. The Taylor expansion of $f(\cdot,\cdot)$ in the limit $N\to\infty$ is,

$$\begin{aligned}&f\left(-\frac{g\beta}{\sqrt{N}}\sum_a t^a r_j^a\,,\,-\frac{g\beta}{\sqrt{N}}\sum_a s^a r_j^a\right)\\ &= 1-\frac{g^2\beta^2}{2N}\left[\left(\sum_a t^a r_j^a\right)^2+2\mu^k\left(\sum_a t^a r_j^a\right)\left(\sum_a s^a r_j^a\right)+\left(\sum_a s^a r_j^a\right)^2\right]+O(N^{-2}).\end{aligned} \qquad \text{(S42)}$$

Using this we get,

$$\begin{aligned}&\left\langle e^{-\frac{ig\beta}{\sqrt{N}}\left[(\sum_a t^a \boldsymbol{r}^a)\cdot\boldsymbol{w}+(\sum_a s^a \boldsymbol{r}^a)\cdot\boldsymbol{v}\right]}\right\rangle\\ &\underset{N\to\infty}{\longrightarrow} \prod_{j=1}^{N}\left[1-\frac{g\beta}{2N}\left[\left(\sum_a t^a r_j^a\right)^2+2\mu^k\left(\sum_a t^a r_j^a\right)\left(\sum_a s^a r_j^a\right)+\left(\sum_a s^a r_j^a\right)^2\right]\right]\\ &\underset{e^x\approx 1+x}{\longrightarrow} \exp\left[-\frac{g\beta}{2}\sum_{a,b}\left(\frac{\sum_{j=1}^{N} r_j^a r_j^b}{N}\right)\left(t^a t^b+2\mu^k t^a s^b+s^a s^b\right)\right]\\ &= \exp\left[-\frac{g\beta}{2}\sum_{a,b} q^{ab}\left(t^a t^b+2\mu^k t^a s^b+s^a s^b\right)\right]. \qquad \text{(S43)}\end{aligned}$$

In the last line we have introduced the usual definition of the order parameter,

$$q^{ab}=\frac{1}{N}\sum_{j=1}^{N} r_j^a r_j^b. \qquad \text{(S44)}$$

Collecting terms, we find that Eq. (S40) becomes,

$$\begin{aligned}
&\left\langle \exp\left\{-\frac{g\beta}{2N}\sum_a \sum_{k=K+1}^{P}\left[(\boldsymbol{w}^k\cdot\boldsymbol{r}^a)^2+(\boldsymbol{v}^k\cdot\boldsymbol{r}^a)^2\right]\right\}\right\rangle \\
&=\prod_{k=K+1}^{P}\left\{\left(\frac{g\beta}{2\pi}\right)^n\int\prod_a \mathrm{d}t^a\mathrm{d}s^a\right. \\
&\qquad\left.\times\exp\left[-\frac{g\beta}{2}\left(\sum_a\left[(t^a)^2+(s^a)^2\right]+g\beta\sum_{a,b}q^{ab}\left(t^at^b+2\mu^k t^a s^b+s^as^b\right)\right)\right]\right\} \\
&=\prod_{k=K+1}^{P}\left\{\left(\frac{1}{2\pi}\right)^n\int \mathrm{d}\boldsymbol{t}\mathrm{d}\boldsymbol{s}\exp\left[-\frac{1}{2}\begin{pmatrix}\boldsymbol{t}\\ \boldsymbol{s}\end{pmatrix}^\top\begin{pmatrix}I_n+g\beta q & \mu^k g\beta q\\ \mu^k g\beta q & I_n+g\beta q\end{pmatrix}\begin{pmatrix}\boldsymbol{t}\\ \boldsymbol{s}\end{pmatrix}\right]\right\} \\
&=\sqrt{\prod_{k=K+1}^{P}\det\begin{pmatrix}I_n+g\beta q & \mu^k g\beta q\\ \mu^k g\beta q & I_n+g\beta q\end{pmatrix}^{-1}},
\end{aligned}\tag{S45}$$

Here q is an $n\times n$ matrix [Eq. (S44)] and I_n is the $n\times n$ identity matrix. In the next to last step of Eq. (S45) we rescaled the integration variables t, s by $\sqrt{g\beta}$.

The term in Eq. (S38) corresponding to the K presented pairs can be calculated in a similar fashion, which yields,

$$\begin{aligned}
&\left\langle \exp\left\{g\beta\sum_a\sum_{k=1}^{K}\left[-\frac{1}{2N}(\boldsymbol{w}^k\cdot\boldsymbol{r}^a)^2-\frac{1}{2N}(\boldsymbol{v}^k\cdot\boldsymbol{r}^a)^2+x^k\boldsymbol{w}^k\cdot\boldsymbol{r}^a+y^k\boldsymbol{v}^k\cdot\boldsymbol{r}^a\right]\right\}\right\rangle \\
&=\int(g\beta)^{nK}\prod_a\prod_{k=1}^{K}\frac{\mathrm{d}t^{k,a}\mathrm{d}s^{k,a}}{2\pi}e^{+\frac{g\beta N}{2}\sum_{k,a}\left[(t^{k,a})^2+(s^{k,a})^2\right]}\left\langle e^{g\beta\sum_{k,a,i}\left[(x^kt^{k,\alpha})w_i^kr_i^a+(y^k-s^{k,a})v_i^kr_i^a\right]}\right\rangle.
\end{aligned}\tag{S46}$$

We introduce the delta function to enforce the definition of the order parameter q,

$$\delta\left(q^{ab}-\frac{1}{N}\sum_{j=1}^{N}r_j^ar_j^b\right)=N\int\frac{\mathrm{d}\hat{q}^{ab}}{2\pi}e^{q^{ab}(N\hat{q}^{ab}-\sum_j r_j^ar_j^b)}.\tag{S47}$$

Putting all terms together Eq. (S38) gives,

$$\begin{aligned}
\langle Z^n\rangle=&(g\beta)^{nP}N^{\frac{n^2}{2}}\int\prod_{k,a}\frac{\mathrm{d}t^{k,a}\mathrm{d}s^{k,a}}{2\pi}\prod_{a,b}\frac{\mathrm{d}\hat{q}^{ab}\mathrm{d}q^{ab}}{2\pi}e^{N\sum_{a,b}\hat{q}^{ab}\hat{q}^{ab}-\frac{g\beta N}{2}\sum_{k,a}[(t^{k,a})^2+(s^{k,a})^2]} \\
&\times\left\{\int\prod_a\mathrm{d}r^a\left\langle e^{-\beta\sum_a F(r^a)+g\beta\sum_{k,a}\left[(x^k-it^{k,a})w^kr^a+(y^k-is^{k,a})v^kr^a\right]-\sum_{a,b}\hat{q}^{ab}r^ar^b}\right\rangle\right\}^N \\
&\times\prod_{k=K+1}^{P}\sqrt{\det\begin{pmatrix}I_n+g\beta q & \mu^k g\beta q\\ \mu^k g\beta q & I_n+g\beta q\end{pmatrix}^{-1}}
\end{aligned}$$

$$= \int \prod_{k,\alpha} \frac{\mathrm{d}t^{k,a}\mathrm{d}s^{k,a}}{2\pi} \prod_{a,b} \frac{\mathrm{d}\hat{q}^{ab}\mathrm{d}q^{ab}}{2\pi} e^{N\mathcal{F}(q^{ab},\hat{q}^{ab},t^{k,a},s^{k,a})}. \tag{S48}$$

In the last line we have defined $\mathcal{F}(q^{ab},\hat{q}^{ab},t^{k,a},s^{k,a})$ as,

$$\begin{aligned}
&\mathcal{F}(q^{ab},\hat{q}^{ab},t^{k,a},s^{k,a}) = \\
&\frac{nK}{N}\ln(g\beta) + \frac{n^2}{2}\frac{\ln N}{N} + \sum_{a,b} q^{ab}\hat{q}^{ab} - \frac{g\beta}{2}\sum_{a,k}[(t^{a,k})^2 + (s^{a,k})^2] \\
&+ \ln\left\{\int \prod_a \mathrm{d}r^a \left\langle e^{-\beta\sum_a F(r^a) + g\beta\sum_{k,a}\left[(x^k - it^{k,a})w^k r^a + (y^k - is^{k,a})v^k r^a\right] - \sum_{a,b}\hat{q}^{ab}r^a r^b} \right\rangle\right\} \\
&- \frac{1}{2N}\sum_{k=K+1}^{P} \ln\det\begin{pmatrix} I_n + g\beta q & \mu^k g\beta q \\ \mu^k g\beta q & I_n + g\beta q \end{pmatrix}.
\end{aligned} \tag{S49}$$

186 In the limit $N \to \infty$, we use the saddle point approximation to compute the integral in the 187 last line of Eq. (S48). Furthermore, because the Lyapunov function E_0 is convex [Eq. (S33)], 188 the saddle point solution is replica symmetric, i.e.,

$$q^{ab} = q_0\delta_{ab} + q_1(1-\delta_{ab}), \qquad \hat{q}^{ab} = \hat{q}_0\delta_{ab} + \hat{q}_1(1-\delta_{ab}), \qquad t^{a,k} = t^k, \qquad s^{a,k} = s^k. \tag{S50}$$

We then simplify the terms in $\mathcal{F}$,

$$\begin{aligned}
\sum_{a,b} q^{ab}\hat{q}^{ab} &= nq_0\hat{q}_0 + n(n-1)q_1\hat{q}_1, \\
\sum_{a,k}[(t^{a,k})^2 + (s^{a,k})^2] &= n\sum_k[(t^k)^2 + (s^k)^2].
\end{aligned} \tag{S51}$$

$$\begin{aligned}
\ln\det\begin{pmatrix} I_n + g\beta q & \mu^k g\beta q \\ \mu^k g\beta q & I_n + g\beta q \end{pmatrix} &= \ln\det[I_n + g\beta(1-\mu^k)q] + \ln\det[I_n + g\beta(1+\mu^k)q] \\
&= n\ln\left[(1 + g\beta(1-\mu^k)(q_0 - q_1))(1 + g\beta(1+\mu^k)(q_0 - q_1))\right] \\
&\quad + \ln\left[\left(1 + \frac{g\beta(1-\mu^k)nq_1}{1 + g\beta(1-\mu^k)(q_0 - q_1)}\right)\left(1 + \frac{g\beta(1+\mu^k)nq_1}{1 + g\beta(1+\mu^k)(q_0 - q_1)}\right)\right].
\end{aligned}$$

Simplifying the term in the third line of Eq. (S49) requires a number of additional steps. Using the integral representation of Gaussian function we write,

$$\begin{aligned}
e^{-\sum_{a,b}\hat{q}^{ab}r^a r^b} &= e^{-(\hat{q}_0 - \hat{q}_1)\sum_a (r^a)^2 - \hat{q}_1(\sum_a r^a)^2} \\
&= e^{-(\hat{q}_0 - \hat{q}_1)\sum_a (r^a)^2} \int \frac{\mathrm{d}z}{\sqrt{2\pi}} e^{-\frac{z^2}{2} + i\sum_a \sqrt{2\hat{q}_1} r^a z}.
\end{aligned} \tag{S52}$$

Substituting this into the integral in Eq. (S49) gives,

$$\begin{aligned}
&\int \prod_a \mathrm{d}r^a \left\langle e^{-\beta \sum_a F(r^a)+g\beta \sum_{k,a}\left[(x^k-it^{k,a})w^k r^a+(y^k-is^{k,a})v^k r^a\right]-\sum_{a,b}\hat{q}^{ab}r^a r^b} \right\rangle \\
&= \int \frac{\mathrm{d}z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \prod_a \left[\int \mathrm{d}\nu_\beta(r^a) e^{-(\hat{q}_0-\hat{q}_1)(r^a)^2+i\sqrt{2\hat{q}}r^a z} \left\langle e^{g\beta(x^k-it^{k,a})w^k r^a+(y^k-is^{k,a})v^k r^a} \right\rangle \right] \\
&= \int Dz \left[\int \mathrm{d}\nu_\beta(r) e^{-(\hat{q}_0-\hat{q}_1)r^2+i\sqrt{2\hat{q}}rz} \left\langle e^{g\beta(x^k-it^k)w^k r+(y^k-is^k)v^k r} \right\rangle \right]^n . \qquad \text{(S53)}
\end{aligned}$$

Here we have introduced the notation,

$$Dz = \frac{\mathrm{d}z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \qquad \mathrm{d}\nu_\beta(r) = \mathrm{d}r\, e^{-\beta F(r)}. \qquad \text{(S54)}$$

Therefore, under the replica symmetric ansatz, Eq. (S49) becomes

$$\begin{aligned}
&\mathcal{F}(q_0, q_1, \hat{q}_0, \hat{q}_1, t^k, s^k) = \\
&\quad \frac{n^2}{2}\frac{\ln N}{N} + nq_0\hat{q}_0 + n(n-1)q_1\hat{q}_1 - \frac{ng\beta}{2}\sum_k [(t^k)^2+(s^k)^2] \\
&\quad + \ln \int Dz \left[\int \mathrm{d}\nu_\beta(r) e^{-(\hat{q}_0-\hat{q}_1)r^2+i\sqrt{2\hat{q}_1}rz} \left\langle \prod_k e^{b\beta(x^k-it^k)w^k r+(y^k-is^k)v^k r} \right\rangle \right]^n \\
&\quad - \frac{1}{2N} \sum_{k=K+1}^{P} \Bigg\{ n \ln \left[(1+g\beta(1-\mu^k)(q_0-q_1))(1+g\beta(1+\mu^k)(q_0-q_1)) \right] \\
&\qquad + \ln \left[\left(1+\frac{g\beta(1-\mu^k)nq_1}{1+g\beta(1-\mu^k)(q_0-q_1)}\right)\left(1+\frac{g\beta(1+\mu^k)nq_1}{1+g\beta(1+\mu^k)(q_0-q_1)}\right) \right] \Bigg\}. \qquad \text{(S55)}
\end{aligned}$$

