Optimal routing to cerebellum-like structures

Samuel Muscinelli[1], Mark Wagner[2], and Ashok Litwin-Kumar[1]

[1]*Mortimer B. Zuckerman Mind Brain Behavior Institute, Department of Neuroscience, Columbia University, New York, NY, United States of America*
[2]*National Institute of Neurological Disorders and Stroke, NIH, Bethesda, MD, United States of America*

Abstract

The vast expansion from mossy fibers to cerebellar granule cells produces a neural representation that supports functions including associative and internal model learning. This motif is shared by other cerebellum-like structures, including the insect mushroom body, electrosensory lobe of electric fish, and mammalian dorsal cochlear nucleus, and has inspired numerous theoretical models of its functional role. Less attention has been paid to structures immediately presynaptic to granule cell layers, whose architecture can be described as a "bottleneck" and whose functional role is not understood. We therefore develop a general theory of cerebellum-like structures in conjunction with their afferent pathways. This theory predicts the role of the pontine relay to cerebellar cortex and the glomerular organization of the insect antennal lobe. It also reconciles theories of nonlinear mixing with recent observations of correlated granule cell activity. More generally, it shows that structured compression followed by random expansion is an efficient architecture for flexible computation.

Introduction

In the cerebral cortex, multiple layers of densely connected, recurrent networks process input to form sensory representations. Theoretical models and studies of artificial neural networks have shown that such architectures are capable of extracting features from structured input spaces relevant for the production of complex behaviors [1]. However, the vertebrate cerebellum and cerebellum-like structures including the insect mushroom body, the electrosensory lobe of the electric fish, and the mammalian dorsal cochlear nucleus, operate with very different architectural principles [2]. In these areas, sensory and motor inputs are routed in a largely feedforward manner to a sparsely connected granule cell layer, whose neurons lack lateral recurrent interactions. These features suggest that such areas exploit a different strategy than the cerebral cortex to form their neural representations, despite being involved in many adaptive behaviors.

Many theories have focused on the computational role of the expanded granule cell representation in the cerebellum and cerebellum-like systems [3–7]. However, these theories have assumed a set of independent inputs, neglecting upstream areas that construct them. As we show, this assumption severely limits the learning performance of such systems for structured inputs. We hypothesized that this limitation is overcome by the specialized regions presynaptic to granule cell layers that process inputs to facilitate downstream learning. These regions have an architecture that can be described as a "bottleneck." In the mammalian cerebellum, inputs to granule cells originating from the cerebral cortex arrive primarily via the pontine nuclei in the brainstem, which compresses the cortical representation [8]. In the insect olfactory system, about 50 classes of olfactory projection neurons in the antennal lobe route input from thousands of olfactory sensory neurons to the roughly 2000 Kenyon cells in the mushroom body, the analogs of cerebellar granule cells. Other cerebellum-like structures exhibit a similar architecture [2].

Theoretical analyses have provided explanations for both the large expansion in number of granule cells and their small number of incoming connections (in-degree) without explicitly modeling the areas upstream of granule

cells [6, 7]. On the other hand, some of these upstream areas have been studied in isolation from the downstream granule cell layer. A number of studies have focused on the function of the insect antennal lobe, as well as the olfactory bulb, an analogous structure in mammals. Some have proposed that its main function is to de-noise olfactory sensory neuron signals [9], while others have argued for whitening the statistics of these responses [10, 11]. The pontine nuclei upstream of cerebellar granule cells have received less attention. Recent experiments suggest that the pontine nuclei do not only relay the cortical representation received from layer-5 pyramidal cells but instead integrate and reshape it [12].

Here, we use a combination of simulations, analytical calculations, and data analysis to develop a general theory of cerebellum-like structures and their afferent pathways. We propose that the bottleneck architecture of regions presynaptic to granule-like layers can be understood from the twofold goal of increasing dimensionality and minimizing noise. Our theory predicts that incoming weights to these regions should be tuned to the input statistics, and we show that this is particularly beneficial when input neurons are noisy and correlated. When applied to the insect olfactory system, our theory explains its glomerular organization and inter-glomerular interactions. The same objective, in the presence of distributed inputs from the motor cortex, implies that the pontine nuclei perform subspace selection. Furthermore, this analysis provides an explanation for recent observations of high correlations among granule cells [13]. More generally, our analysis reveals principles that relate the statistical properties of a neural representation to the architectures that optimally transform the representation to facilitate learning.

Results

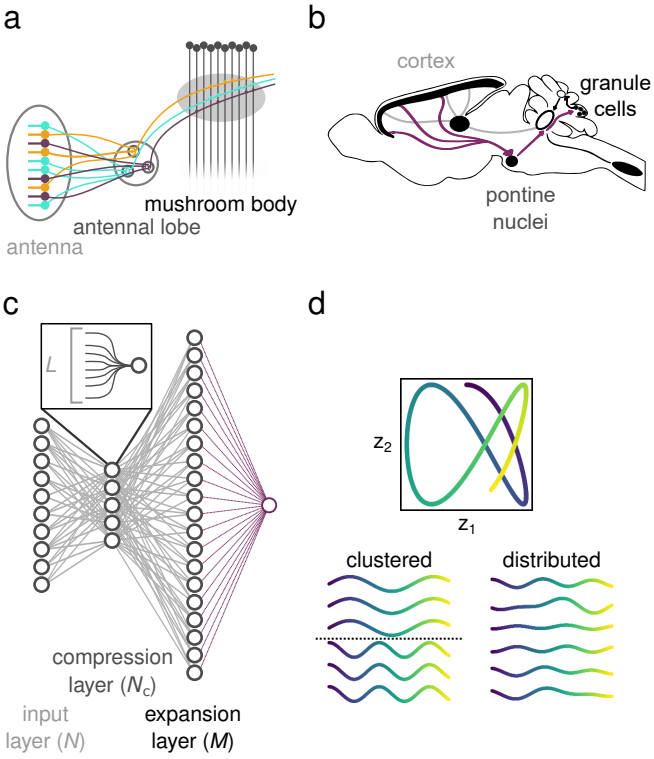

Figure 1: **Similar routing architecture to expanded representations.**
a: Schematic of the architecture of the insect olfactory system.
b: Schematic of the cortico-ponto-cerebellar pathway.
c: Schematic of the bottleneck model architecture. The input representation is compressed in the compression layer (N_c neurons). Each neuron in the compression layer receives L inputs, so that when $L = N$ the compression is fully-connected. From the compression layer, the representation undergoes a sparse, random and nonlinear expansion to the expansion layer (M neurons). Readout weights from the expansion layer are adjusted by synaptic plasticity to produce the appropriate output for a specified task.
d: Example of clustered and distributed input representations. A smooth trajectory in a two-dimensional task space (top) is embedded in an input representation of six neurons (bottom). Bottom left: examples of input neuron responses in a clustered representation (each row is a neuron). The dotted line separates the two clusters. Bottom right: examples of input neuron responses in a distributed representation.

The pathways to cerebellum-like structures, such as the mushroom body in the insect olfactory system (Fig. 1a) and the mammalian cerebellum itself (Fig. 1b) are characterized by an initial compression, in which the number of neurons is reduced, followed by an expansion. We model this "bottleneck" motif as a three-layer feedforward neural network (Fig. 1c). Information flows from N input layer neurons to M granule cells via a "compression layer" of N_c neurons, each of which samples L inputs. Since there are fewer compression layer neurons than inputs the compression ratio $N/N_c > 1$, while the expansion ratio $M/N_c > 1$, consistent with the large expansion to granule cells in cerebellum-like structures. In the insect olfactory system (Fig. 1a) tens of olfactory receptor neurons project to an individual glomerulus in the antennal lobe. In the cortico-cerebellar pathway (Fig. 1b) the compression ratio between cortico-pontine projection neurons and neurons in the pontine nuclei is estimated to be between two and

ten [8]. In contrast to neurons in expansion layers, which typically emit sparse bursts of action potentials, neurons in compression layers typically have higher firing rates. For this reason, in our model we consider either linear or rectified linear neurons for the compression layer, while for most of our results we use binary neurons to model the expansion layer.

What is the computational role of the compression layer? As a benchmark for comparing different architectures, we begin with the ability of a readout of the expansion layer representation (e.g. a Purkinje cell in the cerebellar cortex) to learn a categorization task in which input layer patterns are associated with positive or negative labels (which could represent positive and negative valences with which conditioned stimuli are associated [6, 7]). If the readout weights are learned using a supervised Hebbian rule, performance on this task increases with the dimension of the expansion layer representation and decreases with its noise strength [7,14]. We developed a theory describing how these quantities depend on properties of the compression and expansion layer connectivities.

In contrast to previous work [6,7,15], we do not assume that the input patterns are simply random and uncorrelated. Instead, we define the task as a mapping from patterns in a D-dimensional *task subspace* to the labels. The task subspace represents the portion of the input space where inputs relevant to the task will tend to lie. For example, in an odor classification task, olfactory sensory neurons of the same type respond similarly, thus defining a subspace in the space of all possible receptor firing rates. Hence, in our model, principal components analysis (PCA) performed on the input representation would reveal, in the absence of noise, D nonzero eigenvalues, with $D \ll N$. These eigenvalues decay (as a power law with exponent p; see section 3), so that our model exhibits representations similar to those in experiments, for example recordings of motor cortical activity that exhibit such decay [13, 16]. We consider two classes of input representations, each of which represents a different organization of selectivities of input neurons to the task variables. In a *clustered* representation, input neurons are organized in distinct groups, each of which is selective to a specific task variable (Fig. 1d, left). Such a representation can arise from a "labeled line" wiring organization and leads to high within-group correlations. In contrast, in a *distributed* representation, each neuron is tuned to different linear combinations of many task variables (Fig. 1d, right).