Now we take the limits $P, N \to \infty$ and $n \to 0$, and identify $\alpha = P/N$, which gives,

$$\begin{aligned}
\lim_{\substack{N\to\infty \\ n\to 0}} \frac{\ln\langle Z^n \rangle}{nN} &= \lim_{\substack{N\to\infty \\ n\to 0}} \frac{\mathcal{F}(q_0, q_1, \hat{q}_0, \hat{q}_1, t^k, s^k)}{n} \\
&= q_0\hat{q}_0 - q_1\hat{q}_1 - g\beta \sum_k \frac{(t^k)^2+(s^k)^2}{2} \\
&\quad + \int Dz \ln \int \mathrm{d}\nu_\beta(r) e^{-(\hat{q}_0-\hat{q}_1)r^2+i\sqrt{2\hat{q}_1}rz} \left\langle \prod_k e^{g\beta(x^k-it^k)w^k r+(y^k-is^k)v^k r} \right\rangle \\
&\quad - \frac{\alpha}{2} \Bigg\langle \ln \left[(1+g\beta(1-\mu^k)(q_0-q_1))(1+g\beta(1+\mu^k)(q_0-q_1)) \right] \\
&\qquad + \frac{g\beta(1-\mu^k)q_1}{1+g\beta(1-\mu^k)(q_0-q_1)} + \frac{g\beta(1+\mu^k)q_1}{1+g\beta(1+\mu^k)(q_0-q_1)} \Bigg\rangle_\mu . \qquad \text{(S56)}
\end{aligned}$$

In the third line of Eq. (S56) we used the fact that for a well behaved function $A(z)$,

$$\lim_{n\to 0} \frac{1}{n} \ln \int Dz A^n(z) = \int Dz \ln A(z). \qquad \text{(S57)}$$

The last term in Eq. (S56) (proportional to $\alpha/2$) was obtained by taking the limit over $P, N \to \infty$ and introducing α, and assuming that the number of presented stimulus-pairs K is finite (necessary for the neural activity to remain finite, justified below). The average $\langle \cdots \rangle_\mu$ is over the distribution of correlation values μ^k, $k = 1, \ldots, P$ (i.e., over all learned stimulus-pairs).

To simplify the calculations, we define new variables

$$q' = g\beta(q_0 - q_1), \qquad \hat{q}' = 2\frac{\hat{q}_0 - \hat{q}_1}{\alpha g\beta}, \qquad \hat{q} = -\frac{2\hat{q}_1}{\alpha g^2\beta^2}, \qquad q = \frac{q_1}{g\beta}. \tag{S58}$$

and further make the change of variables, $t^k \to it^k$ and $s^k \to is^k$. With these simplifications, we rewrite Eq. (S56) as,

$$\begin{aligned}
&\lim_{\substack{N\to\infty \\ n\to 0}} \frac{\mathcal{F}(q', q, \hat{q}', \hat{q}, t^k, s^k)}{n} = \\
&\quad \frac{\alpha g\beta}{2}\left(\hat{q}'q - \hat{q}q'\right) + \alpha\frac{\hat{q}'q'}{2} + g\beta\sum_k \frac{(t^k)^2 + (s^k)^2}{2} \\
&\quad + \int Dz \ln\left[\int \mathrm{d}\nu_\beta(r) e^{-g\beta\frac{\alpha\hat{q}'r^2}{2} + g\beta\sqrt{\alpha\hat{q}}rz}\left\langle e^{g\beta r\sum_k\left[(x^k - t^k)w^k + (y^k - s^k)v^k\right]}\right\rangle\right] \\
&\quad - \frac{\alpha}{2}\left\langle \ln(1 + (1-\mu)q')(1 + (1+\mu)q') + \frac{(1-\mu)g\beta q}{1 + (1-\mu)q'} + \frac{(1+\mu)g\beta q}{1 + (1+\mu)q'}\right\rangle_\mu.
\end{aligned} \tag{S59}$$

To extract information about the network's response properties as $N \to \infty$, we evaluated these expressions at the saddle point of $\mathcal{F}(q_0, q_1, \hat{q}_0, \hat{q}_1, t^k, s^k)$. The saddle point satisfies,

$$\begin{aligned}
0 &= \frac{\partial\mathcal{F}}{\partial t^k} = g\beta\left(t^k - \int \mathrm{d}Q_\beta\, w^k r\right), \\
0 &= \frac{\partial\mathcal{F}}{\partial s^k} = g\beta\left(s^k - \int \mathrm{d}Q_\beta v^k r\right), \\
0 &= \frac{\partial\mathcal{F}}{\partial q} = \frac{\alpha g\beta}{2}\left[\hat{q}' - \left\langle \frac{1-\mu}{1 + (1-\mu)q'} + \frac{1+\mu}{1 + (1+\mu)q'}\right\rangle_\mu\right], \\
0 &= \frac{\partial\mathcal{F}}{\partial q'} = \frac{\alpha g\beta}{2}\left[\frac{\hat{q}'}{g\beta} - \hat{q} + \left\langle \frac{(1-\mu)^2}{[1 + (1-\mu)q']^2}q + \frac{(1+\mu)^2}{[1 + (1+\mu)q']^2}q\right\rangle_\mu\right], \\
0 &= \frac{\partial\mathcal{F}}{\partial\hat{q}'} = \frac{\alpha g\beta}{2}\left(\frac{q'}{g\beta} + q - \int \mathrm{d}Q_\beta r^2\right), \\
0 &= \frac{\partial\mathcal{F}}{\partial\hat{q}} = \frac{\alpha g\beta}{2}\left(-q' + \frac{1}{\sqrt{\alpha\hat{q}}}\int \mathrm{d}Q_\beta\, rz\right).
\end{aligned} \tag{S60}$$

Here we have defined the probability measure $\mathrm{d}Q_\beta$ as,

$$\left\langle \int \mathrm{d}Q_\beta(\ldots)\right\rangle = \int Dz \frac{\int \mathrm{d}\nu_\beta(r) e^{-g\beta\frac{\alpha\hat{q}'r^2}{2} + g\beta\sqrt{\alpha\hat{q}}rz}\left\langle e^{g\beta r\sum_k\left[(x^k - t^k)w^k + (y^k - s^k)v^k\right]}(\ldots)\right\rangle}{\int \mathrm{d}\nu_\beta(r) e^{-g\beta\frac{\alpha\hat{q}'r^2}{2} + g\beta\sqrt{\alpha\hat{q}}rz}\left\langle e^{g\beta r\sum_k\left[(x^k - t^k)w^k + (y^k - s^k)v^k\right]}\right\rangle}. \tag{S61}$$

Indeed, the probability measure $\mathrm{d}Q_\beta$ contains a Boltzmann distribution with the corresponding Hamiltonian,

$$\mathcal{H}(r) = F(r) + \frac{g\alpha\hat{q}' r^2}{2} - g\sqrt{\alpha\hat{q}}zr - gr\sum_{k=1}^{K}\left[(x^k - t^k)w^k + (y^k - s^k)v^k\right]. \tag{S62}$$

In the limit $\beta \to \infty$ ('zero temperature'), the Boltzmann distribution is dominated by the minimum of $\mathcal{H}(r)$ (i.e., the 'ground-state'). Since $\mathcal{H}(r)$ is strictly convex, there is a unique minimum $r^\star \geq 0$.

If $r^\star > 0$, the ground state satisfies $\mathcal{H}'(r^\star) = 0$, or equivalently,

$$0 = \phi_+^{-1}(r^\star) + \theta + g\alpha\hat{q}' r^\star - g\sqrt{\alpha\hat{q}}z - g\sum_{k=1}^{K}\left[(x^k - t^k)w^k + (y^k - s^k)v^k\right]. \tag{S63}$$

Otherwise, $r^\star = 0$. Indeed, the two cases can be written in a compact way,

$$r^\star = \phi\left(-g\alpha\hat{q}' r^\star + g\sqrt{\alpha\hat{q}}z + g\sum_{k=1}^{K}\left[(x^k - t^k)w^k + (y^k - s^k)v^k)\right]\right). \tag{S64}$$

The solution of above equation defines a function $r^\star(w^k, v^k, z)$. We recognize the argument of ϕ as the total input to each neuron and $r^\star$ as its nonlinear firing-rate response. It is important to note that the solution $r^\star$ depends on the Gaussian integration variable (z), the random synaptic weights (w, v), and the variables indicating the stimuli being presented (x, y), so overall the saddle point equations are expected to give a *distribution* of firing-rates, not a single value.

Substituting the ground-state solution $r^\star$ into the saddle point equation, we get at $\beta \to \infty$,

$$\begin{aligned}
t^k &= \left\langle w^k r^\star(w^k, v^k, z)\right\rangle_{w^k, v^k, z},\\
s^k &= \left\langle v^k r^\star(w^k, v^k, z)\right\rangle_{w^k, v^k, z},\\
\hat{q}' &= \left\langle \frac{1-\mu}{1+(1-\mu)q'}\right\rangle_\mu + \left\langle \frac{1+\mu}{1+(1+\mu)q'}\right\rangle_\mu,\\
\hat{q} &= \left\langle \left[\frac{1-\mu}{1+(1-\mu)q'}\right]^2\right\rangle_\mu q + \left\langle \left[\frac{1+\mu}{1+(1+\mu)q'}\right]^2\right\rangle_\mu q,\\
q &= \left\langle r^\star(w^k, v^k, z)^2\right\rangle_{w^k, v^k, z},\\
q' &= \frac{1}{\sqrt{\alpha\hat{q}_1}}\left\langle r^\star(w^k, v^k, z)z\right\rangle_{w^k, v^k, z}.
\end{aligned} \tag{S65}$$

Notice that the order parameters t^k and s^k coincide with the internal predictions $\hat{x}^k$ and $\hat{y}^k$ [Eq. (S1)], and the order parameter q is the second moment of the firing-rate distribution.

2.2. Single-neuron and population statistics

We summarize the main results obtained from the above calculations: Given the distribution of synaptic weights $\{w^k, v^k\}$ and a standard normal random variable z, the firing-rate distribution $p(r)$ is the same as the distribution of the ground-state firing-rate $r^\star(w^k, v^k, z)$ [Eq. (S34)]. The order parameters $q, q', \hat{q}, \hat{q}', t^k = \hat{x}^k, s^k = \hat{y}^k$ which appear in $r^\star(w^k, v^k, z)$ need to be solved from the saddle point equations [Eq. (S60)]. Moreover, the voltage distribution of the neurons in the network is simply the distribution of the argument of the firing-rate transfer function ϕ in Eq. (S64), i.e.,

$$h^\star(w^k, v^k, z) \equiv -b\alpha\hat{q}' r^\star(w^k, v^k, z) + b\sqrt{\alpha\hat{q}}z + b\sum_{k=1}^{K}\left[(x^k - t^k)w^k + (y^k - s^k)v^k)\right]. \qquad \text{(S66)}$$

Below we restrict our analysis to the special case where $\{w^k, v^k\}$ follow a multivariate Gaussian distribution; all the stimulus-pairs are learned equally well $\mu^k = \mu$; and the activation function is ReLU, $\phi = [x - \theta]_+$.

2.2.1. *The high-dimensional case, $P/N \to \alpha > 0$*

Under the above assumptions, Eq. (S34) can be solved exactly, giving neurons' firing-rate and voltage distributions,

$$\begin{aligned} r^\star(w^k, v^k, z) &= \frac{b}{1+\alpha b\hat{q}'}\left[\sqrt{\alpha\hat{q}}z + \sum_{k=1}^{K}\left[(x^k - t^k)w^k + (y^k - s^k)v^k)\right] - \frac{\theta}{b}\right]_+, \\ &\equiv \frac{b}{1+\alpha\hat{q}'b}\left[I - \frac{\theta}{b}\right]_+. \\ h^\star(w^k, v^k, z) &= I - \frac{\alpha\hat{q}'b}{1+\alpha\hat{q}'b}\left[I - \frac{\theta}{b}\right]_+. \end{aligned} \qquad \text{(S67)}$$

For convenience, we denote the Gaussian variable $I = \sqrt{\alpha\hat{q}}z + \sum_{k=1}^{K}[(x^k - t^k)w^k + (y^k - s^k)v^k)]$. Each neuron receives input with mean 0, and variance (denoted σ^2) that depends on the stimuli presented – how many, and whether they are matched or mismatched. From the above equation we see that neurons' firing-rates follow a truncated Gaussian distribution.

The saddle point equations [Eq. (S65)] can be simplified into,

$$\begin{aligned}
\hat{q}' &= \frac{1-\mu}{1+(1-\mu)q'} + \frac{1+\mu}{1+(1+\mu)q'}, \\
\hat{q} &= \left[\left(\frac{1-\mu}{1+(1-\mu)q'}\right)^2 + \left(\frac{1+\mu}{1+(1+\mu)q'}\right)^2\right] q, \\
q' &= \frac{bH\left(\frac{\theta}{b\sigma}\right)}{1+\alpha b\hat{q}'}, \\
q &= \frac{(q')^2}{H\left(\frac{\theta}{b\sigma}\right)}\left[\sigma^2 + \left(\frac{\theta}{b}\right)^2 - \frac{\sigma\theta}{\sqrt{2\pi}b}\frac{e^{-\frac{\theta^2}{2b^2\sigma^2}}}{H\left(\frac{\theta}{b\sigma}\right)}\right], \\
\delta x^k &= x^k - \hat{x}^k = \frac{(1+q')x^k - \mu q' y^k}{1+2q'+(1-\mu^2)(q')^2}, \\
\delta y^k &= y^k - \hat{y}^k = \frac{-\mu q' x^k + (1+q')y^k}{1+2q'+(1-\mu^2)(q')^2},
\end{aligned} \tag{S68}$$

where $H(x) = \int_x^\infty Dz$ is related to the complementary error function. Since $q' \geq 0$ and $x^k = y^k = 1$ in the match condition, $\delta x^k = \delta y^k \geq 0$. The variance σ^2 in the above equations is given by,

$$\begin{aligned}
\sigma^2 &= \alpha\hat{q} + \sum_{k=1}^{K}\left[(\delta x^k)^2 + (\delta y^k)^2 + 2\mu\delta x^k \delta y^k\right] \\
&= \frac{2[(1-\mu^2)(1+q')^2+\mu^2]S + 2\mu[1-(1-\mu^2)(q')^2]T}{[1+2q'+(1-\mu^2)(q')^2]^2} \\
&+ \frac{2\alpha(q')^2}{H\left(\frac{\theta}{b\sigma}\right)}\left[\sigma^2 + \left(\frac{\theta}{b}\right)^2 - \frac{\sigma\theta}{\sqrt{2\pi}b}\frac{e^{-\frac{\theta^2}{2b^2\sigma^2}}}{H\left(\frac{\theta}{b\sigma}\right)}\right]\frac{1+\mu^2+2(1-\mu^2)q'+(1-\mu^2)+(1-\mu^2)(q')^2}{[1+2q'+(1-\mu^2)(q')^2]^2}.
\end{aligned} \tag{S69}$$

Here, we define variables that quantify the number of stimuli presented and whether their presentation is matched or mismatched: $S = \frac{1}{2}\sum_{k=1}^{K}[(x^k)^2 + (y^k)^2]$ and $T = \sum_{k=1}^{K} x^k y^k$. Combining Eq. (S69) with the first and third lines of Eq. (S68) gives a solution for σ, q', $\hat{q}'$. By substituting these into the other saddle point equations, we get all the order parameters.

In this case, the mean and variance of the firing-rate distribution are,

$$\begin{aligned}
\langle r^\star \rangle &= \frac{1}{1+\alpha b\hat{q}'}\left[\frac{b\sigma}{\sqrt{2\pi}} - \theta H\left(\frac{\theta}{b\sigma}\right)\right], \\
\mathrm{Var}(r^\star) &= \frac{1}{(1+\alpha b\hat{q}')^2}\left[b^2\sigma^2\left(H\left(\frac{\theta}{b\sigma}\right) - \frac{1}{2\pi}e^{-\frac{\theta^2}{b^2\sigma^2}}\right) - \frac{\theta b\sigma}{\sqrt{2\pi}}\left(1-2H\left(\frac{\theta}{b\sigma}\right)\right)\right. \\
&\qquad \left. + \theta^2 H\left(\frac{\theta}{b\sigma}\right)\left(1-H\left(\frac{\theta}{b\sigma}\right)\right)\right].
\end{aligned} \tag{S70}$$

2.2.2. The case $\alpha \to 0$

When $\alpha \to 0$, the saddle point equations reduce to,

$$\begin{aligned} q' &= bH\left(\frac{\theta}{b\sigma}\right), \\ \sigma^2 &= \frac{2[(1-\mu^2)(1+q')^2+\mu^2]S + 2\mu[1-(1-\mu^2)(q')^2]T}{[1+2q'+(1-\mu^2)(q')^2]^2}. \end{aligned} \tag{S71}$$

Once the values of q' and σ are obtained from Eq. (S71), other order parameters in Eq. (S68) can be computed directly. Note that when $\theta = 0$, then $q' = b/2$. For a general threshold value $\theta \geq 0$, q' is proportional to the gain parameter b and can thus be regarded as an order parameter quantifying the 'effective gain parameter' in the network. We see from Eq. (S71) that q' depends on σ, which is scaled in turn by the quantities measuring the total stimulus strength, S and T. Thus, the changes of q' in the match versus mismatch condition can be viewed as a global gain component in the predictive signal.

The single neuron firing-rate [Eq. (S34)] is now,

$$r^\star = \phi(bI) = [bI-\theta]_+, \qquad I = \sum_{k=1}^{K}\left[w^k\delta x^k + v^k\delta y^k\right] \sim \mathcal{N}(0,\sigma^2). \tag{S72}$$

Notice that the mean and variance of the firing-rate [Eq. (S72)] can be obtained from Eq. (S70) by setting $\alpha = 0$, and that the variable I in this case coincides the voltage level of neurons in the network [Eq. (S67)]. These results are used to generated the firing-rate statistics in Fig. 1.

In the case where only one stimulus-pair is presented ($K = 1$), the Pearson correlation between firing-rate vectors in the mismatch and match conditions can be calculated as follows. We denote by I_x, I_y, I_{xy} the voltage levels in the x-only, y-only mismatch and match conditions, respectively. The I's are multivariate Gaussian variables with mean 0. We computed the correlations between inputs to neurons in the different mismatch conditions $\rho^I_{x,y} = (\langle I_x I_y\rangle - \langle I_x\rangle\langle I_y\rangle)/(\sigma_x\sigma_y)$ and between the mismatch and match conditions $\rho^I_{x,xy} = (\langle I_x I_{xy}\rangle - \langle I_x\rangle\langle I_{xy}\rangle)/(\sigma_x\sigma_{xy})$. Here σ_x^2, σ_y^2, σ_{xy}^2, are the variances of I_x, I_y, I_{xy}, respectively. We found,

$$\begin{aligned} \rho^I_{x,y} &= -\frac{2\mu[1+(1-\mu^2)(q')^2]}{(1-\mu^2)(1+q')^2+\mu^2}, \\ \rho^I_{x,xy} &= \frac{2(1+\mu)[1+(1-\mu)q']^2}{\sqrt{[(1-\mu^2)(1+q')^2+\mu^2][\mu(1+\mu)+(1-\mu^2)(1+2q'+(1-\mu)(q')^2)]}}. \end{aligned} \tag{S73}$$

From the symmetry in the model we have $\rho^I_{x,xy} = \rho^I_{y,xy}$.

In most cases, the experimentally accessible quantity is the firing-rate rather than the input current, so we also computed the Pearson correlation between firing-rates. We denote this correlation as $\rho^r_{m,n}$, where m, n can refer to the conditions x, y, xy, and write its formal definition,

$$\rho^r_{m,n} = \frac{\langle [bI_m - \theta]_+ [bI_n - \theta]_+ \rangle - \langle [bI_m - \theta]_+ \rangle \langle [bI_n - \theta]_+ \rangle}{\sqrt{\mathrm{Var}([bI_m - \theta]_+)\mathrm{Var}([bI_n - \theta]_+)}}. \tag{S74}$$

When $\theta = 0$, the cross covariance between firing-rates can be worked out as,

$$\langle [bI_m - \theta]_+ [bI_n - \theta]_+ \rangle = \frac{b^2 \sigma_m \sigma_n}{2\pi} \left(\frac{\pi}{2} \rho^I_{m,n} + \rho^I_{m,n} \arctan \frac{\rho^I_{m,n}}{\sqrt{1 - (\rho^I_{m,n})^2}} + \sqrt{1 - (\rho^I_{m,n})^2} \right). \tag{S75}$$

Together with the firing-rate mean and variance [Eq. (S70)], we obtained an explicit expression at for the firing-rate Pearson correlation, $\rho^r_{m,n}$. In the case of $\theta = 0$, Eq. (S74) becomes,

$$\rho^r_{m,n} = \frac{1}{\pi - 1} \left[\frac{\pi}{2} \rho^I_{m,n} + \rho^I_{m,n} \arctan \frac{\rho^I_{m,n}}{\sqrt{1 - (\rho^I_{m,n})^2}} + \sqrt{1 - (\rho^I_{m,n})^2} - 1 \right]. \tag{S76}$$

2.3. Balance level distribution

The balance level for neuron i in the network is defined as,

$$B_i = \left| \frac{I^F_i}{I^F_i - I^R_i} \right| = \left| \frac{\sum_{k=1}^P (w^k_i x^k + v^k_i y^k)}{\sum_{k=1}^P (w^k_i \delta x^k + v^k_i \delta y^k)} \right|. \tag{S77}$$

The B_i's are i.i.d. random variables for each i. Here the denominator is the net input $\delta I_i = I^F_i - I^R_i$ to neuron i, i.e., the difference between feedforward and recurrent input currents,

$$I^F = \sum_{k=1}^K (w^k x^k + v^k y^k), \qquad I^R = \sum_{k=1}^K (w^k \hat{x}^k + v^k \hat{y}^k). \tag{S78}$$

From Eq. (S66), δI can be expressed as

$$\delta I = \sum_{k=1}^P (w^k \delta x^k + v^k \delta y^k) = I - \frac{\alpha \hat{q}' b}{1 + \alpha \hat{q}' b} \left[I - \frac{\theta}{b} \right]_+ , \tag{S79}$$

where $I = \sqrt{\alpha\hat{q}}z + \sum_{k=1}^{K}(w^k\delta x^k + v^k\delta y^k)$ is defined in Eq. (S67). To simplify the notation we drop the subscript i from δI. Thus, to sample from the distribution of balance levels, one can first sample (w^k, v^k, z) from their corresponding distributions and then compute I^F and δI. The ratio between I^F and δI gives a sample of the balance level.

When the synaptic weights have Gaussian distribution and $\alpha = 0$, the pair $(I^F, \delta I)$ is jointly Gaussian,

$$(I^F, \delta I) \sim \mathcal{N}\left(0, \begin{pmatrix} \sigma_F^2 & \rho_B\sigma_F\sigma_\delta \\ \rho_B\sigma_F\sigma_\delta & \sigma_\delta^2 \end{pmatrix}\right). \tag{S80}$$

The coefficients of the covariance matrix of $(I^F, \delta I)$ are,

$$\begin{aligned} \sigma_F^2 &= 2\left(S + \mu T\right), \\ \sigma_\delta^2 &= \frac{2[(1-\mu^2)(1+q')^2 + \mu^2]S + 2\mu[1-(1-\mu^2)(q')^2]T}{[1+2q'+(1-\mu^2)(q')^2]^2}, \\ \rho_B\sigma_F\sigma_\delta &= \frac{2[1+(1-\mu^2)q']S + 2\mu T}{1+2q'+(1-\mu^2)(q')^2}. \end{aligned} \tag{S81}$$

The balance level in this case can be expressed using a Cauchy random variable ξ as,

$$B = \frac{\sigma_F}{\sigma_\delta}|\xi|, \tag{S82}$$

where the probability density function for $\xi \in \mathbb{R}$ is,

$$p(\xi) = \frac{1}{\pi}\frac{\sqrt{1-\rho_B^2}}{(\xi-\rho_B)^2 + 1 - \rho_B^2}. \tag{S83}$$

This result means that the average of the balance level distribution diverges. We use the quantiles to measure the magnitude of the balance level in the network (Fig. 2).

3. CHARACTERIZING DIFFERENT FUNCTIONAL NEURON TYPES

3.1. Firing-rate correlations from two-body replica calculations

In this section we compute the probabilities of single neurons belonging to the different functional cell types for *two* stimulus-pairs. Since the stimulus-pairs and the neurons are statistically equivalent, we focus on the responses of neuron i to the first two stimulus-pairs, $(\, h_i^{x_1},\, h_i^{y_1},\, h_i^{x_1y_1},\, h_i^{x_2},\, h_i^{y_2},\, h_i^{x_2y_2}\,)$. To mathematically characterize those voltage responses, we consider the joint distribution of the neurons' firing-rates in two different stimulus conditions,

$$p(r_1, r_2) = \frac{1}{N}\sum_{i=1}^{N} \delta(r_1 - r_i^A)\delta(r_2 - r_i^B). \tag{S84}$$

The superscripts A, B denote the stimulus conditions, i.e., A and B are chosen from $\{x_1, y_1, x_1y_1, x_2, y_2, x_2y_2\}$. We will show that at the limit $N \to \infty$, the joint distribution for all different combinations of stimulus conditions can be obtained from the calculation of pairwise firing-rate correlations [Eq. (S84)].

To evaluate Eq. (S84), we consider two identical networks driven by different stimulus inputs. The energy function of the 1st system with firing-rates $\boldsymbol{r}^A$ is,

$$E_0^A(\boldsymbol{r}^A; \{\boldsymbol{w}^k, \boldsymbol{v}^k\}) = \sum_{k=1}^{P} b\left(-x_A^k \boldsymbol{w}^k \cdot \boldsymbol{r}^A - y_A^k \boldsymbol{v}^k \cdot \boldsymbol{r}^A + \frac{1}{2N}\left[(\boldsymbol{w}^k \cdot \boldsymbol{r}^A)^2 + (\boldsymbol{v}^k \cdot \boldsymbol{r}^A)^2\right]\right) \\ + \sum_{i=1}^{N} F(r_i^A), \tag{S85}$$

and similarly for the energy function of the 2nd system, $E_0^B(\boldsymbol{r}^B; \{\boldsymbol{w}^k, \boldsymbol{v}^k\})$. Note that in stimulus conditions A and B, only the first two stimulus-pair inputs are nonzero.