Selectivity to task-relevant dimensions determines learning performance

We compared the performance of a network without a compression layer, in which input layer neurons are randomly sampled by expansion layer neurons directly, to two networks with compression: one with random compression weights and one with learned compression weights. The learned compression weights are trained using error backpropagation [17], under the assumption that expansion weights are random and that readout weights are learned using Hebbian plasticity (see section 5). There is a substantial performance improvement from learning the compression weights, even though the subsequent expansion weights are fixed and random (Fig. 2a, b). However, the network with random compression performs worse than the network without a compression layer. These results suggest that compression can be highly beneficial, but only if the compression weights are appropriately tuned, and if they are not, compression instead degrades performance. To understand the principles underlying these effects, we developed a theory with which we will investigate the regimes in which compression is advantageous and explicitly specify the compression weights.

According to our theory, performance is increased when compression layer neurons are specifically tuned to the task-relevant principal components of the input representation. These components correspond, in a clustered representation, to groups of similarly tuned neurons, while in a distributed representation, to patterns of activity across the input layer. We represent the preferred stimulus of a compression layer neuron as a "tuning vector" in the N-dimensional space of input layer activity. When these vectors lie within the task subspace and are as different as possible across compression layer neurons, the compressed representation is denoised and decorrelated, improving the efficacy of the subsequent random expansion (see Methods). Indeed, if a tuning vector lies outside the task subspace, the corresponding compression layer neuron will be in part tuned to task-irrelevant activity, increasing noise (Fig. 2c, d). Furthermore, increased overlap of tuning vectors leads to correlation among compression layer neurons, reducing dimension (Fig. 2c, e).

In addition to alignment with task-relevant principal components, the magnitude, or gain, of compression layer neuron tuning affects dimension. If the activities of compression layer neurons are scaled so that sub-leading principal components are amplified, dimension is increased. If all principal components are equally strong, the

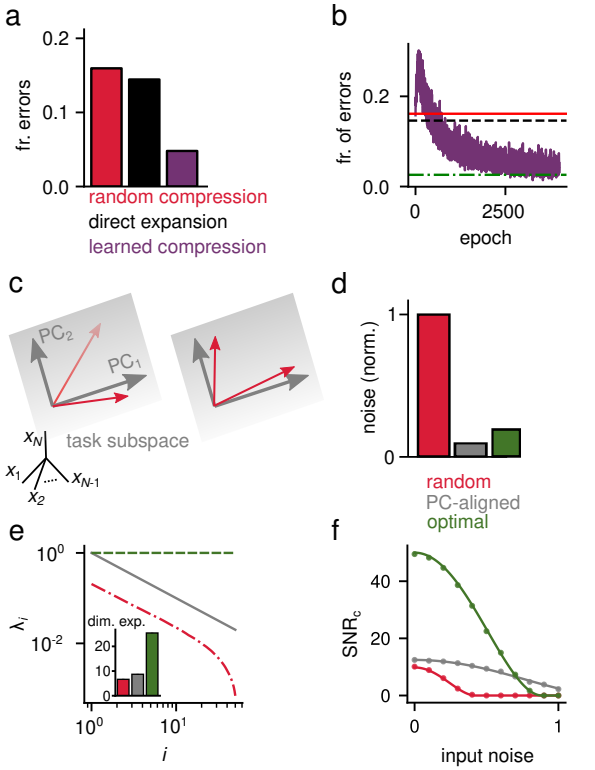

Figure 2: **Selectivity to task-relevant dimensions determines learning performance.**
a: Comparison of the performance of three networks: one in which only the readout weights are learned (random compression), one in which there is no compression and the input layer directly projects to the expansion layer via random weights (direct expansion), and one in which both the readout weights and compression weights are learned (learned compression).
b: Fraction of errors during training (purple), compared to other strategies not based on gradient descent. Color code is the same as in **a**.
c: Different geometrical arrangements of tuning vectors (red) with respect to the input principal components (grey arrows) and task subspace (grey plane), in the N-dimensional input space. Left: tuning vectors do not lie in the task subspace. Right: tuning vectors lie in the task subspace, but are not mutually orthogonal.
d: Noise strength at the compression layer, for random compression (red), optimal compression (green), and compression which leads to PC-aligned tuning vectors (gray). Noise is normalized to its strength at the input layer.
e: PCA eigenvalue spectrum at the compression layer and corresponding dimension expansion $\dim(m)/\dim(x)$ (inset).
f: Performance (SNR) of a Hebbian classifier on a random classification task, trained to read out from the *compression layer* representation, i.e. $\mathrm{SNR}_c = \dim(c)(1-\Delta_c)^2$. Dots indicates simulations results, lines indicate analytical results.

compression layer implements a whitening transformation, which results in the maximum expansion of dimension. However, we find a trade-off between maximizing dimension and denoising: amplification of sub-leading principal components also causes noise amplification (see section 11), and whitening ceases to be the best strategy above a certain noise intensity (Fig. 2f). We refer to the network that optimizes the trade-off between dimension and noise as an *optimal compression* network. Consistent with our theory, an optimal compression network's performance on a random classification task is a lower bound on the performance of a network whose compression weights are trained with backpropagation as described above (Fig. 2b). In the following sections, we describe the architecture of optimal compression networks for two cerebellum-like systems.

So far we assumed that the responses of the compression layer neurons are linear, meaning that the dimension of the compressed representation can be no larger than D. However, introducing a nonlinearity at the compression layer can increase the expansion layer dimension (Fig. S1a). Can two layers of nonlinear neuronal responses lead to improved performance? Surprisingly, in our setting with nonlinear compression followed by random nonlinear expansion, we find they cannot. While nonlinear compression layer neurons do indeed increase dimension, the additional noise introduced by the compression nonlinearity leads to an overall reduced performance. The fact that responses in the antennal lobe and pontine nuclei are substantially denser than those of Kenyon cells or granule cells is consistent with these neurons operating closer to a linear regime, in agreement with our analysis.

Optimal compression in the insect olfactory system

In the insect olfactory system, olfactory sensory neurons (OSNs) that express the same receptor project to the same olfactory glomerulus in the antennal lobe (Fig. 3a; [19]). Projections from the glomeruli are randomly mixed by Kenyon cells in the mushroom body [20, 21]. We hypothesized that evolutionary and developmental processes

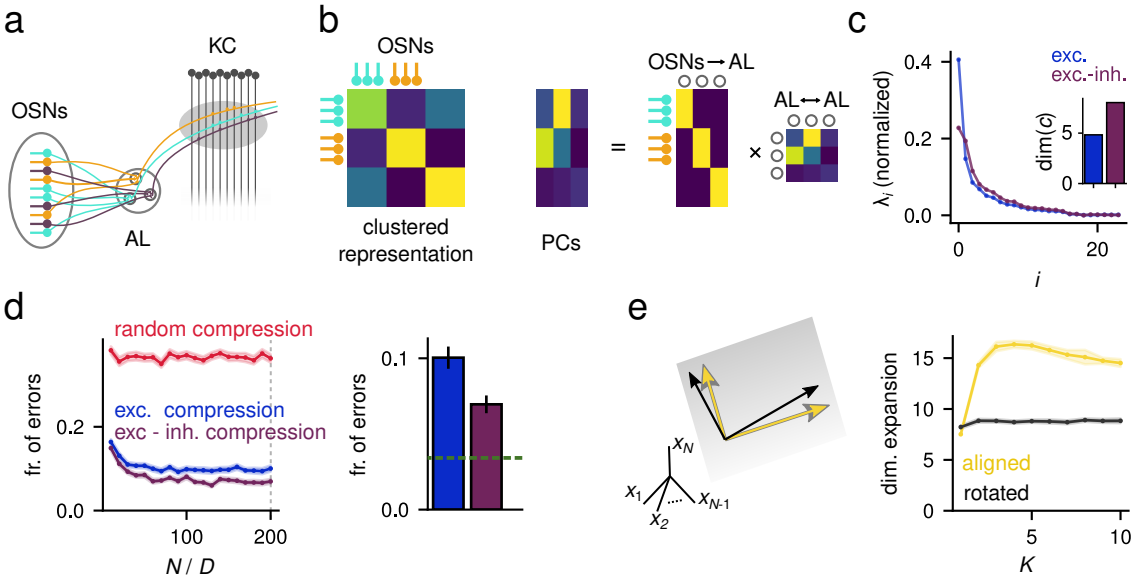

Figure 3: **Optimal compression in the insect olfactory system.**
a: Schematic of the insect olfactory system.
b: Left: Simplified representation of the olfactory receptor neuron covariance matrix. Rows and columns are ordered by receptor type. Center: Matrix whose columns are the principal components corresponding to the covariance matrix on the left, which corresponds to the transpose of the optimal compression matrix $\mathrm{G}^{\mathrm{opt}}$. Right: Factorization of the principal component matrix in the center panel as a block-diagonal rectangular matrix multiplying a square matrix. The two matrices have a natural biological interpretation as the OSN $\rightarrow$ AL connectivity and inter-glomeruli interactions, respectively.
c: Comparison of the compression layer eigenvalue spectrum when inputs are constructed using experimental recordings [18], with and without global inhibition (purple and blue, respectively). The inset shows the corresponding compression layer dimension.
d: Odor classification performance as a function of the input redundancy N/D (left). The bar plot on the right highlights the improvement due to recurrent inhibition (for fixed $N/D = 200$, see vertical dashed line on the left panel) and compares it to the benefit of optimal compression (green dashed line). Shaded region and error bars indicate standard error of the mean.
e: Left: Geometrical representation of tuning vectors which are aligned (yellow) versus not aligned (black) with principal components, corresponding to clustered and distributed compression layer representations, respectively. Right: Corresponding dimension expansion at the expansion layer plotted against the in-degree of expansion layer neurons K.

optimize the connectivity of the antennal lobe to facilitate a readout of the Kenyon cell representation.