The partition function of the whole system is defined as

$$Z_{\text{total}} = \int_{\mathbb{R}_+^{2N}} e^{-\beta E_0^A(\boldsymbol{r}^A;\{\boldsymbol{w}^k,\boldsymbol{v}^k\}) - \beta E_0^B(\boldsymbol{r}^B;\{\boldsymbol{w}^k,\boldsymbol{v}^k\})} \mathrm{d}\boldsymbol{r}^A \mathrm{d}\boldsymbol{r}^B = Z_A \cdot Z_B. \tag{S86}$$

Again we use the replica trick,

$$\lim_{N\to\infty} \frac{\ln Z_{\text{total}}}{2N} = \lim_{N\to\infty} \left\langle \frac{\ln Z_{\text{total}}}{2N} \right\rangle_{\boldsymbol{w},\boldsymbol{v}} = \lim_{n\to 0}\lim_{N\to\infty} \frac{\ln \langle Z_{\text{total}}^n \rangle}{2nN} = \lim_{n\to 0}\lim_{N\to\infty} \frac{\ln \langle Z_A^n Z_B^n \rangle}{2nN}. \tag{S87}$$

Note that the neural activities $\boldsymbol{r}^A$ and $\boldsymbol{r}^B$ of the two *separate but identical* networks are in fact statistically coupled due to the replica-average over $(\boldsymbol{w}^k, \boldsymbol{v}^k)$. The calculation for $\langle Z_A^n Z_B^n \rangle$ is similar to the one shown in §2. We denote the order parameters under replica symmetric ansatz as,

$$\begin{aligned} q_A^{ab} =& \frac{1}{N}\sum_j r_j^{A,a} r_j^{A,b} = q_0^A \delta_{ab} + q_1^A(1 - \delta_{ab}), \\ q_B^{ab} =& \frac{1}{N}\sum_j r_j^{B,a} r_j^{B,b} = q_0^B \delta_{ab} + q_1^B(1 - \delta_{ab}), \\ q_c^{ab} =& \frac{1}{N}\sum_j r_j^{A,a} r_j^{B,b} = q_{c,0} \delta_{ab} + q_{c,1}(1 - \delta_{ab}). \end{aligned} \tag{S88}$$

The last order parameter represents the overlap between replicas in system A and system B. Thus, the calculation of firing-rate correlations is very similar to the one-step replica symmetry-breaking calculation where the overlap between replicas within the same system is different from the overlap between the systems [79].

In the $N, P \to \infty, P/N \to \alpha, n \to 0$ limit, with similar changes of variables as before [Eq. (S58)], we write the result of the calculation as,

$$\frac{\ln \langle Z_A^n Z_B^n \rangle}{nN} = \mathcal{F}_{\text{total}}(q, \hat{q}, q', \hat{q}', t, s). \tag{S89}$$

Each order parameter in the function $\mathcal{F}_{\text{total}}$ has 3 components. For example, q has the components (q_A, q_B, q_c). The calculation gives the function $\mathcal{F}_{\text{total}}$,

$$\begin{aligned}
&\mathcal{F}_{\text{total}}(q, \hat{q}, q', \hat{q}', t, s) = \\
&\quad \frac{g\beta}{2}\sum_k \left((t_A^k)^2 + (s_A^k)^2 + (t_B^k)^2 + (s_B^k)^2\right) + \frac{\alpha}{2}\left(\hat{q}'_A q'_A + \hat{q}'_B q'_B + \hat{q}'_c q'_c\right) \\
&\quad + \frac{\alpha g\beta}{2}(\hat{q}_A q'_A - \hat{q}_A q'_A + \hat{q}'_B q_B - \hat{q}_B q'_B + \hat{q}'_c q_c - \hat{q}_c q'_c) \\
&\quad + \int D\boldsymbol{z} \ln\left[\int \mathrm{d}\nu_\beta(r_A)\mathrm{d}\nu_\beta(r_B)\left\langle e^{-\beta\mathcal{G}(r_A, r_B, \boldsymbol{z}, w^k, v^k)}\right\rangle\right] - \lim_{n\to 0}\frac{1}{2n}\left\langle \ln\det\mathcal{A}(\mu, q)\right\rangle_\mu .
\end{aligned} \tag{S90}$$

We introduced the functions,

$$\begin{aligned}
&\mathcal{G}(r_A, r_B, \boldsymbol{z}, w^k, v^k) = \\
&\quad \frac{g\alpha}{2}\hat{q}'_A r_A^2 - g r_A\sqrt{\alpha}\left(\sqrt{\hat{q}_c}z_1 + \sqrt{\hat{q}_A - \hat{q}_c}z_2\right) - g r_A \sum_k \left[(x_A^k - t_A^k)w^k + (y_A^k - s_A^k)v^k\right] \\
&\quad + \frac{g\alpha}{2}\hat{q}'_B r_B^2 - g r_B\sqrt{\alpha}\left(\sqrt{\hat{q}_c}z_1 + \sqrt{\hat{q}_B - \hat{q}_c}z_3\right) - g r_B \sum_k \left[(x_B^k - t_B^k)w^k + (y_B^k - s_B^k)v^k\right] \\
&\quad - g\alpha\hat{q}'_c r_A r_B, \\
&\mathcal{A}(\mu, q) = \begin{pmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{12} & \mathcal{A}_{22} \end{pmatrix} \\
&\quad \mathcal{A}_{11} = \begin{pmatrix} (1 + q'_A)I_n + g\beta q_A \mathbf{1}\mathbf{1}^\top & \mu(q'_A I_n + g\beta q_A \mathbf{1}\mathbf{1}^\top) \\ \mu(q'_A I_n + g\beta q_A \mathbf{1}\mathbf{1}^\top) & (1 + q'_A)I_n + g\beta q_A \mathbf{1}\mathbf{1}^\top \end{pmatrix} \\
&\quad \mathcal{A}_{12} = \begin{pmatrix} q'_c I_n + g\beta q_c \mathbf{1}\mathbf{1}^\top & \mu(q'_c I_n + g\beta q_c \mathbf{1}\mathbf{1}^\top) \\ \mu(q'_c I_n + g\beta q_c \mathbf{1}\mathbf{1}^\top) & q'_c I_n + g\beta q_c \mathbf{1}\mathbf{1}^\top \end{pmatrix}
\end{aligned} \tag{S91}$$

From symmetry, $\mathcal{A}_{22}$ is obtained by replacing $A \leftrightarrow B$ in $\mathcal{A}_{11}$. Here, I_n is the $n \times n$ identity matrix and $\mathbf{1}$ is an n-dimensional vector of 1's. All the order parameters in Eq. (S90) should be evaluated at the saddle point, in the limit $\beta \to \infty$. We obtain these order parameters as follows.

First, we find the Hamiltonian corresponding to this system,

$$\begin{aligned}\mathcal{H}(r_A, r_B) &= \mathcal{G}(r_A, r_B, \boldsymbol{z}, w^k, v^k) + F(r_A) + F(r_B)\\ &= \frac{g\alpha}{2}\hat{q}'_A r_A^2 - g r_A \sqrt{\alpha}\left(\sqrt{\hat{q}_c} z_1 + \sqrt{\hat{q}_A - \hat{q}_c} z_2\right) - g r_A \sum_k \left[(x_A^k - t_A^k) w^k + (y_A^k - s_A^k) v^k\right]\\ &\quad + \frac{g\alpha}{2}\hat{q}'_B r_B^2 - g r_B \sqrt{\alpha}\left(\sqrt{\hat{q}_c} z_1 + \sqrt{\hat{q}_B - \hat{q}_c} z_3\right) - g r_B \sum_k \left[(x_B^k - t_B^k) w^k + (y_B^k - s_B^k) v^k\right]\\ &\quad - g\alpha \hat{q}'_c r_A r_B + F(r_A) + F(r_B). \end{aligned} \tag{S92}$$

The extra terms $F(r_A)$ and $F(r_B)$ come from the probability measure $\mathrm{d}\nu_\beta$. When $\beta \to \infty$, the unique minimum $(r_A^\star, r_B^\star)$ is given by,

$$\begin{aligned} r_A^\star &= \phi\left(-g\alpha\hat{q}'_A r_A^\star + g\sqrt{\alpha}\left(\sqrt{\hat{q}_c} z_1 + \sqrt{\hat{q}_A - \hat{q}_c} z_2\right) + g\sum_{k=1}^{K}\left[(x_A^k - t_A^k) w^k + (y_A^k - s_A^k) v^k)\right] + g\alpha\hat{q}'_c r_B^\star\right),\\ r_B^\star &= \phi\left(-g\alpha\hat{q}'_B r_B^\star + g\sqrt{\alpha}\left(\sqrt{\hat{q}_c} z_1 + \sqrt{\hat{q}_B - \hat{q}_c} z_3\right) + g\sum_{k=1}^{K}\left[(x_B^k - t_B^k) w^k + (y_B^k - s_B^k) v^k)\right] + g\alpha\hat{q}'_c r_A^\star\right). \end{aligned} \tag{S93}$$

At the saddle point, the derivative of $\mathcal{F}_{\text{total}}$ [Eq. (S90)] with respect to $\hat{q}_c$ is set to 0, giving

$$0 = \frac{\alpha g \beta}{2}\left\{-q'_c + \frac{1}{\sqrt{\alpha}}\int \mathrm{d}Q_\beta \left[\left(\frac{z_1}{\sqrt{\hat{q}_c}} - \frac{z_2}{\sqrt{\hat{q}_A - \hat{q}_c}}\right) r_A + \left(\frac{z_1}{\sqrt{\hat{q}_c}} - \frac{z_3}{\sqrt{\hat{q}_B - \hat{q}_c}}\right) r_B\right]\right\}. \tag{S94}$$

In the limit $\beta \to \infty$ we find that,

$$\left\langle \left(\frac{z_1}{\sqrt{\hat{q}_c}} - \frac{z_2}{\sqrt{\hat{q}_A - \hat{q}_c}}\right) r_A^\star \right\rangle_z = \left\langle \frac{1}{\sqrt{\hat{q}_c}}\frac{\partial r_A^\star}{\partial z_1} - \frac{1}{\sqrt{\hat{q}_A - \hat{q}_c}}\frac{\partial r_A^\star}{\partial z_2}\right\rangle_z \overset{\text{Eq. (S93)}}{=} 0. \tag{S95}$$

Similarly, the average over the term proportional to $r_B^\star$ in Eq. (S94) is also 0. Substituting this into Eq. (S94), we get at the saddle point,

$$q'_c = 0. \tag{S96}$$

Next, we simplify the determinant of $\mathcal{A}(\mu, q)$. It is useful to write the submatrices as,

$$\begin{aligned} \mathcal{A}_{11} &= I_{2n} + q'_A \begin{pmatrix} 1 & \mu \\ \mu & 1 \end{pmatrix} \otimes I_n + g\beta q_A \begin{pmatrix} 1 & \mu \\ \mu & 1 \end{pmatrix} \otimes \mathbf{1}\mathbf{1}^\top,\\ \mathcal{A}_{12} &= q'_c \begin{pmatrix} 1 & \mu \\ \mu & 1 \end{pmatrix} \otimes I_n + g\beta q_c \begin{pmatrix} 1 & \mu \\ \mu & 1 \end{pmatrix} \otimes \mathbf{1}\mathbf{1}^\top. \end{aligned} \tag{S97}$$

The symbol $\otimes$ denotes the Kronecker product between two matrices. For this product, we have the identity, $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. Therefore the submatrices $\mathcal{A}_{11}$ and $\mathcal{A}_{12}$ commute and the determinant of $\mathcal{A}(\mu, q)$ becomes,

$$\det \mathcal{A}(\mu, q) = \det \begin{pmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{12} & \mathcal{A}_{22} \end{pmatrix} = \det(\mathcal{A}_{11}\mathcal{A}_{22} - \mathcal{A}_{12}^2). \tag{S98}$$

The two terms are equal to,

$$\begin{aligned} \mathcal{A}_{11}\mathcal{A}_{22} = I_{2n} + (q'_A + q'_B) \begin{pmatrix} 1 & \mu \\ \mu & 1 \end{pmatrix} \otimes I_n + q'_A q'_B \begin{pmatrix} 1+\mu^2 & 2\mu \\ 2\mu & 1+\mu^2 \end{pmatrix} \otimes I_n \\ + g\beta(ng\beta q_A q_B + q_A q'_B + q'_A q_B) \begin{pmatrix} 1+\mu^2 & 2\mu \\ 2\mu & 1+\mu^2 \end{pmatrix} \otimes \mathbf{1}\mathbf{1}^\top, \\ \mathcal{A}_{12}^2 = (2g\beta q'_c q_c + ng^2\beta^2 q_c^2 + O(q_c'^2)) \begin{pmatrix} 1+\mu^2 & 2\mu \\ 2\mu & 1+\mu^2 \end{pmatrix} \otimes \mathbf{1}\mathbf{1}^\top, \end{aligned} \tag{S99}$$

where we used $q'_c = 0$. Note that we have kept the term linear in q'_c in $\mathcal{A}_{12}^2$, anticipating that we will need to evaluate derivatives with respect to q'_c below. We find that the determinant has the following form,

$$\begin{aligned} \frac{1}{n} \ln \det(\mathcal{A}_{11}\mathcal{A}_{22} - \mathcal{A}_{12}^2) &= \frac{1}{n} \ln \det(Q_0 \otimes I_n + Q_1(n) \otimes \mathbf{1}\mathbf{1}^\top) \\ &= \ln \det Q_0 + \frac{1}{n} \ln \frac{\det(Q_0 + nQ_1(n))}{\det Q_0} \\ &\overset{n \to 0}{=} \ln \det Q_0 + \mathrm{tr}[Q_0^{-1} Q_1(0)]. \end{aligned} \tag{S100}$$

Here, Q_0 and $Q_1(n)$ are 2×2 matrices that depend on the order parameters,

$$\begin{aligned} Q_0 &= \begin{pmatrix} 1+q'_A & \mu q'_A \\ \mu q'_A & 1+q'_A \end{pmatrix} \begin{pmatrix} 1+q'_B & \mu q'_B \\ \mu q'_B & 1+q'_B \end{pmatrix}, \\ Q_1(n) &= \left[g\beta(q'_A q_B + q_A q'_B) + ng^2\beta^2(q_A q_B - q_c^2) - 2g\beta q'_c q_c + O(q_c'^2)\right] \begin{pmatrix} 1+\mu^2 & 2\mu \\ 2\mu & 1+\mu^2 \end{pmatrix}. \end{aligned} \tag{S101}$$