The OSN representation of odors is clustered, since neurons expressing the same receptor have identical odor tuning. As a result, the covariance matrix of OSN responses is a block matrix, with strong within-block correlations (Fig. 3b, left). According to our theory, in an optimal compression network the compression weights are aligned with the principal components of this covariance matrix. When the responses of OSNs expressing different receptor types are uncorrelated, the compression layer neurons of such a network pool inputs from all OSNs of a particular type, consistent with the anatomical convergence of OSNs to antennal lobe glomeruli.

However, when there are correlations among the responses of different receptor types, the principal components have a block structure that is not purely block-diagonal (Fig. 3b, center). We show (see section 2) that, in this case, the optimal compression weights can be factored into a feedforward matrix that represents convergence of OSNs to antennal lobe glomeruli and a square matrix that represents recurrent inter-glomerular interactions (Fig. 3b, right)

$$\mathrm{G}^{\mathrm{opt}} = \left(\mathrm{I} - \mathrm{G}_{\mathrm{AL}}^{\mathrm{opt}}\right)^{-1} \mathrm{G}^{\mathrm{OSN}\rightarrow\mathrm{AL}} \quad , \tag{1}$$

where $\mathrm{G}^{\mathrm{opt}}$ is the optimal compression matrix, $\mathrm{G}_{\mathrm{AL}}^{\mathrm{opt}}$ is the matrix of inter-glomeruli effective connectivity that

would lead to optimal compression, while $G^{OSN\to AL}$ contains the OSN to AL compression connectivity.

We reanalyzed experimental recordings of single odor receptors to different odorants [18] to estimate the correlations among OSN types and found that they are more positive than expected by chance (Fig. S3). We show analytically that when these correlations are uniformly positive, global lateral inhibition across antennal lobe glomeruli is sufficient for optimal compression (see section 10). Consistent with this result and with studies that propose inter-glomerular interactions perform pattern decorrelation and normalization [10, 11, 22], global inhibition considerably increases the dimension of the antennal lobe representation when using the recorded responses as input to our model (Fig. 3c). This increase in dimension leads to improved performance in an odor classification task (Fig. 3d). However, we note that the network with global inhibition does not reach the performance of a network in which the antennal lobe representation is perfectly decorrelated, Fig. 3d, right), since specific lateral interactions among glomeruli further increase dimension when correlations are not uniformly positive (see Methods). Future studies should analyze whether the specific structure of lateral connectivity in the antennal lobe is consistent with this role.

In contrast to networks with specific convergence of OSN types onto glomeruli, a network in which OSNs are randomly mixed in the antennal lobe performs poorly (Fig. 3d, left). It may seem counterintuitive that such convergence is needed for optimal performance when antennal lobe responses are subsequently randomly mixed by Kenyon cells. Our theory illustrates that this difference is a consequence of both denoising and dimension. When OSN responses are noisy, pooling OSNs of the same type reduces noise by a factor N/D as compared to random compression (see section 11). Even in the absence of noise, the dimension of the compression layer is higher for clustered compression than for random compression. This arises from the random distortion of the input layer representation introduced by random compression weights. This distortion can only be avoided by ensuring weights onto compression layer neurons are orthogonal, a more stringent requirement that cannot be assured by independent random sampling of inputs (see section 9). In fact, if input to compression layer weights are constrained to be excitatory, the observed OSN convergence is the only possible weight matrix that has this property.

While the above benefits of convergence can be understood in terms of a linear mapping from input to compression layer, an additional and more subtle benefit of clustered compressed representations arises from the sparse, nonlinear expansion layer. Even when both representations are constructed with orthogonal compression weights, a sparse expansion leads to a higher dimension for a clustered representation than for a distributed one (Fig. 3e). This result arises from the ability of sparsely connected expansion layer neurons to select different principal components. If these neurons sample from a distributed representation, their input is always dominated by leading components and sub-leading ones are discarded, decreasing dimension. In contrast, if they sample only a few inputs from a clustered representation, some neurons mix sub-leading components, increasing dimension. In total, our theory reveals analytically that OSN convergence and inter-glomerular inhibition can be viewed as consequences of optimizing the Kenyon cell representation to facilitate downstream learning.

Optimal compression in the cortico-cerebellar pathway

In the cortico-ponto-cerebellar pathway, inputs from motor cortex are relayed to cerebellar granule cells via a compressed representation in the pontine nuclei. The motor cortex representation is low-dimensional and distributed (Fig. 4a; [23]). Its principal components thus contain both positive and negative elements. Since direct cortico-pontine projection are excitatory, effective input layer to compression layer weights could align with these components only if effectively inhibitory weights result from either di-synaptic feedforward inhibition or recurrent inhibition. However, unlike the antennal lobe, the pontine nuclei lack strong lateral inhibition. In rodents, lateral inhibition seems to be completely absent, while for primates and larger mammals, it appears to play only a limited role [8].

We therefore asked whether purely excitatory compression can approximate optimal compression for a distributed representation. We use gradient descent to train excitatory feedforward and inhibitory recurrent weights simultaneously, with the objective of maximizing dimension and minimizing noise at the compression layer (see section 5). Surprisingly, when the input representation is distributed and redundant ($N \gg D$), lateral inhibition does not improve the compressed representation (Fig. 4b, top). This is because, in this scenario, with high probability it is possible to find input neurons that encode each dimension of the task subspace. Projections from these neurons

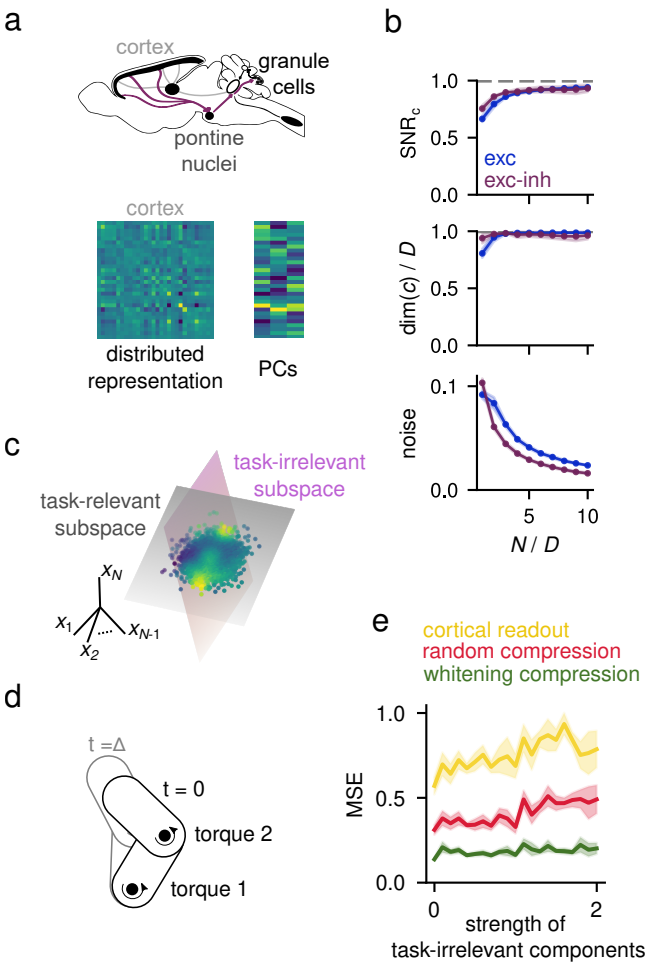

Figure 4: **Optimal compression in the cortico-cerebellar pathway.**
a: Top: Schematic of the cortico-ponto-cerebellar pathway. Bottom left: Example of covariance matrix for a distributed representation. Bottom right: Matrix of corresponding principal components (one column per component). Since principal components have positive and negative entries, optimal compression weights requires both excitatory and inhibitory connections.
b: Sign-constrained compression. Panels show the results of gradient descent training for a distributed input representation. The training objective is to maximize the SNR of a Hebbian classifier reading out from the compression layer, i.e. $\mathrm{SNR}_c = \dim(c)(1-\Delta_c)^2$. Top: SNR (training objective) as a function of the input redundancy N/D. The blue curve is for purely excitatory compression, while the purple curve is for simultaneous training of excitatory compression weights and recurrent inhibitory weights. Shading indicates standard deviation across input realizations and weight initializations. Middle: Same as top, but for dimension of the compression layer representation. Bottom: Same as top, but for noise strength Δ_c in the compression layer. Parameters: $D = 10$, input noise strength: 0.2 , $p = 1$, $N_c = 10$.
c: Illustration of a continuously varying target in the cortical representation. The high-dimensional cortical representation consists of orthogonal task-relevant and task-irrelevant subspaces. The cerebellum learns to map cortical activations (dots) to an output (color code) via a smooth nonlinear function.
d: Schematic of two-joint arm task. Given the state of the arm, defined by two joint angles and two angular velocities, and the torque applied to both joints, the system predicts the state of the arm at a time point Δ time units in the future.
e: Mean-squared error (MSE) of the model in predicting the future state of the arm, plotted against the strength of task-irrelevant activity. Network parameters: $M = 1000$, $N = 500$, $N_c = 100$, $L = 50$. Task parameters: $\Delta = 0.1s$

to the compression layer are sufficient to avoid a decrease in dimension (Fig. 4b, middle) and introducing lateral inhibition does not produce any additional benefit (Fig. 4b, bottom). Thus, even in the absence of lateral inhibition, excitatory cortico-pontine projections can be adjusted to maximize classification performance at the Purkinje cell readout. This result stands in contrast to our previous conclusion for systems with clustered input representations for which lateral inhibition is beneficial, such as the antennal lobe, and provides an explanation for this difference in architecture.