We use the above simplification to evaluate the derivative of $\mathcal{F}_{\text{total}}$ [Eq. (S90)] with respect to q_c and set it to 0 at the saddle point,

$$0 = \frac{\alpha g \beta}{2} \left[\hat{q}'_c - \frac{1}{2g\beta} \left\langle \mathrm{tr} \left[Q_0^{-1} \frac{\partial Q_1(0)}{\partial q_c} \right] \right\rangle_\mu \right]. \tag{S102}$$

Using Eq. (S101), we find that $\frac{\partial Q_1(0)}{\partial q_c}|_{q_c'=0} = 0$. Therefore,

$$\hat{q}_c' = 0. \tag{S103}$$

Substituting this result into Eq. (S93), we see this is consistent with the one-body replica results in Eq. (S34). Moreover, all the saddle point equations in the one-body scenario [Eq. (S60)] will hold in the two-body scenario. The two new equations when taking derivatives of $\mathcal{F}_{\text{total}}$ with respect to $\hat{q}_c'$ and q_c' are,

$$\begin{aligned} 0 &= \frac{\partial \mathcal{F}_{\text{total}}}{\partial \hat{q}_c'} = \frac{\alpha g \beta}{2}\left(\frac{q_c'}{g\beta} + q_c - \int \mathrm{d}Q_\beta r_A r_B\right), \\ 0 &= \frac{\partial \mathcal{F}_{\text{total}}}{\partial q_c'} = \frac{\alpha g \beta}{2}\left[-\frac{\hat{q}_c'}{g\beta} - \hat{q}_c - \frac{1}{2g\beta}\left\langle \operatorname{tr}\left[Q_0^{-1}\frac{\partial Q_1(0)}{\partial q_c'}\right]\right\rangle_\mu\right]. \end{aligned} \tag{S104}$$

The second equation can be further simplified as follows. From Eq. (S101), we find

$$\left.\frac{\partial Q_1(0)}{\partial q_c'}\right|_{q_c'=0} = -2q_c g\beta \begin{pmatrix} 1+\mu^2 & 2\mu \\ 2\mu & 1+\mu^2 \end{pmatrix}. \tag{S105}$$

Combining with Eq. (S101), we get

$$\begin{aligned} \frac{1}{2g\beta}\operatorname{tr}\left[Q_0^{-1}\frac{\partial Q_1(0)}{\partial q_c'}\right]\Bigg|_{q_c'=0} &= -\frac{(1-\mu)^2 q_c}{[1+(1-\mu)q_A'][1+(1-\mu)q_B']} - \frac{(1+\mu)^2 q_c}{[1+(1+\mu)q_A'][1+(1+\mu)q_B']} \\ &\equiv -C(\mu, q_A', q_B')q_c. \end{aligned} \tag{S106}$$

Therefore, using Eqs. (S104-S106), in the limit $\beta \to \infty$, we get,

$$\hat{q}_c = \langle C(\mu, q_A', q_B')\rangle_\mu \, q_c = \langle C(\mu, q_A', q_B')\rangle_\mu \, \langle r_A^\star r_B^\star\rangle_{w^k, v^k, \boldsymbol{z}}\,. \tag{S107}$$

Below we consider the case where the activation function ϕ is ReLU and μ's are the same for all learned stimulus-pairs. In this case $r_A^\star$ and $r_B^\star$ can be solved in closed form,

$$\begin{aligned} r_A^\star(w^k, v^k, \boldsymbol{z}) &= \frac{g}{1+\alpha \hat{q}_A' g}\left[I_A - \frac{\theta}{g}\right]_+, \\ r_B^\star(w^k, v^k, \boldsymbol{z}) &= \frac{g}{1+\alpha \hat{q}_B' g}\left[I_B - \frac{\theta}{g}\right]_+, \\ \frac{\hat{q}_c}{C(\mu, q_A', q_B')} &= q_c = \frac{g^2}{(1+\alpha g \hat{q}_A')(1+\alpha g \hat{q}_B')}\left\langle \left[I_A - \frac{\theta}{g}\right]_+ \left[I_B - \frac{\theta}{g}\right]_+ \right\rangle_{I_A, I_B}. \end{aligned} \tag{S108}$$

The random variables representing the currents can be read from Eq. (S93),

$$\begin{aligned} I_A &= \sqrt{\alpha}\left(\sqrt{\hat{q}_c} z_1 + \sqrt{\hat{q}_A - \hat{q}_c} z_2\right) + \sum_{k=1}^{K}\left[(x_A^k - t_A^k)w^k + (y_A^k - s_A^k)v^k\right], \\ I_B &= \sqrt{\alpha}\left(\sqrt{\hat{q}_c} z_1 + \sqrt{\hat{q}_B - \hat{q}_c} z_3\right) + \sum_{k=1}^{K}\left[(x_B^k - t_B^k)w^k + (y_B^k - s_B^k)v^k\right]. \end{aligned} \tag{S109}$$

In summary, to obtain the joint distribution of neural activity under two stimulus conditions A and B, we first sample w^k, v^k, $\boldsymbol{z}$ from their corresponding distributions, and calculate I_A and I_B from Eq. (S109). The order parameters $q_A, q'_A, q_B, q'_B, \hat{q}_A, \hat{q}'_A, \hat{q}_B, \hat{q}'_B$ are then solved from the one-body replica equations [Eq. (S60)] and order parameters introduced in the two-body calculation $(\hat{q}_c, \hat{q}'_c, q_c, q'_c)$ are obtained from Eq. (S108). Finally, $r^\star_A$ and $r^\star_B$ are calculated from Eq. (S93), which gives a random sample from the joint distribution.

The joint distribution of voltage levels can be computed similarly using the following formula for $h^\star_A$,

$$h^\star_A(w^k, v^k, \boldsymbol{z}) = I_A - \frac{\alpha \hat{q}'_A g}{1 + \alpha \hat{q}'_A g}\left[I_A - \frac{\theta}{g}\right]_+ . \tag{S110}$$

A similar formula holds for $h^\star_B$ with the corresponding input current and order parameters. These results can be generalized to scenarios with more than two stimulus conditions. This joint distribution was be used to calculate the fraction of different functional neuronal types as shown in Fig. 3b,d.

3.2. Explicit formulas in the Gaussian case

When the weights w^k and v^k are Gaussian, the input currents I_A and I_B are also Gaussian variables. Moreover, their variances $\sigma_A^2 = \langle I_A^2\rangle$ and $\sigma_B^2 = \langle I_B^2\rangle$ are given by the one-body calculation [Eq. (S69)]. We denote the input current covariance as $\sigma_{AB} = \langle I_A I_B\rangle$. This covariance is given by,

$$\sigma_{AB} = \alpha\hat{q}_c + \sum_{k=1}^{K}\left[\delta x^k_A \delta x^k_B + \mu^k(\delta x^k_A \delta y^k_B + \delta x^k_B \delta y^k_A) + \delta y^k_A \delta y^k_B\right] \equiv \alpha \hat{q}_c + \sigma^0_{AB}. \tag{S111}$$

Here σ^0_{AB} can be obtained from the one-body replica equations [Eq. (S60)].

Substituting σ_{AB} this into Eq. (S108) (i.e., averaging over the correlated Gaussians I_A, I_B), and defining $\rho_{AB} = \sigma_{AB}/(\sigma_A\sigma_B)$, we get a self-consistent equation for ρ_{AB},

$$\begin{aligned}\rho_{AB} - \frac{\sigma^0_{AB}}{\sigma_A\sigma_B} = {} & \frac{\alpha b^2 C(\mu, q'_A, q'_B)\sqrt{1-\rho^2_{AB}}}{(1+\alpha b\hat{q}'_A)(1+\alpha b\hat{q}'_B)} \\ & \times \int_{\frac{\theta}{b}}^{+\infty} Dz \left[\frac{1}{2\pi} e^{-\frac{(b\rho_{AB}z-\theta)^2}{2b^2(1-\rho^2_{AB})}} + \frac{b\rho_{AB}z-\theta}{b\sqrt{2\pi(1-\rho^2_{AB})}} H\left(-\frac{b\rho_{AB}z-\theta}{b\sqrt{1-\rho^2_{AB}}}\right)\right]\left(z - \frac{\theta}{b}\right).\end{aligned} \tag{S112}$$

When $\theta = 0$, the above equation for ρ_{AB} simplifies into

$$\rho_{AB} - \frac{\sigma^0_{AB}}{\sigma_A \sigma_B} = \frac{\alpha b^2 C(\mu, q'_A, q'_B)}{2\pi(1+\alpha b \hat{q}'_A)(1+\alpha b \hat{q}'_B)} \left(\frac{\pi}{2}\rho_{AB} + \rho_{AB} \arctan \frac{\rho_{AB}}{\sqrt{1-\rho^2_{AB}}} + \sqrt{1-\rho^2_{AB}} \right). \tag{S113}$$

Note that the quantity ρ_{AB} calculated here is the high-dimensional counterpart of Eq. (S73), i.e., ρ_{AB} reduces to $\rho^I_{m,n}$ in the limit $\alpha \to 0$. We computed the firing rate correlations in the high-dimensional regime based on Eqs. (S74-S76) which give the correlation between the input current ρ_{AB}.

The fraction of different functional neuronal types can also be obtained from the statistics σ^2_A, σ^2_B and ρ_{AB}. Specifically, we set the stimulus conditions $A =$'x-only mismatch condition' and $B =$ 'match condition'. The fraction of PE and R neurons are defined as (Methods),

$$\begin{aligned} f_{PE} &= \mathbb{P}\left\{ h_A > \frac{\sigma}{2}, h_A - h_B > \frac{\sigma}{2} \right\}, \\ f_R &= \mathbb{P}\left\{ h_A > \frac{\sigma}{2}, |h_A - h_B| < \frac{\sigma}{2} \right\}, \end{aligned} \tag{S114}$$

where h_A, h_B are given by Eq. (S110). In the low-dimensional limit $\alpha \to 0$, $h_A = I_A, h_B = I_B$ and have multivariate Gaussian distribution. There the fractions of PE and R neurons have the explicit formulas,

$$\begin{aligned} f_{PE} &= \int_{\frac{\sigma}{2\sigma_A}}^{+\infty} Dz\, H\left(\frac{\frac{\sigma}{2\sigma_A} - (\frac{\sigma_A}{\sigma_B} - \rho_{AB})z}{\sqrt{1 - (\frac{\sigma_A}{\sigma_B} - \rho_{AB})^2}} \right), \\ f_R &= \int_{\frac{\sigma}{2\sigma_A}}^{+\infty} Dz \left[1 - H\left(\frac{\frac{\sigma}{2\sigma_A} - (\frac{\sigma_A}{\sigma_B} - \rho_{AB})z}{\sqrt{1 - (\frac{\sigma_A}{\sigma_B} - \rho_{AB})^2}} \right) \right]. \end{aligned} \tag{S115}$$

3.3. Imperfect match of paired stimuli

We consider a network that learns a single stimulus association, and is presented with a 'probe' stimulus that is an imperfect match to the expected (learned) stimulus. This difference is modeled by letting the recurrent weight vector $\boldsymbol{w}$ be different from the feedforward weight vector $\boldsymbol{w}'$, giving the dynamics,

$$\frac{\mathrm{d}h_i(t)}{\mathrm{d}t} = -h_i(t) - \frac{b}{N}\sum_{j=1}^{N}(w_i w_j + v_i v_j)\phi(h_j(t)) + b\,(w'_i x + v_i y). \tag{S116}$$

We used this model to understand recent experimental findings, where a motor-auditory association was learned, and animals were probed with sounds that differed from the learned tone [13]. We assume that the components of $\boldsymbol{w}'$ have mean 0 and unit variance [similarly to $\boldsymbol{w}$ and $\boldsymbol{v}$, Eq. (S9)], and the following cross terms,

$$\langle w_i w'_j \rangle = \delta_{ij}\kappa, \qquad \langle v_i w'_j \rangle = \delta_{ij}\kappa\mu. \tag{S117}$$

Here $0 \le \kappa \le 1$ indicates the similarity between the learned stimulus input x and the one used as a probe. When $\kappa = 1$, the learned and probe stimuli are equal.

This network is very similar to the special case $\alpha \to 0$ of the network studied in §2.2.2. To understand its steady-state response, we use Eq. (S34) and define similarly,

$$r^\kappa = \phi(I^\kappa) = [bI^\kappa - \theta]_+, \qquad I^\kappa = w'x - w\hat{x} + v(y - \hat{y}). \tag{S118}$$

Here ϕ is assumed to be the ReLU function, $\hat{x}$ and $\hat{y}$ are the internal predictions [Eq. (S1)] and are given by the saddle point equations [Eq. (S68)],

$$\begin{aligned} \hat{x} &= \frac{\kappa[q' + (1-\mu^2)(q')^2]x + \mu q' y}{1 + 2q' + (1-\mu^2)(q')^2}, \\ \hat{y} &= \frac{\kappa\mu q' x + [q' + (1-\mu^2)(q')^2]y}{1 + 2q' + (1-\mu^2)(q')^2}. \end{aligned} \tag{S119}$$

Note that we have modified them accordingly to account for fact that stimulus-pairing is 'imperfect'. When all the weights have Gaussian distributions, the order parameters q', σ satisfy [similarly to Eq. (S71)],

$$\begin{aligned} q' &= bH\left(\frac{\theta}{b\sigma^\kappa}\right), \\ (\sigma^\kappa)^2 &= 2(1-\kappa^2) + \frac{2\kappa^2[(1-\mu^2)(1+q')^2 + \mu^2]S + 2\kappa\mu[1 - (1-\mu^2)(q')^2]T}{[1 + 2q' + (1-\mu^2)(q')^2]^2}. \end{aligned} \tag{S120}$$

We computed the representation similarity between stimuli semi-analytically by first solving q', σ^κ, sampling I^κ from $\mathcal{N}(0, (\sigma^\kappa)^2)$, and finally calculating the Pearson correlation coefficient [Eq. (S74)] between $r^{\kappa=1}$ and r^κ for different values of κ.