So far, our theory and results have focused on optimizing performance for a classification task. Some of the tasks that the cerebellum is involved in, such as eye-blink conditioning or timing tasks, may be reasonably interpreted in this way, but others may not be. An influential hypothesis of cerebellar function is that the cerebellum predicts the sensory consequences of motor commands, implementing a so-called forward model [24]. In this view, the cerebellum integrates representations of the current motor command and sensory state to estimate future sensory states (Fig. S6). We cast the problem of learning a forward model as a nonlinear regression task, assigning each point in the task subspace (representing the combination of motor command $\boldsymbol{u}$ and sensory state $\boldsymbol{s}$) a predicted sensory state $\boldsymbol{s}' = \boldsymbol{s} + f(\boldsymbol{s}, \boldsymbol{u})$ (Fig. 4c, see also 13). In this case, the goal of cerebellar Purkinje cells is to learn the

nonlinear function $f(\boldsymbol{s}, \boldsymbol{u})$. We consider a planar two link arm model, characterized by two joints at which torques can be applied (Fig. 4d). Optimal compression leads to substantially better performance than random compression when learning a forward model for this system, showing that the benefits we have described are not specific to discrete classification tasks (Fig. 4e).

Bottleneck architecture is more efficient than a single layer network

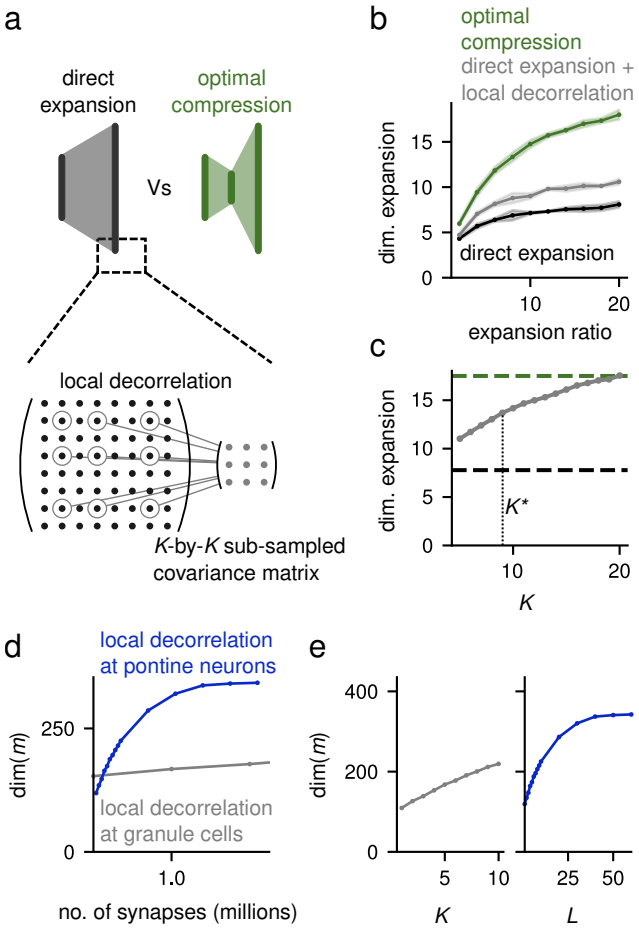

Figure 5: **Bottleneck architecture is more efficient than a single layer network.**
a: Direct expansion (left) and optimal compression (right) architectures. Inset: Illustration of local decorrelation at granule cells in the direct expansion architecture. Each expansion layer neuron only has access to a sub-sampled version of the full input covariance matrix (3-by-3 in the illustration). By local decorrelation, we mean that each expansion layer neuron can decorrelate this sub-sampled representation and nonlinearly mix the resulting representation.
b: Dimension expansion $\dim(m)/\dim(x)$ for different network architectures plotted against the expansion ratio. Local decorrelation yields only a small improvement.
c: Dimension expansion yielded by local decorrelation for increasing K. The total number of synapses quickly increases with K, and $K^\star$ (dotted line) indicates the value at which it becomes larger than the number of synapses needed with optimal compression.
d: Dimension of the expansion layer plotted against the total number of synapses used in the bottleneck architecture with local decorrelation at the compression layer (blue) and for a direct expansion architecture with local decorrelation at the expansion layer (grey).
e: Same as **d**, but plotted against the granule cell in-degree K for the direct expansion architecture (left) and against the pontine in-degree L for the bottleneck architecture (right).

Could we attain the benefits of optimal compression without requiring two layers of processing from input to expanded representation? Since optimal compression, as we have defined it, involves a linear transformation, it is indeed possible to generate an equivalent single-layer expansion by computing the product of the optimal compression and random expansion weights. What then is the advantage of performing these operations in two distinct steps? The answer is a consequence of the sparsity of the expansion layer weights.

For the expansion layer to implement both optimal compression and dimensional expansion, its neurons must be equipped with a local decorrelation mechanisms at their afferent synapses (Fig. 5a, inset). However, due to their sparse connectivity, each expansion layer neuron receives input only from a subset of input neurons. Minimizing correlations of this sub-sampled representation does not necessarily lead to decorrelation of the full representation. As a result, the benefit of adding local whitening to the single-layer expansion architecture is limited (Fig. 5b) if the in-degree of the expansion layer is small and neurons are not permitted to use non-local information (information about neurons to which they are not connected) in order to set synaptic weights.

If we increase the expansion layer in-degree in the model, local decorrelation better approximates optimal compression (Fig. 5c). However, the total number of synapses necessary to implement this architecture is larger than that required for optimal compression. The wiring cost of performing local decorrelation at the expansion layer is particularly high when considering parameters consistent with cerebellar cortex, i.e. $M \simeq 200,000$ and $N_c \simeq 7000$.

In this situation, it is much more efficient to perform local decorrelation at the cortico-pontine connections than at the granule cell layer (Fig. 5d,e). In summary, our results show that a dedicated compression layer provides an efficient implementation of this computation, both in terms of number of synapses and wiring complexity.

Biologically plausible learning of cortico-pontine compression

Activity in motor cortex is task-dependent and exhibits a steady turnover of the neurons representing a stable latent dynamics [25]. Unlike the genetically determined, clustered representation of olfactory sensory neurons, such activity may therefore have a covariance structure that changes over time. We therefore extend our theory from the case of fixed covariance to one that must be learned through experience-dependent synaptic plasticity.

Hebbian plasticity rules are a natural candidate for learning of compression weights, as they enable downstream neurons to extract the leading principal component of upstream population activity [26–28]. In many models, recurrent inhibitory interactions among downstream neurons are introduced to ensure that each neuron extracts a different principal component. Due to the lack of inhibition in the pontine nuclei, we asked whether sparsity of compression connectivity instead introduces the necessary diversity among pontine neuron afferents to achieve high task performance (Fig. 6a).

We find that the SNR of a Hebbian classifier trained on the expansion layer representation has a non-monotonic dependence on L, the in-degree of pontine neurons. The SNR is low for very small L, increases quickly, and finally decays slowly as L becomes very large (Fig. 6b, top). This behavior is a result of the trade-off between denoising and dimension. Noise strength at the expansion layer decays with L, as increased L permits more accurate estimation of leading principal components (Fig. 6b, center). On the other hand, dimension decreases with L, since as L increases, compression layer neurons estimate similar components (Fig. 6b, bottom). The value $L^\star$ that yields the best performance lies between ten and a hundred incoming inputs. $L^\star$ is only weakly affected by architectural parameters such as the number of input neurons N or expansion layer neurons M (Fig. S4c). Instead, it depends on features of the input representation, such as its dimension and the noise strength, with stronger noise and lower-dimensional representations favoring large in-degrees (Fig. 6c,d).

Thus, in the absence of recurrent inhibition in the compression layer, an intermediate in-degree leads to the best performance. Importantly, this is not only true for random classification tasks, but also for nonlinear regression (Fig. S5b). Furthermore, when L is set near its optimal value, performance for learning a forward model approaches that of an optimal compression network (Fig. 6f). Given the typically low-dimensional motor cortex representation, we predict that the optimal in-degree of rodent pontine nuclei neurons should be between ten and a hundred. To our knowledge, this in-degree has not been measured, but the large dendritic arbor of these neurons [29], suggests that it is at least much larger than the in-degree of granule cells.

We also tested whether further improvement could be achieved when we include recurrent inhibition using a recent model that implements a combination of Hebbian and anti-Hebbian plasticity rules (see 12, [28, 30]). After learning, the average principal component overlaps of sub-leading components decay more slowly than without recurrent inhibition (Fig. S4e), improving performance (Fig. S4f). As a consequence, we predict that mammals with more recurrent inhibition in the pontine nuclei may exhibit larger pontine in-degree.

One limitation of tuning the compression weights using Hebbian plasticity is that by construction it is an unsupervised method, meaning that it extracts leading principal components, but not necessarily only task-relevant ones. This might not present a problem when the input is only corrupted with independent random noise, since in this case the PCA spectrum is dominated by signal components (see 11.1). However, it will reduce the performance when leading components are task-irrelevant. Fortunately, the anatomy of the cortico-cerebellar system suggests a solution to this problem: in addition to cortical input, the pontine nuclei also receive feedback from the deep cerebellar nuclei (DCN), the output structure of the cerebellum [29] (Fig. 6e). Previous theories have largely ignored these connections. We provide a novel interpretation of this motif and suggest that it provides a supervisory signal that aids identification of task-relevant inputs. To test this hypothesis, we extended our model to include the target output as an external input to the pontine neurons that is used as a supervisory signal for synaptic plasticity, while otherwise leaving the network dynamics unchanged (see section 12.1). With this teaching signal, Hebbian plasticity is biased towards components of the input which correlate with the target and are therefore likely task-relevant. This mechanism leads to substantially improved performance when the dominant input components to the pontine

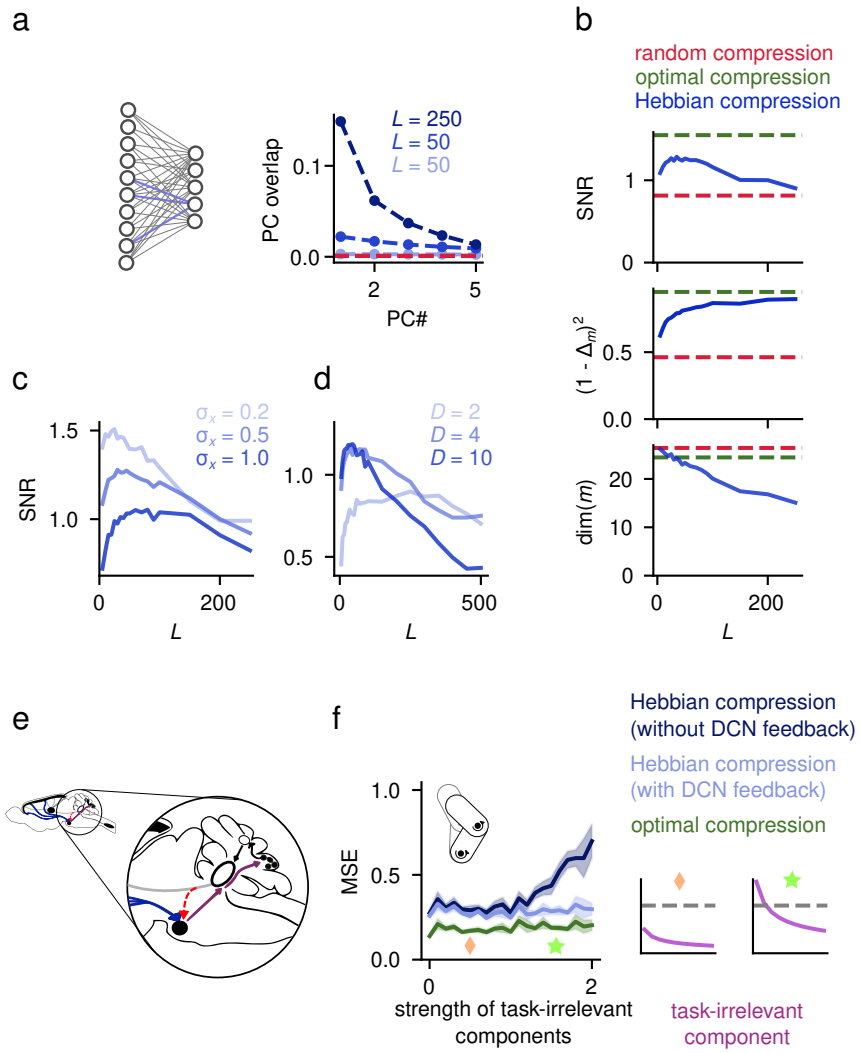

Figure 6: **Biologically plausible learned compression.**
a: Hebbian compression with sparse connectivity. Left: Incoming weights onto a post-synaptic neurons (blue) are learned independently from those onto other post-synaptic neurons, using a Hebbian rule. Right: Mean squared overlap between weight vector and principal components, after Hebbian learning. Overlaps are defined as $\mu_{ij} = \sum_{k=0}^{N} G_{ik} u_k^{(j)}$, where $\boldsymbol{u}^{(j)}$ is the j-th PC. Red dashed line indicates the overlaps obtained for random compression. Here and in subsequent plots, $p = 0.1$ and $D = 5$.
b: SNR (top), magnitude of de-noising (middle) and dimension (bottom) at the expansion layer representation for random, PCA and Hebbian compression, as a function of the compression layer in-degree L.
c: Effect of noise strength on SNR. Larger noise results in larger $L^\star$.
d: Effect of input dimension on SNR. In panels **b**, **c** and **d** we used $N = 1000$, $N_c = 500$, $M = 2000$, $D = 5$, $p = 0.1$ and $\sigma_x = 0.5$ unless otherwise stated.
e: Illustration of the DCN-pontine feedback (in red).
f: Feedback from DCN improves selection of the task-relevant dimensions. Mean-squared error (MSE) of the bottleneck network on the two-joint arm forward model, as in Fig. 4e. Inset: variance explained by task-irrelevant components (violet), in decreasing order, for two example values of the task-irrelevant component strength (green star and orange diamond). For comparison, grey dashed line indicates the variance explained by the leading task-relevant component. Network parameters: $M = 1000$, $N = 500$, $N_c = 100$, $L = 50$. Task parameters: $\Delta = 0.1s$

nuclei are task-irrelevant (Fig. 6f).

Hebbian compression explains correlation and selectivity of granule cells

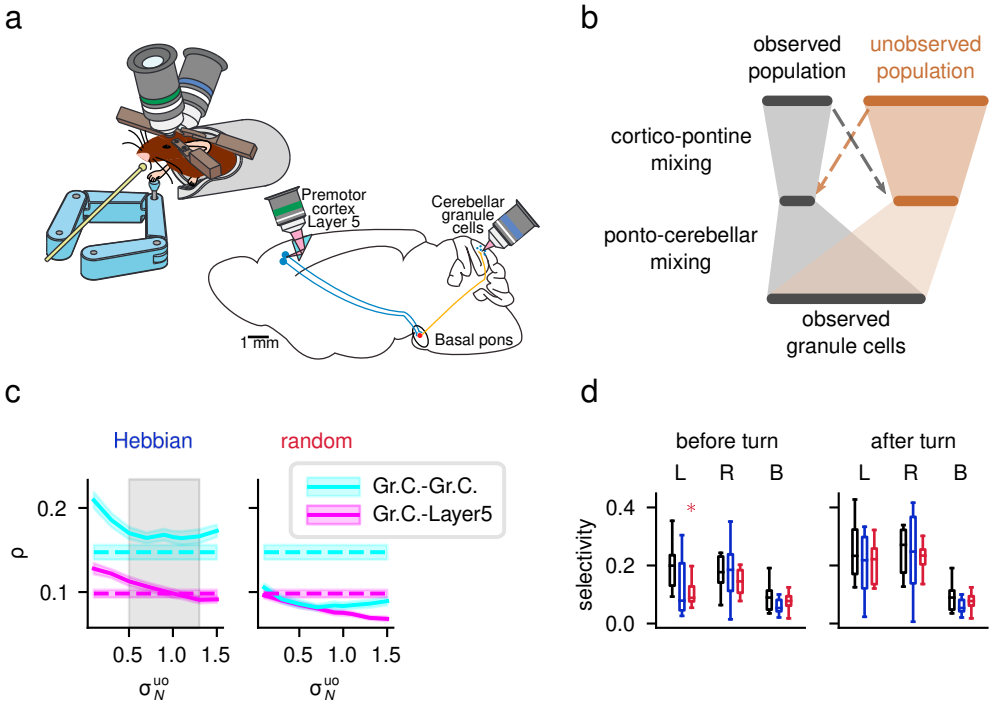

Figure 7: **Bottleneck model explains correlations and selectivity of recorded granule cells.**
a: Illustration of the experimental design of [13]. Mice performed a forelimb control task (left) while layer-5 pyramidal neurons and cerebellar granule cells were simultaneously recorded using two-photon calcium imaging (right).
b: Schematic illustrating how the bottleneck model is extended to reproduce the data. The dashed line indicates little or no mixing in the cortico-pontine pathway, while the shaded areas indicate strong mixing in the ponto-cerebellar pathway.
c: Layer 5-granule cell (magenta) and granule cell-granule cell (cyan) correlations, for Hebbian (left) and random (right) compression strategies. Mean correlations across neurons are averaged across animals and plotted against σ_N^{uo}, the noise strength in the unobserved population. Colored shaded area indicates standard error of the mean, computed across animals. Gray shaded area indicates the region in which correlations in the model are not statistically different from those in the data for both areas ($p > 0.05$, Wilcoxon signed-rank test). For this panel, the signal strength of the unobserved population is $\sigma_S^{uo} = 1$.
d: Average selectivity to left (L) and right (R) turns, or to turns without direction preference (B), for granule cells in the data (black) and models (Hebbian: blue; random: red). Selectivity is measured separately for the time-window before the turn (left) and after the turn (right). Boxes indicate 25th and 75th percentiles across mice, while whiskers indicate the full range of the mean selectivities across mice. Asterisks indicate cases in which model and data are not compatible, color coded according to the compression type. For panel **d**, $\sigma_S^{uo} = 1$ and $\sigma_N^{uo} = 0.5$ for the Hebbian model and $\sigma_N^{uo} = 0.1$ for the random model. For all panels, $N_c = N/2$, $M = 10N$, $N_{uo} = 2N$, $f = 0.1$, $L = 20$.