To get the segregation index, we considered the difference between mismatch and match responses Δ for an arbitrary κ and $\kappa = 1$,

$$\begin{aligned} \Delta^\kappa &= [bI^\kappa_x - \theta]_+ - [bI^\kappa_{xy} - \theta]_+, \\ \Delta^{\kappa=1} &= [bI^{\kappa=1}_x - \theta]_+ - [bI^{\kappa=1}_{xy} - \theta]_+. \end{aligned} \tag{S121}$$

Note that I^κ_x, I^κ_{xy}, $I^{\kappa=1}_x$ and $I^{\kappa=1}_{xy}$ are random variables that depend on the random weights w', w and v, order parameters $\hat{x}$ and $\hat{y}$, and the inputs x and y. The inputs were chosen according to the stimulus condition (match/mismatch). The segregation index (as a function of κ) is defined as the Pearson correlation between the two random variables Δ^{κ} and $\Delta^{\kappa=1}$, which is shown in Fig. 4f.

4. THE E/I NETWORK MODEL

4.1. Derivation of the E/I connectivity in the model

We consider a network with two separate populations of excitatory and inhibitory neurons. The time-dependent voltages of E and I neurons are given by the following system of differential equations,

$$\begin{aligned} \tau_E \frac{\mathrm{d}h_i^E}{\mathrm{d}t} &= -h_i^E + \sum_{j=1}^{N_E} J_{ij}^{EE}\phi(h_j^E) - \sum_{j=1}^{N_I} J_{ij}^{EI}\phi_I(h_j^I) + I_i^E, \\ \tau_I \frac{\mathrm{d}h_i^I}{\mathrm{d}t} &= -h_i^I + \sum_{j=1}^{N_E} J_{ij}^{IE}\phi(h_j^E) - \sum_{j=1}^{N_I} J_{ij}^{II}\phi_I(h_j^I) + I_i^I. \end{aligned} \tag{S122}$$

We assume that the activation function of inhibitory neurons is ReLU with threshold value equal to zero, $\phi_I(x) = \max\{x, 0\}$. Notice the negative sign of the third term in both equations. This implies that the connectivity matrices J^{EE}, J^{EI}, J^{IE} and J^{II} are non-negative. We now derive these matrices, and the inputs I^E and I^I, by matching the steady state activity of E neurons in the E/I network to the neural activity in the original network [Eq. (S4)]. At steady state, Eq. (S122) reads,

$$\begin{aligned} h_i^E &= \sum_{j=1}^{N_E} J_{ij}^{EE}\phi(h_j^E) - \sum_{j=1}^{N_I} J_{ij}^{EI}\phi_I(h_j^I) + I_i^E, \\ h_i^I &= \sum_{j=1}^{N_E} J_{ij}^{IE}\phi(h_j^E) - \sum_{j=1}^{N_I} J_{ij}^{II}\phi_I(h_j^I) + I_i^I. \end{aligned} \tag{S123}$$

We restrict ourselves to choices of connectivity in which inhibitory neurons operate in the linear regime, i.e., $h_i^I \geq 0 \Rightarrow \phi_I(h_i^I) = h_i^I$. Substituting h_i^I into h_i^E in Eq. (S123) we get,

$$h_i^E = \sum_{j=1}^{N_E} \left[J_{ij}^{EE} - (J^{IE}(I_{N_I} + J^{II})^{-1}J^{EI})_{ij} \right] \phi(h_j^E) + I_i^E - \sum_{j=1}^{N_I} J_{ij}^{EI} I_j^I. \tag{S124}$$

One can be check that the steady state solution is stable when $\tau_I \ll \tau_E$. Here $(I_{N_I} + J^{II})$ is assumed to be invertible. From now on we suppress the subscript N_I indicating the dimension of the identity matrix I_{N_I}. Equating this with the steady state in the original network [Eq. (S8)] gives the constraints on the connectivity and input,

$$J_{ij}^{EE} - [J^{EI}(I + J^{II})^{-1}J^{IE}]_{ij} = -\frac{b}{N}\sum_{k=1}^{P}\left(w_i^k w_j^k + v_i^k v_j^k\right),$$
$$I_i^E - [J^{EI}(I + J^{II})^{-1}I^I]_i = b\sum_{k=1}^{P}(w_i^k x^k + v_i^k y^k). \tag{S125}$$

Following a scheme for separating E/I connectivity used in previous work [54], we define positive random variables $\xi_i^k, \eta_i^k \geq 0$ such that the variables w_i^k, v_i^k are retrieved when the mean is subtracted from the new variables. Mathematically,

$$w_i^k = \xi_i^k - \bar{\xi}, \qquad v_i^k = \eta_i^k - \bar{\eta}. \tag{S126}$$

The means $\bar{\xi}$, $\bar{\eta}$ are chosen to be independent of the neuron and pattern indices i, k. Using the same trick as Ref. [54], the first equation in Eq. (S125) can be separated into two parts,

$$\begin{aligned} J_{ij}^{EE} &= \frac{\gamma b}{N}\sum_{k=1}^{P}\left(\xi_i^k \xi_j^k + \eta_i^k \eta_j^k\right) \\ &\quad + \frac{bP}{N}\left[\left(\sum_{k=1}^{P}\xi_i^k\right)\left(\sum_{k=1}^{P}\xi_j^k\right) + \left(\sum_{k=1}^{P}\eta_i^k\right)\left(\sum_{k=1}^{P}\eta_j^k\right)\right] \\ [J^{EI}(I + J^{II})^{-1}J^{IE}]_{ij} &= \frac{(\gamma+1)b}{N}\sum_{k=1}^{P}\left(\xi_i^k \xi_j^k + \eta_i^k \eta_j^k\right) \end{aligned} \tag{S127}$$

340 Here γ is an arbitrary positive number, which we set to 1 in all later results.

We make two additional assumptions: (*i*) 'Feedforward' stimulus input exclusively target excitatory neurons ($I_i^I = 0$); and (*ii*) I-to-E connectivity has the form $J^{EI} = \tilde{J}^{EI}(I + J^{II})$, where $\tilde{J}^{EI}$ is a nonnegative matrix. Given these, Eqs. (S125, S127) become,

$$[\tilde{J}^{EI}J^{IE}]_{ij} = \frac{2b}{N}\sum_{k=1}^{P}\left(\xi_i^k \xi_j^k + \eta_i^k \eta_j^k\right),$$
$$I_i^E = b\sum_{k=1}^{P}(w_i^k x^k + v_i^k y^k). \tag{S128}$$

To obtain the E/I balance level for excitatory neurons in this network, we write the total excitatory input $I_i^{E,\mathrm{tot}}$ as the sum of different contributions,

$$\begin{aligned}
\frac{I_i^{E,\mathrm{tot}}}{b} &= \sum_{k=1}^{P}(w_i^k\hat{x}^k + v_i^k\hat{y}^k) + \sum_{k=1}^{P}(w_i^k x^k + v_i^k y^k) && \text{(stimulus-specific, local)}\\
&+ 2\left(\bar{\xi}\sum_{k=1}^{P}\hat{x}^k + \bar{\eta}\sum_{k=1}^{P}\hat{y}^k\right) && \text{(stimulus-specific, global)}\\
&+ \frac{2}{N}\left(\bar{\xi}\sum_{k=1}^{P}w_i^k + \bar{\eta}\sum_{k=1}^{P}v_i^k\right)\sum_{i=1}^{N}\phi(r_i^E) && \text{(stimulus-nonspecific, local)}\\
&+ 2\alpha\left(\bar{\xi}^2 + \bar{\eta}^2\right)\sum_{i=1}^{N}\phi(r_i^E) + \bar{\xi}\sum_{k=1}^{P}x^k + \bar{\eta}\sum_{k=1}^{P}y^k. && \text{(stimulus-nonspecific, global)}
\end{aligned} \tag{S129}$$

Taking the ratio between the stimulus-specific, local component and the net input to each excitatory neuron, we get,

$$B_i^{E/I} = \left|\frac{I_i^R + I_i^F}{\delta I_i}\right| = \left|-1 + 2B_i\right|, \tag{S130}$$

where I_i^F, I_i^R, δI_i and B_i are those defined in the original network model [without separation of E and I; Eq. (S77)]. Therefore, for moderate values of $B_i > 1/2$, up to a scaling factor and shift, the stimulus-specific, local component of the E/I balance level is the same as the balance level we analyzed in Figs. 2, 3. Note that in the range of α values analyzed in Fig. 2, the fraction of neurons with $B_i < 1/2$ is negligible in both match and mismatch conditions.

4.2. Interpolation via nonnegative matrix factorization

Solving for $\tilde{J}^{EI}$ and J^{IE} in Eq. (S128) is equivalent to a nonnegative matrix factorization problem [53]. Using the shifted, nonnegative weight vectors, we define the matrices Ξ, H, S,

$$\Xi = \frac{1}{N}\begin{pmatrix}\boldsymbol{\xi}^{1\top}\\ \vdots \\ \boldsymbol{\xi}^{P\top}\end{pmatrix} = \frac{1}{N}\begin{pmatrix}\xi_1^1 & \dots & \xi_N^1\\ \vdots & \ddots & \vdots\\ \xi_1^P & \dots & \xi_N^P\end{pmatrix} \in \mathbb{R}^{P\times N},$$

$$H = \frac{1}{N}\begin{pmatrix}\boldsymbol{\eta}^{1\top}\\ \vdots \\ \boldsymbol{\eta}^{P\top}\end{pmatrix} = \frac{1}{N}\begin{pmatrix}\eta_1^1 & \dots & \eta_N^1\\ \vdots & \ddots & \vdots\\ \eta_1^P & \dots & \eta_N^P\end{pmatrix} \in \mathbb{R}^{P\times N}, \qquad S = \begin{pmatrix}\Xi\\ H\\ \mathbf{0}\end{pmatrix} \in \mathbb{R}^{N\times N}. \tag{S131}$$

Throughout this section, we will assume $2P \leq N$, and '$\mathbf{0}$' pads with 0's such that S is a square matrix. Thus, the connectivity equation [Eq. (S128)] can be rewritten as,

$$\tilde{J}^{EI} J^{IE} = 2b(\Xi^\top \Xi + H^\top H) = b(\gamma+1) S^\top S. \tag{S132}$$

For each choice of a nonnegative matrix J^{IE}, the above equation has a nonnegative solution J^{EI} if and only if the convex cone formed by the row vectors of J^{IE} contains the convex cone formed by the row vectors of S [formally denoted as $\mathrm{cone}(J^{IE}) \supseteq \mathrm{cone}(S)$]. This condition can be derived from the definition of matrix multiplication [53]. Based on this condition, we identify a family of solutions $\{J^{EI}(\lambda),\, J^{IE}(\lambda)\}$ parameterized by $\lambda \in [0,1]$ as follows. At one end, we choose J^{IE} equal to the identity ($J^{IE}(\lambda=0) = I_N$). At the other end, $J^{IE}(\lambda=1) = S'$, where S' is defined such that its first $2P$ rows are the same as the nonzero rows of S and the rest of its rows are randomly sampled from the vectors $\boldsymbol{\xi}^k/N$, $\boldsymbol{\eta}^k/N$. This ensures that $\mathrm{cone}(S') \supseteq \mathrm{cone}(S)$. This family of solutions assumes that the number of inhibitory neurons equal to the number of excitatory neurons.

The firing-rates of inhibitory neurons are given by,

$$r_i^I(\lambda) \equiv \phi_I(h_i^I) = h_i^I = \sum_{j=1}^N J_{ij}^{IE}(\lambda) r_j^E. \tag{S133}$$

At the two ends, this reduces to,

$$\begin{aligned} r_i^I(\lambda=0) &= r_i^E(0), \\ r_i^I(\lambda=1) &= \begin{cases} \hat{x}^k + \frac{\bar{\xi}}{N}\sum_{i=1}^N \phi(h_i^E), & \text{if the } i\text{th row of } S' \text{ is } \boldsymbol{\xi}^{k\top} \\ \hat{y}^k + \frac{\bar{\eta}}{N}\sum_{i=1}^N \phi(h_i^E), & \text{if the } i\text{th row of } S' \text{ is } \boldsymbol{\eta}^{k\top} \end{cases} \end{aligned} \tag{S134}$$

Based on these equations, we call $\lambda = 0$ the 'private' solution and $\lambda = 1$ the 'internal prediction' scenario. For $\lambda = 1$, the second term in $r_i^I(1)$ can be canceled by a global disinhibtory input. For intermediate λ's, it may seem natural to choose a linear interpolation between the two solutions, $J^{IE}(\lambda) = \lambda J^{IE}(1) + (1-\lambda) J^{IE}(0)$. We find however that this choice does not ensure that the solution for J^{EI} is nonnegative.

Instead, we choose E-to-I connectivity as follows. Two intermediate points within the segment $[0,1]$ are denoted as $\lambda = 0^+$ and $\lambda = 1^-$, thereby dividing the segment into three.

At those points we choose J^{IE} to be,

$$J^{IE}(0^+) = \left(\begin{array}{c|c} \Xi_{P,2P} & \mathbf{0} \\ H_{P,2P} & \mathbf{0} \\ \hline \mathbf{0} & NI_{N-2P} \end{array}\right), \qquad J^{IE}(1^-) = \left(\begin{array}{c|c} \Xi_{P,2P} & \Xi_{P,N-2P} \\ H_{P,2P} & H_{P,N-2P} \\ \hline \mathbf{0} & N\mathrm{diag}(\boldsymbol{a}) \end{array}\right). \tag{S135}$$

Here, $\Xi_{P,2P}, H_{P,2P}$ consist of the first P rows and first $2P$ columns of Ξ and H, respectively; $\Xi_{P,N-2P}, H_{P,N-2P}$ consist of the first P rows and last $N-2P$ columns of Ξ and H, respectively; $\mathrm{diag}(\boldsymbol{a})$ is a diagonal matrix, with diagonal elements given by the $N-2P$ components of the vector $\boldsymbol{a}$ which is specified below. Again the $\mathbf{0}$'s are used for padding.