Recent recordings have shown that cerebellar granule cells in mice exhibit high selectivity to task variables and strong correlations in activity, both with each other and with cortical neurons [13]. Both findings are, at first sight, at odds with theories of random mixing [6,7,31] which predict that granule cells are decorrelated and encode nonlinear combinations of task variables. We show that optimal compression in the cortico-pontine pathway provides an alternative explanation for the experimental findings that preserves mixing in the granule cell layer.

We developed a model based on simultaneous two-photon calcium recordings of layer-5 pyramidal cells in motor cortex and cerebellar granule cells (Fig. 7a; [13]). During recording sessions, mice performed a skilled forelimb task (see 14) that required them to move a joystick in a L-shaped trajectory either to the left or to the right. We used recorded calcium traces of layer-5 pyramidal cells as inputs $\boldsymbol{x}$ to the cortico-ponto-cerebellar model described

above. In the model, cortico-pontine synapses undergo unsupervised learning via Hebbian plasticity, as detailed in the previous section. Similar to [13], we modeled unrecorded neurons by including an unobserved layer-5 population and associated pontine subpopulation. The latter projects to both the observed granule cell layer and other unobserved granule cells that we do not include in the model (Fig. 7b).

Since we do not have access to the unobserved population, we introduce two model parameters σ_S^{uo}, σ_N^{uo}, which control the strength of task-relevant (signal) and task-irrelevant (noise) components of the unobserved cortical population (see section 14). We systematically varied both parameters and measured average correlations in the model, both among granule cells and between granule and layer-5 cells. The model and data are compatible with respect to both measures, provided that task-irrelevant activity in the unobserved cortical population is strong enough (Fig. 7c, left). Notably, a model with random, non-plastic compression weights is not compatible with the data and exhibits lower correlations even for very small σ_N^{uo} (Fig. 7c, right).

We also quantified the selectivity of granule cell subpopulations responsive to left and right turns, or responsive to both turn directions, before and after the turn ([13], see section 14). We fix the parameters of the unobserved population to those yielding the best match of the granule cell correlations, with no further tuning. The model accounts for the selectivities observed in the data (Fig. 7d). A model without Hebbian compression can also explain the selectivity profile, but only if task-irrelevant activity in the unobserved population is assumed to be extremely weak.

Our analysis shows that the results of [13] are consistent with granule cell responding to mixtures of mossy fiber activity. Due to Hebbian plasticity, neurons in the pontine nuclei filter out task-irrelevant activity, hence becoming more selective to task variables than what would be expected from random compression and forming a lower-dimensional task representation. This decrease in dimension is not detrimental, but rather a consequence of discarding high-dimensional, task-irrelevant activity and preserving task-relevant activity. Since the latter is low-dimensional, random mixing at the granule cell layer yields only a moderate dimension expansion and high correlations. Altogether, these results show that for low-dimensional tasks, unsupervised learning at the cortico-pontine synapses can lead to a selective and highly correlated granule cell representation, despite mixing of distinct mossy fiber inputs.

Discussion

Our results demonstrate that specialized processing in "bottleneck" structures presynaptic to granule-like expansion layers substantially improves the quality of expanded representations. This two-stage architecture, with a structured bottleneck followed by a disordered expansion, is also more efficient, in terms of wiring cost, than a single-stage architecture. In the insect olfactory system, our theory shows that the glomerular organization of the antennal lobe is optimized to maximize the dimensionality of the Kenyon cell representation. In the cortico-cerebellar pathway, it shows that Hebbian learning of excitatory cortical inputs improves performance and provides a novel prediction for the in-degree of pontine neurons as well as the role of feedback from the deep cerebellar nuclei. It also accounts for the magnitude of correlations among simultaneously recorded granule cells and motor cortical neurons, arguing that for tasks with low-dimensional input representations, low-dimensional granule cell representations are optimal.

Other pathways to the cerebellum and other cerebellum-like structures

We focused on the cortico-ponto-cerebellar pathway and the insect olfactory system due to the ability to characterize the statistics of their inputs, but other cerebellar regions and cerebellum-like structures also exhibit bottleneck architectures. In addition to the cerebrocerebellum that is targeted by pontine inputs, the spinocerebellum is innervated by inputs from the spinal cord that integrate proprioceptive information from skeletal muscles and joints [32]. Characterizing the statistics of this ensemble of proprioceptive inputs to predict the optimal organization and connectivity of this pathway is an interesting direction for future research.

In electric fish, the nucleus praeeminentialis (PE) is a major source of input to the electrosensory lobe [2]. PE receives input from the midbrain and cerebellum, while sending output solely to the electrosensory lobe. Inter-

estingly, PE also receives inputs from the electrosensory lateral line lobe (ELL) neurons, a pathway analogous to DCN-pontine nuclei feedback connections. This suggests that our hypothesized supervisory role of DCN-pontine nuclei feedback could be an instance of a more general motif across cerebellum-like structures.

Response properties of compression layer neurons

While in previous work [13] pontine neurons have been modeled as binary, here we consider linear neurons, which we argue is more consistent with the graded firing rates they exhibit. Indeed, pontine neurons have higher firing rates and denser responses than cerebellar granule cells [12, 33, 34]. The latter are modeled here and in previous work as binary units, motivated by their tendency to reliably respond with sparse spikes or bursts to combinations of inputs [33]. Similar arguments are valid for the insect olfactory system, when comparing projection neurons, which have denser responses, to Kenyon cells [34].

We tested that our results are robust when pontine neurons are modeled using rectified linear units (ReLUs) and showed that the introduction of nonlinear responses at the level of the compression layer does not improve the performance of the bottleneck network. This reflects that a linear transformation is well-suited to maximize the performance of the subsequent random expansion. However, it is also possible that, for specific input statistics, nonlinear compression layer neurons lead to an improvement. Furthermore, non-random expansion architectures, such as deep networks, can benefit substantially from multiple nonlinear layers.

Feedforward and lateral inhibition

In most mammalian brain areas, long-range projections are predominantly excitatory. This is true of cortico-pontine projections from layer-5 pyramidal cells and the cholinergic projections of OSNs to the antennal lobe [35]. When the input representation is clustered into groups of neurons that exhibit high correlations within each cluster and are uncorrelated across clusters, we showed that convergence of projections from each cluster, such as the glomerular organization of the antennal lobe, is optimal. When correlations between clusters exist, we showed that either di-synaptic feedforward or lateral inhibition is necessary to maximize performance.

In the antennal lobe, both types of inhibition are present. However, di-synaptic inhibition is believed to largely mediate interactions among different glomeruli [10], suggesting that lateral inhibition dominates. We showed that global lateral inhibition is sufficient to effectively denoise and decorrelate OSNs whose response properties are constrained by experimental data [18]. It is an interesting future direction to investigate whether the pattern of correlations across specific pairs of glomeruli are reflected in lateral antennal lobe connections [36].

While inhibitory di-synaptic pathways to the pontine nuclei do exist [8,12], our results suggest that purely excitatory compression weights can perform near-optimally when the input representation is redundant and distributed, rather than clustered. However, we find that lateral inhibition might play a role in learning. Lateral inhibition promotes competition to ensure heterogeneous responses even when compression layer neurons share many inputs [28]. While lateral inhibition is almost absent from the pontine nuclei in rodents, its presence increases in larger mammals, such as cats and primates [8]. This suggests that in species where lateral inhibition is more abundant, pontine neurons may be more specifically tuned to task-relevant input dimensions and may exhibit larger in-degrees.

Cortico-pontine learning and topographical organization of the pontine nuclei

Our theory highlights the importance of plasticity at cortico-pontine synapses, thanks to which pontine neurons select task-relevant subspaces within the cortical input space. This could support, for example, distinguishing between movement preparation and execution [37] or compensating for representational drift in motor cortex [25]. We also showed that such subspace selection can be further improved by supervisory feedback from the DCN. Each pontine neuron *approximately* extracts task-relevant principal components, which requires a large number of pontine neurons to achieve good performance (Fig. S4a,b). This result provides a motivation for the smaller compression ratio observed in the pontine nuclei compared to the antennal lobe: when subspace selection is imperfect, it must

by compensated by a larger number of neurons in the compression layer. Another possibility is that part of the learning process that enables subspace selection is carried out by layer-5 pyramidal cells. These neurons receive a large number of inputs from other cortical neurons and therefore are capable of performing some of the computations we have described.

At a larger scale, the pontine nuclei exhibit a topographical organization, perhaps genetically encoded, that largely reflects the cortical organization [38, 39]. For example, motor cortical neurons responsible for different body parts project to distinct regions of the pontine nuclei. Moreover, there is evidence of convergence of motor and somatosensory cortical neurons coding for the same body part onto neighboring pontine regions [40]. It is therefore likely that both hard-coded connectivity and experience-dependent Hebbian plasticity control the compression statistics.

Random mixing and correlations in low-dimensional tasks

Our theory is consistent with data collected using simultaneous two-photon imaging from layer-5 pyramidal cells in motor cortex and cerebellar granule cells [13]. We showed that the level of correlations and selectivity of granule cells can be explained if cortico-pontine connections are tuned to task-relevant dimensions, but not if they are fixed and random. A previous theory proposed that the data could be accounted for by a model in which, during the course of learning, random mixing in the granule cell layer is reduced and for each granule cell a single mossy fiber input comes to dominate its response [13]. Our alternative preserves random mixing in granule cells and instead emphasizes the role of low-dimensional granule cell representations when animals are engaged in behaviors with low-dimensional structure. Such an interpretation may generally account for recordings of granule cells that exhibit low dimensionality and suggest the importance of complex behavioral tasks or multiple behaviors to probe the computations supported by these neurons [41].

Acknowledgements

We would like to thank Marjorie Xie, Adam Hantman, Britton Sauerbrei, Jonathan Kadmon, and Rick Warren for helpful discussions and comments. We would also like to thank Larry Abbott, Nate Sawtell, Manuel Beiran, and Kaushik Lakshminarasimhan for their comments on the manuscript. The M.J.W. laboratory is supported by the NINDS Intramural Research Program. A.L.-K. and S.M. were supported by the Gatsby Charitable Foundation, NSF award DBI-1707398, and the Simons Collaboration on the Global Brain. S.M. was also supported by the Swartz Foundation. A.L.-K. was also supported by the Burroughs Wellcome Foundation, the McKnight Endowment Fund, and NIH award R01EB029858.

References

[1] D. L. K. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature Neuroscience*, vol. 19, pp. 356–365, Mar. 2016.

[2] C. C. Bell, V. Han, and N. B. Sawtell, "Cerebellum-Like Structures and Their Implications for Cerebellar Function," *Annual Review of Neuroscience*, vol. 31, no. 1, pp. 1–24, 2008.

[3] D. Marr, "A theory of cerebellar cortex," *The Journal of Physiology*, vol. 202, no. 2, pp. 437–470, 1969.

[4] J. S. Albus, "A theory of cerebellar function," *Mathematical Biosciences*, vol. 10, pp. 25–61, Feb. 1971.

[5] M. Ito, "Neural design of the cerebellar motor control system," *Brain Research*, vol. 40, pp. 81–84, May 1972.

[6] B. Babadi and H. Sompolinsky, "Sparseness and Expansion in Sensory Representations," *Neuron*, vol. 83, pp. 1213–1226, Sept. 2014.

[7] A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, and L. F. Abbott, "Optimal Degrees of Synaptic Connectivity," *Neuron*, vol. 93, pp. 1153–1164.e7, Mar. 2017.

[8] P. Brodal and J. G. Bjaalie, "Organization of the pontine nuclei," *Neuroscience Research*, vol. 13, pp. 83–118, Mar. 1992.

[9] W. R. Chen and G. M. Shepherd, "The olfactory glomerulus: A cortical module with specific functions," *Journal of Neurocytology*, vol. 34, pp. 353–360, Sept. 2005.

[10] S. R. Olsen and R. I. Wilson, "Lateral presynaptic inhibition mediates gain control in an olfactory circuit," *Nature*, vol. 452, pp. 956–960, Apr. 2008.

[11] S. R. Olsen, V. Bhandawat, and R. I. Wilson, "Divisive Normalization in Olfactory Population Codes," *Neuron*, vol. 66, pp. 287–299, Apr. 2010.

[12] J.-Z. Guo, B. A. Sauerbrei, J. D. Cohen, M. Mischiati, A. R. Graves, F. Pisanello, K. M. Branson, and A. W. Hantman, "Disrupting cortico-cerebellar communication impairs dexterity," *eLife*, vol. 10, p. e65906, July 2021.

[13] M. J. Wagner, T. H. Kim, J. Kadmon, N. D. Nguyen, S. Ganguli, M. J. Schnitzer, and L. Luo, "Shared Cortex-Cerebellum Dynamics in the Execution and Learning of a Motor Task," *Cell*, vol. 177, pp. 669–682.e24, Apr. 2019.

[14] 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, pp. 585–590, May 2013.

[15] N. A. Cayco-Gajic, C. Clopath, and R. A. Silver, "Sparse synaptic connectivity is required for decorrelation and pattern separation in feedforward networks," *Nature Communications*, vol. 8, p. 1116, Oct. 2017.

[16] A. A. Russo, S. R. Bittner, S. M. Perkins, J. S. Seely, B. M. London, A. H. Lara, A. Miri, N. J. Marshall, A. Kohn, T. M. Jessell, L. F. Abbott, J. P. Cunningham, and M. M. Churchland, "Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response," *Neuron*, vol. 97, pp. 953–966.e8, Feb. 2018.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.

[18] E. A. Hallem and J. R. Carlson, "Coding of Odors by a Receptor Repertoire," *Cell*, vol. 125, pp. 143–160, Apr. 2006.

[19] L. B. Vosshall, A. M. Wong, and R. Axel, "An Olfactory Sensory Map in the Fly Brain," *Cell*, vol. 102, pp. 147–159, July 2000.

[20] S. J. C. Caron, V. Ruta, L. F. Abbott, and R. Axel, "Random convergence of olfactory inputs in the Drosophila mushroom body," *Nature*, vol. 497, pp. 113–117, May 2013.

[21] E. Gruntman and G. C. Turner, "Integration of the olfactory code across dendritic claws of single mushroom body neurons," *Nature Neuroscience*, vol. 16, pp. 1821–1829, Dec. 2013.

[22] R. W. Friedrich and M. T. Wiechert, "Neuronal circuits and computations: Pattern decorrelation in the olfactory bulb," *FEBS Letters*, vol. 588, pp. 2504–2513, Aug. 2014.

[23] K. V. Shenoy, M. Sahani, and M. M. Churchland, "Cortical Control of Arm Movements: A Dynamical Systems Perspective," *Annual Review of Neuroscience*, vol. 36, pp. 337–359, July 2013.

[24] D. M. Wolpert, R. C. Miall, and M. Kawato, "Internal models in the cerebellum," *Trends in Cognitive Sciences*, vol. 2, pp. 338–347, Sept. 1998.

[25] J. A. Gallego, M. G. Perich, R. H. Chowdhury, S. A. Solla, and L. E. Miller, "Long-term stability of cortical population dynamics underlying consistent behavior," *Nature Neuroscience*, vol. 23, pp. 260–270, Feb. 2020.

[26] E. Oja, "Simplified neuron model as a principal component analyzer," *Journal of Mathematical Biology*, vol. 15, pp. 267–273, Nov. 1982.

[27] P. Foldiak, "Adaptive network for optimal linear feature extraction," in *Proceedings of the IEEE/INNS International Joint Conference on Neural Networks*, 1989.

[28] C. Pehlevan and D. B. Chklovskii, "Optimization theory of Hebbian/anti-Hebbian networks for PCA and whitening," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1458–1465, Sept. 2015.

[29] C. Schwarz and P. Thier, "Binding of signals relevant for action: Towards a hypothesis of the functional role of the pontine nuclei," *Trends in Neurosciences*, vol. 22, pp. 443–451, Oct. 1999.

[30] C. Pehlevan, T. Hu, and D. B. Chklovskii, "A Hebbian/Anti-Hebbian Neural Network for Linear Subspace Learning: A Derivation from Multidimensional Scaling of Streaming Data," *Neural Computation*, vol. 27, pp. 1461–1495, July 2015.

[31] O. Barak, M. Rigotti, and S. Fusi, "The Sparseness of Mixed Selectivity Neurons Controls the Generalization–Discrimination Trade-Off," *Journal of Neuroscience*, vol. 33, pp. 3844–3856, Feb. 2013.

[32] R. Apps and M. Garwicz, "Anatomical and physiological foundations of cerebellar information processing," *Nature Reviews Neuroscience*, vol. 6, pp. 297–311, Apr. 2005.

[33] P. Chadderton, T. W. Margrie, and M. Häusser, "Integration of quanta in cerebellar granule cells during sensory processing," *Nature*, vol. 428, pp. 856–860, Apr. 2004.

[34] I. Ito, R. C.-y. Ong, B. Raman, and M. Stopfer, "Sparse odor representation and olfactory learning," *Nature Neuroscience*, vol. 11, pp. 1177–1184, Oct. 2008.

[35] H. Kazama and R. I. Wilson, "Origins of correlated activity in an olfactory circuit," *Nature Neuroscience*, vol. 12, pp. 1136–1144, Sept. 2009.

[36] N. M. Chapochnikov, C. Pehlevan, and D. B. Chklovskii, "Normative and mechanistic model of an adaptive circuit for efficient encoding and feature extraction," Sept. 2021.

[37] M. T. Kaufman, M. M. Churchland, S. I. Ryu, and K. V. Shenoy, "Cortical activity in the null space: Permitting preparation without movement," *Nature Neuroscience*, vol. 17, pp. 440–448, Mar. 2014.