The interpolation of $J^{IE}(\lambda)$ from $\lambda = 0$ to $\lambda = 1$ thus consists of three regions:

(I) λ from 0 to 0^+: The upper left block of J^{IE} changes from an identity matrix to a matrix of stimulus input vectors.

(II) λ from 0^+ to 1^-: The upper and lower right blocks linearly interpolate the matrices shown in Eq. (S135). Results in the main text are taken from here.

(III) λ from 1^- to 1: The lower part of the matrix changes to contain stimulus vectors.

We start with solutions in Region (II) which we found to be the most relevant to the empirical measurements in [12], since we estimated $\lambda \approx 0.6$. Network properties for a range of λ values between 0 and 1 (Figs. 5, S6, S7, S8) are also based on the results in Region (II). The connectivity matrices $J^{EI}(\lambda)$ and $J^{IE}(\lambda)$ in Region (II) are given by,

$$\begin{aligned} J^{IE}(\lambda) &= \left(\begin{array}{c|c} \Xi_{P,2P} & \lambda\Xi_{P,N-2P} \\ H_{P,2P} & \lambda H_{P,N-2P} \\ \hline \mathbf{0} & N[(1-\lambda)I_{N-2P} + \lambda\mathrm{diag}(\boldsymbol{a})] \end{array}\right), \\ \tilde{J}^{EI}(\lambda) &= 2b\begin{pmatrix} \Xi & H & \mathcal{J}(\lambda) \end{pmatrix}. \end{aligned} \tag{S136}$$

Here $\mathcal{J}(\lambda)$ is a $N \times (N-2P)$ matrix whose elements are given by

$$[\mathcal{J}(\lambda)]_{ij} = \frac{(1-\lambda)(\xi_i\xi_j + \eta_i\eta_j)}{(\lambda a_j + 1 - \lambda)N}, \quad i = 1, \ldots, N, \quad j = N-2P+1, \ldots, N. \tag{S137}$$

One can check that $\mathrm{cone}(J^{IE}(\lambda)) \supseteq \mathrm{cone}(S)$, and thus Eq. (S132) is satisfied and the elements of J^{EI} are nonnegative for every λ.

The interpolation in Region (I) requires smoothly 'morphing' the upper left block of the connectivity matrix involving Ξ and H to the identity matrix. This can be done by replacing the last row and last column with 0 and then setting the last diagonal element to be 1. Repeating this replacement P times yields the identity matrix. We note that in the low-dimensional case $[P = O(1)]$, this procedure only changes the E connections to P out of N inhibitory neurons. Thus its effect on the overall statistics of inhibitory neurons' activity is negligible. In the high-dimensional case $[P = O(N)]$, the distributions of neural activity and synaptic weights themselves change smoothly along this interpolation path. Similarly, in Region (III), we replace every row in the lower part of the matrix with one of the randomly sampled vectors that appear in the matrix S'.

4.3. Plasticity of inhibitory weights during learning

The interpolation solutions presented in the last section are valid for any set of positive real numbers a_i, $i = 2P + 1, \ldots, N$. In Fig. 5 we choose the a_i's as follows,

$$a_i(\mu) = \begin{cases} 1.4 + 12\exp[1.5 s_i(\mu)] & \text{if } s_i(\mu) \leq 0 \\ 0.002 & \text{if } 0 < s_i(\mu) < 0.97 \\ 2.002 & \text{if } s_i(\mu) \geq 0.97 \end{cases} \tag{S138}$$

where

$$s_i(\mu) = [r^E_{x,i}(\mu) - \langle r^E_x(\mu)\rangle][r^E_{xy,i}(\mu) - \langle r^E_{xy}(\mu)\rangle]. \tag{S139}$$

Here $r^E_{x,i}(\mu)$ and $r^E_{xy,i}(\mu)$ are the firing-rates of the i-th excitatory neuron in the x-only mismatch and match conditions for a given value of μ. $\langle r^E_x(\mu)\rangle$ and $\langle r^E_{xy}(\mu)\rangle$ are the average firing-rates over all the E neurons in the two conditions. This mathematical form for a_i is chosen to match the experimental data on fast spiking neurons (Fig. 5c,d).

To track individual synapses during learning, we generate the kth stimulus input vectors $\boldsymbol{\xi}^k$ and $\boldsymbol{\eta}^k$ as follows: (1) We first generate two independent isotropic Gaussian vectors $\boldsymbol{a}^k_0$, $\boldsymbol{b}^k_0$, with mean equal to 3 and standard deviation equal to 1; (2) Then we form the a linear combination to generate two correlated Gaussian random variables,

$$\boldsymbol{a}^k = \boldsymbol{a}^k_0, \qquad \boldsymbol{b}^k = \mu \boldsymbol{a}^k_0 + \sqrt{1-\mu^2}\boldsymbol{b}^k_0. \tag{S140}$$

(3) Finally, we clip both variables to positive and define them as $\boldsymbol{\xi}^k$ and $\boldsymbol{\eta}^k$. In this case, the resulting vectors $\boldsymbol{w}^k$ and $\boldsymbol{v}^k$ [Eq. (S126)] will be approximately correlated Gaussian variables with mean 0. These procedures are used to produce the plots in Fig. 5c,e,f,g.

5. PARAMETER VALUES USED IN THE FIGURES

Unless specified, in all the main and supplementary figures, $\boldsymbol{w}^k$ and $\boldsymbol{v}^k$ have joint Gaussian distribution and satisfy Eq. (S9). The number of neurons in the network is $N = 2000$.

Figure 1: We set $\alpha = 0$, $\theta = 0$ and $b = 150$ throughout this figure.

Panel b: We use Eq. (S72) to generate $N = 2000$ samples of 2D random variables (I_x, I_{xy}) and compute the corresponding firing-rates.

Panel d: The theory lines for the Pearson correlation between different stimulus conditions are calculated from Eqs. (S71, S73, S76). The simulation points are calculated by sampling the neurons' firing-rates as described in the Panel b caption. As each vector represents the mean-subtracted firing-rate vectors, the cosine of the angle is equivalent to the Pearson correlation coefficient between the original firing-rate vectors.

Panel f: The firing-rate distribution on the left ('Our model') is generated in the same way as in Panel b. As the neural responses to two stimulus-pairs are mutually independent at $\alpha = 0$, the joint distribution is a product of the corresponding marginal distributions. The firing-rate distribution on the right ('Segregated model') is generated by using the same marginal distributions (as in the plot of 'Our model'), but adding a nonzero correlation (which equals to 0.9) in the input variables (I_x, I_{xy}) that are used to calculate the firing-rates.

Figure 2: We set $\theta = 0$ throughout this figure.

Panel b: $\alpha = 0$, $b = 150$. The 'Early' and 'Late' plots for balance level distribution are calculated at $\mu = 0$ and $\mu = 0.9$ respectively.

Panel d: $\alpha = 0$, $b = 150$. For SVM classification, stimulus inputs in the mismatch condition are generated from Gaussian mixtures centered at $(0, 1)$ and $(1, 0)$, both of which are isotropic and have variance 0.05. Similar Gaussian mixtures are used for stimulus input in the match condition, except that the centers are at $(0, 0)$ and $(1, 1)$. The SVM model is fitted using the Matlab function 'fitcsvm'. The classification error is calculated via the matlab functions 'crossval' and 'kfoldLoss'.

Panel e: This figure panel is an illustration and the parameters are $\alpha = 0$, $\mu = 0.7$.

Panel f: The threshold on the firing-rate for determining the optimal b is chosen such that at $\alpha = 0$, the optimal balance level is the same as the one fitted to experimental data [20] in Figure 3 ($B^\star \approx 162$).

Figure 3: We fit both sets of experimental data [12, 20] using Eq. (8) (Methods).

Figure 4:

Panel b: $\alpha = 0$, $b = 150$.

Panel c: The values of b in both plots are chosen to be at the optimal values.

Panel d: Plotted on the y axis is the fraction of mixed-representation neurons among all PE neurons for the stimulus pair 1.

Panel f: We set $b = 189$, which is the value extracted from the data [12]. The sparsity levels are defined as the fraction of active neurons in the network and changed by varying the firing-rate threshold θ in the network model. The threshold values corresponding to the three plotted curves are $\theta = 4.5, 6.5, 21.5$.

Figure 5: We set $\theta = 0$, $J^{II} = 0$ throughout this figure. Before and after learning correspond to $\mu = 0$ and $\mu = 0.97$. During learning, the functional cell types of a specific E or I neuron in the network might change. The cell-type-specific synaptic weight statistics shown in Fig. 5f,g only include synapses whose pre- and postsynaptic neurons maintain their identity throughout learning. Other parameter values can be found in §4.3.

Figure 6: $\theta = 0$ throughout this figure. The number of neurons for each module is 400. All the error bars are computed based on 30 random samples of synaptic weight vectors. The steady state of the network is obtained by simulating the ODEs [Eq. (S29)] for total time $t = 4$.

Panel b : $\alpha = 0$, $\mu = 0.97$. The colormap indicates the firing rate averaged over all neurons in all modules in the x-only mismatch condition.

Panel c: $b_1 = b_3 = 50$ and $b_2 = 190$ are the values at the star position in panel b. $\mu = 0.97$.

Panel d-h: $b_1 = b_3 = 50$ and $b_2 = 190$ are the values at the star position in panel b.

Figure S1:

Panel a: $\alpha = 0$, $b = 150$.

Panel c: The threshold on firing-rate for determining optimal b is chosen such that at

$\alpha = 0$, the optimal balance level is the same as the one fitted to experimental data [20] in Fig. 3 ($B^\star \approx 162$). This threshold remains fixed for different values of α.

Figure S3: Throughout this figure, $\alpha = 0$, $b = 150$.

Figure S4: Throughout this figure, $\mu = 0.97, b = 150$.

Figure S5: The threshold value corresponding to the model curve is $\theta = -20$. We set $b = 189$, which is the value extracted from the data [12].

Figure S6: Throughout this figure, we set $\alpha = 0$, $b = 150$.

Figure S7 and S8: We set $\theta = 0$, $J^{II} = 0$ throughout these figures. Before and after learning correspond to $\mu = 0$ and $\mu = 0.97$. Other parameter values are the same as in Fig. 5.

[1] R. P. Rao and D. H. Ballard, “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects,” *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[2] G. B. Keller and T. D. Mrsic-Flogel, “Predictive processing: a canonical cortical computation,” *Neuron*, vol. 100, no. 2, pp. 424–435, 2018.

[3] J. Poort, A. G. Khan, M. Pachitariu, A. Nemri, I. Orsolic, J. Krupic, M. Bauza, M. Sahani, G. B. Keller, T. D. Mrsic-Flogel, *et al.*, “Learning enhances sensory and multiple non-sensory representations in primary visual cortex,” *Neuron*, vol. 86, no. 6, pp. 1478–1490, 2015.

[4] F. C. Widmer, S. M. O’Toole, and G. B. Keller, “NMDA receptors in visual cortex are necessary for normal visuomotor integration and skill learning,” *Elife*, vol. 11, p. e71476, 2022.

[5] S. J. Eliades and X. Wang, “Neural substrates of vocalization feedback monitoring in primate auditory cortex,” *Nature*, vol. 453, no. 7198, pp. 1102–1106, 2008.

[6] G. B. Keller and R. H. Hahnloser, “Neural processing of auditory feedback during vocal practice in a songbird,” *Nature*, vol. 457, no. 7226, pp. 187–190, 2009.

[7] K. S. Walsh, D. P. McGovern, A. Clark, and R. G. O'Connell, "Evaluating the neurophysiological evidence for predictive processing as a model of perception," *Annals of the New York Academy of Sciences*, vol. 1464, no. 1, pp. 242–268, 2020.

[8] A. Nelson, D. M. Schneider, J. Takatoh, K. Sakurai, F. Wang, and R. Mooney, "A circuit for motor cortical modulation of auditory cortical activity," *Journal of Neuroscience*, vol. 33, no. 36, pp. 14342–14353, 2013.

[9] D. M. Schneider, A. Nelson, and R. Mooney, "A synaptic and circuit basis for corollary discharge in the auditory cortex," *Nature*, vol. 513, no. 7517, pp. 189–194, 2014.

[10] B. P. Rummell, J. L. Klee, and T. Sigurdsson, "Attenuation of responses to self-generated sounds in auditory cortical neurons," *Journal of Neuroscience*, vol. 36, no. 47, pp. 12010–12026, 2016.

[11] D. M. Schneider, J. Sundararajan, and R. Mooney, "A cortical filter that learns to suppress the acoustic consequences of movement," *Nature*, vol. 561, no. 7723, pp. 391–395, 2018.

[12] N. J. Audette, W. Zhou, A. La Chioma, and D. M. Schneider, "Precise movement-based predictions in the mouse auditory cortex," *Current Biology*, vol. 32, no. 22, pp. 4925–4940, 2022.

[13] N. J. Audette and D. M. Schneider, "Stimulus-specific prediction error neurons in mouse auditory cortex," *Journal of Neuroscience*, vol. 43, no. 43, pp. 7119–7129, 2023.

[14] G. Iurilli, D. Ghezzi, U. Olcese, G. Lassi, C. Nazzaro, R. Tonini, V. Tucci, F. Benfenati, and P. Medini, "Sound-driven synaptic inhibition in primary visual cortex," *Neuron*, vol. 73, no. 4, pp. 814–828, 2012.

[15] L. A. Ibrahim, L. Mesik, X.-y. Ji, Q. Fang, H.-f. Li, Y.-t. Li, B. Zingg, L. I. Zhang, and H. W. Tao, "Cross-modality sharpening of visual cortical processing through layer-1-mediated inhibition and disinhibition," *Neuron*, vol. 89, no. 5, pp. 1031–1045, 2016.

[16] A. R. Garner and G. B. Keller, "A cortical circuit for audio-visual predictions," *Nature Neuroscience*, vol. 25, no. 1, pp. 98–105, 2022.

[17] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston, "Canonical microcircuits for predictive coding," *Neuron*, vol. 76, no. 4, pp. 695–711, 2012.

[18] S. M. O'Toole, H. K. Oyibo, and G. B. Keller, "Molecularly targetable cell types in mouse visual cortex have distinguishable prediction error responses," *Neuron*, 2023.

[19] R. P. Rao, "A sensory–motor theory of the neocortex," *Nature Neuroscience*, pp. 1–15, 2024.

[20] R. Jordan and G. B. Keller, "Opposing influence of top-down and bottom-up input on excitatory layer 2/3 neurons in mouse primary visual cortex," *Neuron*, vol. 108, no. 6, pp. 1194–1206, 2020.

[21] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation in speech production," *Science*, vol. 279, no. 5354, pp. 1213–1216, 1998.

[22] S. J. Blakemore, S. J. Goodbody, and D. M. Wolpert, "Predicting the consequences of our own actions: the role of sensorimotor context estimation," *Journal of Neuroscience*, vol. 18, no. 18, pp. 7511–7518, 1998.

[23] G. Bouvier, Y. Senzai, and M. Scanziani, "Head movements control the activity of primary visual cortex in a luminance-dependent manner," *Neuron*, vol. 108, no. 3, pp. 500–511, 2020.

[24] C. Büchel, S. Geuter, C. Sprenger, and F. Eippert, "Placebo analgesia: a predictive coding perspective," *Neuron*, vol. 81, no. 6, pp. 1223–1239, 2014.

[25] T. Woo, X. Liang, D. A. Evans, O. Fernandez, F. Kretschmer, S. Reiter, and G. Laurent, "The dynamics of pattern matching in camouflaging cuttlefish," *Nature*, pp. 1–7, 2023.

[26] N. Ulanovsky, L. Las, D. Farkas, and I. Nelken, "Multiple time scales of adaptation in auditory cortex neurons," *Journal of Neuroscience*, vol. 24, no. 46, pp. 10440–10453, 2004.

[27] I. Hershenhoren, N. Taaseh, F. M. Antunes, and I. Nelken, "Intracellular correlates of stimulus-specific adaptation," *Journal of Neuroscience*, vol. 34, no. 9, pp. 3303–3319, 2014.

[28] A. G. Enikolopov, L. Abbott, and N. B. Sawtell, "Internally generated predictions enhance neural and behavioral detection of sensory stimuli in an electric fish," *Neuron*, vol. 99, no. 1, pp. 135–146, 2018.

[29] S. Z. Muller, A. N. Zadina, L. Abbott, and N. B. Sawtell, "Continual learning in a multi-layer network of an electric fish," *Cell*, vol. 179, no. 6, pp. 1382–1392, 2019.

[30] H. Makino and T. Komiyama, "Learning enhances the relative impact of top-down processing in the visual cortex," *Nature Neuroscience*, vol. 18, no. 8, pp. 1116–1122, 2015.

[31] T. S. Yarden, A. Mizrahi, and I. Nelken, "Context-dependent inhibitory control of stimulus-specific adaptation," *Journal of Neuroscience*, vol. 42, no. 23, pp. 4629–4651, 2022.

[32] M. Boerlin, C. K. Machens, and S. Denève, "Predictive coding of dynamical variables in balanced spiking networks," *PLoS Computational Biology*, vol. 9, no. 11, p. e1003258, 2013.

[33] S. Denève and C. K. Machens, "Efficient codes and balanced networks," *Nature Neuroscience*, vol. 19, no. 3, pp. 375–382, 2016.

[34] J. Kadmon, J. Timcheck, and S. Ganguli, "Predictive coding in balanced neural networks with noise, chaos and delays," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16677–16688, 2020.

[35] L. Hertäg and H. Sprekeler, "Learning prediction error neurons in a canonical interneuron circuit," *Elife*, vol. 9, p. e57541, 2020.

[36] L. Hertäg and C. Clopath, "Prediction-error neurons in circuits with multiple neuron types: Formation, refinement, and functional implications," *Proceedings of the National Academy of Sciences*, vol. 119, no. 13, p. e2115699119, 2022.

[37] F. A. Mikulasch, L. Rudelt, and V. Priesemann, "Local dendritic balance enables learning of efficient representations in networks of spiking neurons," *Proceedings of the National Academy of Sciences*, vol. 118, no. 50, p. e2021925118, 2021.

[38] F. A. Mikulasch, L. Rudelt, M. Wibral, and V. Priesemann, "Where is the error? hierarchical predictive coding through dendritic error computation," *Trends in Neurosciences*, vol. 46, no. 1, pp. 45–59, 2023.

[39] Y. Song, B. Millidge, T. Salvatori, T. Lukasiewicz, Z. Xu, and R. Bogacz, "Inferring neural activity before plasticity as a foundation for learning beyond backpropagation," *Nature Neuroscience*, pp. 1–11, 2024.

[40] R. Hodson, M. Mehta, and R. Smith, "The empirical status of predictive coding and active inference," *Neuroscience & Biobehavioral Reviews*, p. 105473, 2023.

[41] E. J. Dennis, A. El Hady, A. Michaiel, A. Clemens, D. R. G. Tervo, J. Voigts, and S. R. Datta, "Systems neuroscience of natural behaviors in rodents," *Journal of Neuroscience*, vol. 41, no. 5, pp. 911–919, 2021.

[42] A. Wallach and N. B. Sawtell, "An internal model for canceling self-generated sensory input in freely behaving electric fish," *Neuron*, 2023.

[43] T. Keck, T. Toyoizumi, L. Chen, B. Doiron, D. E. Feldman, K. Fox, W. Gerstner, P. G. Haydon, M. Hübener, H.-K. Lee, *et al.*, "Integrating Hebbian and homeostatic plasticity: the current state of the field and future research directions," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1715, p. 20160158, 2017.

[44] G. B. Keller, T. Bonhoeffer, and M. Hübener, "Sensorimotor mismatch signals in primary visual cortex of the behaving mouse," *Neuron*, vol. 74, no. 5, pp. 809–815, 2012.

[45] H. Ko, S. B. Hofer, B. Pichler, K. A. Buchanan, P. J. Sjöström, and T. D. Mrsic-Flogel,

"Functional specificity of local synaptic connections in neocortical networks," *Nature*, vol. 473, no. 7345, pp. 87–91, 2011.

[46] L. Cossell, M. F. Iacaruso, D. R. Muir, R. Houlton, E. N. Sader, H. Ko, S. B. Hofer, and T. D. Mrsic-Flogel, "Functional organization of excitatory synaptic strength in primary visual cortex," *Nature*, vol. 518, no. 7539, pp. 399–403, 2015.

[47] S. El-Boustani, J. P. Ip, V. Breton-Provencher, G. W. Knott, H. Okuno, H. Bito, and M. Sur, "Locally coordinated synaptic plasticity of visual cortex neurons in vivo," *Science*, vol. 360, no. 6395, pp. 1349–1354, 2018.

[48] P. Zmarz and G. B. Keller, "Mismatch receptive fields in mouse visual cortex," *Neuron*, vol. 92, no. 4, pp. 766–772, 2016.

[49] K. Friston and S. Kiebel, "Predictive coding under the free-energy principle," *Philosophical transactions of the Royal Society B: Biological sciences*, vol. 364, no. 1521, pp. 1211–1221, 2009.

[50] K. Friston, "The free-energy principle: a unified brain theory?," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[51] Y. Ahmadian and K. D. Miller, "What is the dynamical regime of cerebral cortex?," *Neuron*, vol. 109, no. 21, pp. 3373–3391, 2021.

[52] M. Leinweber, D. R. Ward, J. M. Sobczak, A. Attinger, and G. B. Keller, "A sensorimotor circuit in mouse cortex for visual flow predictions," *Neuron*, vol. 95, no. 6, pp. 1420–1432, 2017.

[53] N. Gillis, *Nonnegative Matrix Factorization.* SIAM, 2020.

[54] T. Haga and T. Fukai, "Extended temporal association memory by modulations of inhibitory circuits," *Physical Review Letters*, vol. 123, no. 7, p. 078101, 2019.

[55] M. V. Srinivasan, S. B. Laughlin, and A. Dubs, "Predictive coding: a fresh view of inhibition in the retina," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 216, no. 1205, pp. 427–459, 1982.

[56] S. Furutachi, A. D. Franklin, T. D. Mrsic-Flogel, and S. B. Hofer, "Cooperative thalamocortical circuit mechanism for sensory prediction errors," *bioRxiv*, pp. 2023–07, 2023.

[57] J. Bolz and C. D. Gilbert, "Generation of end-inhibition in the visual cortex via interlaminar connections," *Nature*, vol. 320, no. 6060, pp. 362–365, 1986.

[58] A. Ayaz, A. Stäuble, M. Hamada, M.-A. Wulf, A. B. Saleem, and F. Helmchen, "Layer-

specific integration of locomotion and sensory information in mouse barrel cortex," *Nature communications*, vol. 10, no. 1, p. 2585, 2019.

[59] D. M. Schneider, "Reflections of action in sensory cortex," *Current Opinion in Neurobiology*, vol. 64, pp. 53–59, 2020.

[60] K. K. Clayton, R. S. Williamson, K. E. Hancock, G.-i. Tasaka, A. Mizrahi, T. A. Hackett, and D. B. Polley, "Auditory corticothalamic neurons are recruited by motor preparatory inputs," *Current Biology*, vol. 31, no. 2, pp. 310–321, 2021.

[61] R. J. Douglas and K. A. Martin, "Neuronal circuits of the neocortex," *Annual Reviews of Neuroscience*, vol. 27, no. 1, pp. 419–451, 2004.

[62] K. D. Harris and G. M. Shepherd, "The neocortical circuit: themes and variations," *Nature Neuroscience*, vol. 18, no. 2, pp. 170–181, 2015.

[63] M. W. Spratling, "Predictive coding as a model of biased competition in visual attention," *Vision Research*, vol. 48, no. 12, pp. 1391–1408, 2008.

[64] H. Ko, L. Cossell, C. Baragli, J. Antolik, C. Clopath, S. B. Hofer, and T. D. Mrsic-Flogel, "The emergence of functional microcircuits in visual cortex," *Nature*, vol. 496, no. 7443, pp. 96–100, 2013.

[65] B. Bathellier, L. Ushakova, and S. Rumpel, "Discrete neocortical dynamics predict behavioral categorization of sounds," *Neuron*, vol. 76, no. 2, pp. 435–449, 2012.

[66] O. Barak, "Recurrent neural networks as versatile tools of neuroscience research," *Current Opinion in Neurobiology*, vol. 46, pp. 1–6, 2017.

[67] U. Pereira-Obilinovic, J. Aljadeff, and N. Brunel, "Forgetting leads to chaos in attractor networks," *Physical Review X*, vol. 13, no. 1, p. 011009, 2023.

[68] B. Wang and J. Aljadeff, "Multiplicative shot-noise: A new route to stability of plastic networks," *Physical Review Letters*, vol. 129, no. 6, p. 068101, 2022.

[69] A. Ororbia, A. Mali, C. L. Giles, and D. Kifer, "Lifelong neural predictive coding: Learning cumulatively online without forgetting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5867–5881, 2022.

[70] V. Zhu and R. Rosenbaum, "Evaluating the extent to which homeostatic plasticity learns to compute prediction errors in unstructured neuronal networks," *Journal of Computational Neuroscience*, vol. 50, no. 3, pp. 357–373, 2022.

[71] R. Engelken, A. Ingrosso, R. Khajeh, S. Goedeke, and L. Abbott, "Input correlations impede

suppression of chaos and learning in balanced firing-rate networks," *PLoS Computational Biology*, vol. 18, no. 12, p. e1010590, 2022.

[72] J. S. Li, A. A. Sarma, T. J. Sejnowski, and J. C. Doyle, "Internal feedback in the cortical perception–action loop enables fast and accurate behavior," *Proceedings of the National Academy of Sciences*, vol. 120, no. 39, p. e2300445120, 2023.

[73] A. Finkelstein, K. Daie, M. Rózsa, R. Darshan, and K. Svoboda, "Connectivity underlying motor cortex activity during naturalistic goal-directed behavior," *bioRxiv*, pp. 2023–11, 2023.

[74] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, "The importance of mixed selectivity in complex cognitive tasks," *Nature*, vol. 497, no. 7451, pp. 585–590, 2013.

[75] S. Fusi, E. K. Miller, and M. Rigotti, "Why neurons mix: high dimensionality for higher cognition," *Current Opinion in Neurobiology*, vol. 37, pp. 66–74, 2016.

[76] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome, "Context-dependent computation by recurrent dynamics in prefrontal cortex," *Nature*, vol. 503, no. 7474, pp. 78–84, 2013.

[77] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27," *Open Review*, vol. 62, 2022.

[78] L. P. Jiang and R. P. Rao, "Predictive coding theories of cortical function," in *Oxford research encyclopedia of neuroscience*, 2022.

[79] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, vol. 9. World Scientific Publishing Company, 1987.

[80] R. Kuhn and S. Bos, "Statistical mechanics for neural networks with continuous-time dynamics," *Journal of Physics A: Mathematical and General*, vol. 26, no. 4, p. 831, 1993.