[38] T. B. Leergaard and J. G. Bjaalie, "Topography of the complete corticopontine projection: From experiments to principal Maps," *Frontiers in Neuroscience*, vol. 1, pp. 211–223, Oct. 2007.

[39] C. F. Kratochwil, U. Maheshwari, and F. M. Rijli, "The Long Journey of Pontine Nuclei Neurons: From Rhombic Lip to Cortico-Ponto-Cerebellar Circuitry," *Frontiers in Neural Circuits*, vol. 11, 2017.

[40] G. A. Mihailoff, H. Lee, C. B. Watt, and R. Yates, "Projections to the basilar pontine nuclei from face sensory and motor regions of the cerebral cortex in the rat," *Journal of Comparative Neurology*, vol. 237, no. 2, pp. 251–263, 1985.

[41] F. Lanore, N. A. Cayco-Gajic, H. Gurnani, D. Coyle, and R. A. Silver, "Cerebellar granule cell axons support high-dimensional representations," *Nature Neuroscience*, vol. 24, pp. 1142–1150, Aug. 2021.

[42] L. F. Abbott, K. Rajan, and H. Sompolinsky, "Interactions between Intrinsic and Stimulus-Evoked Activity in Recurrent Neural Networks," *arXiv:0912.3832 [cond-mat, physics:physics, q-bio]*, Aug. 2010.

[43] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017.

[44] A. Fagg, N. Sitkoff, A. Barto, and J. Houk, "Cerebellar learning for control of a two-link arm in muscle space," in *Proceedings of International Conference on Robotics and Automation*, vol. 3, (Albuquerque, NM, USA), pp. 2638–2644, IEEE, 1997.

Methods

1 Measure of dimension

To quantify the dimension of a representation, we use a measure based on its covariance structure [7, 42]. For a representation $\boldsymbol{x}$ with covariance matrix C^x, which has eigenvalues $\lambda_1, \ldots, \lambda_n$, we define

$$\dim(x) := \frac{\mathrm{Tr}\left(\mathrm{C}^x\right)^2}{\mathrm{Tr}\left(\left(\mathrm{C}^x\right)^2\right)} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2} \quad . \tag{2}$$

In what follows, the problem of calculating the dimension will therefore be equivalent to the problem of calculating the trace of the covariance matrix and of its square.

2 Bottleneck network model

We model the input pathway to cerebellum-like structures as three-layer feedforward neural networks. The input layer representation $\boldsymbol{x}$ features the task subspace (see below) and task-irrelevant activity. The representation $\boldsymbol{x}$ is sent to the compression layer via a compression matrix $\mathrm{G} \in \mathcal{M}^{N_c \times N}$. We consider both linear compression layer units, for which $\boldsymbol{c} = \mathrm{G}\boldsymbol{x}$ and ReLU ones, for which $\boldsymbol{c} = [\mathrm{G}\boldsymbol{x} - \boldsymbol{\theta}]_+$, where the rectification is applied element-wise. The output of the compression layer is sent to the expansion layer via a matrix $\mathrm{J} \in \mathcal{M}^{M \times N_c}$, and we set $\boldsymbol{m} = \phi(\mathrm{J}\boldsymbol{c} - \boldsymbol{\theta})$, once again applied element-wise. In our result, ϕ is typically a Heaviside threshold function, except when considering nonlinear regression and when re-analyzing the data from [13], for which we used a ReLU nonlinearity. The nonzero entries of the expansion matrix J are always sampled independently and i.i.d. from $\mathcal{N}(0, 1/K)$. The thresholds $\boldsymbol{\theta}$ were chosen adaptively and independently for each unit so to obtain the desired coding level (fraction of active units) f or f_c [6], for the expansion and compression layer respectively. The expansion representation $\boldsymbol{m}$ is read out via readout weights $\boldsymbol{w}$, i.e. $y^\mu = \boldsymbol{w}^T(\boldsymbol{m} - f)$ which are set either using a Hebbian rule (Hebbian classifier) or via a pseudoinverse rule. The Hebbian readout weights are set as:

$$\boldsymbol{w} = \sum_{\mu=1}^{P} (\boldsymbol{m}^\mu - f) y^\mu \quad , \tag{3}$$

where y^μ are the target labels and the subtraction is intended element-wise, while the pseudoinverse readout weights are set as:

$$\boldsymbol{w} = \left(\mathrm{M}\mathrm{M}^T + \lambda \mathrm{I}\right)^{-1} \mathrm{M}\boldsymbol{y} \quad , \tag{4}$$

where λ is a regularization parameter.

Recurrent compression layer. We model recurrent interactions in the compression as

$$\tau_c \dot{\boldsymbol{c}} = -\boldsymbol{c} + \mathrm{W}_c \boldsymbol{c} + \mathrm{G}_{FF} \boldsymbol{x} \quad , \tag{5}$$

where W_c is the matrix of recurrent interactions in the compression layer and G_{FF} is the matrix of feedforward interactions from the input to the compression layer. If τ_c is much smaller than the timescale at which the input varies, we can assume stationarity, and we have that

$$\boldsymbol{c} = \mathrm{W}_c \boldsymbol{c} + \mathrm{G}_{FF} \boldsymbol{x} \Rightarrow \boldsymbol{c} = (I - \mathrm{W}_c)^{-1} \mathrm{G}_{FF} \boldsymbol{x} =: \mathrm{G}\boldsymbol{x} \quad , \tag{6}$$

where we defined the effective feedforward matrix as $\mathrm{G} := (I - \mathrm{W}_c)^{-1} \mathrm{G}_{FF}$. Therefore the compression matrix G can be thought as the effective stationary compression matrix in the presence of recurrent interactions and linear units. This factorization of of the compression matrix can be used to interpret optimal compression weights as a combination of feedforward and recurrent interactions. For example, for the antennal lobe we have

$$\mathrm{G} = \left(\mathrm{I} - \mathrm{G}^{\mathrm{AL}}\right)^{-1} \mathrm{G}^{\mathrm{ORNs} \to \mathrm{AL}} \quad , \tag{7}$$

where $\mathrm{G}^{\mathrm{ORNs} \to \mathrm{AL}}$ describes the projections from the antenna to the antennal lobe while G^{AL} captures the inter-glomeruli interactions.

3 Task subspace representation

The task subspace is defined by linearly embedding in the input layer a D-dimensional representation $\boldsymbol{z}$ via a random orthonormal matrix, i.e. $\boldsymbol{x} = \mathrm{A}\boldsymbol{z}$ where $\mathrm{A} \in \mathcal{M}^{N\times D}$ and the columns of A are orthonormal to each other, which is always possible, since we assume that $N > D$. The latent representation $\boldsymbol{z}$ consists of D-dimensional random Gaussian patterns, sampled from $\mathcal{N}(0, \Lambda)$, where Λ is a $D \times D$ diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_D$. The $\{\lambda_i\}$ represent the task subspace PCA eigenvalues, and to control their decay speed we set $\lambda_i = i^{-p}$ and vary the parameter p. Since the columns of A form an orthonormal set, we have that $\mathrm{A}^T\mathrm{A} = \mathrm{I}$, which implies $\dim(\boldsymbol{x}) = \dim(\boldsymbol{z})$. This can easily be proven for any representation $\boldsymbol{z}$ with covariance matrix C^z (even non-diagonal) using the cyclic permutation invariance of the trace:

$$\mathrm{Tr}\left(\mathrm{C}^x\right) = \mathrm{Tr}\left(\mathrm{A}\mathrm{C}^z\mathrm{A}^T\right) = \mathrm{Tr}\left(\mathrm{A}^T\mathrm{A}\mathrm{C}^z\right) = \mathrm{Tr}\left(\mathrm{C}^z\right) \quad , \tag{8}$$

and analogously for $\mathrm{Tr}\left(\left(\mathrm{C}^x\right)^2\right)$.

4 Random classification

The random classification task is defined by first assigning binary labels $y^\mu = \pm 1$ at random to patterns $\bar{\boldsymbol{z}}^\mu$, for $\mu = 1, \ldots, P$, in the task subspace. The network is required to learn these associations and generalize them to patterns which are corrupted by noise. When readout weights are learned using a Hebbian rule (Hebbian classifier). i.e. $\boldsymbol{w} = \sum_{\mu=1}^{P}(\boldsymbol{m}^\mu - f)y^\mu$, where the subtraction is intended element-wise, the probability of a classification error can be expressed in terms of signal-to-noise ratio (SNR) as $P(\text{error}) \simeq \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\mathrm{SNR}}{2}}\right)$ [6]. Previous work [7] has shown that the SNR can be expressed as

$$\mathrm{SNR} \simeq \frac{\dim(\boldsymbol{m})(1-\Delta_m)^2}{P} \quad , \tag{9}$$

where Δ_m is the noise strength at the expansion layer (see section 11), and $\dim(\boldsymbol{m})$ is the *noiseless* dimension, i.e. the dimension of the $\boldsymbol{m}$ representation in the absence of noise.

5 Optimization of compression weights via gradient descent

We used gradient descent to study the network performance for Fig. 2a,b, Fig. 4b, and Fig. S1c. We trained compression weights using backpropagation under the assumption that readout weights are learned using the Hebbian rule of 4. More precisely, for each epoch we sampled a random sparse expansion matrix J and sampled $B = 10$ random classification task, each consisting of $P = D = 50$ target patterns.

For Fig. 2a,b and Fig. S1c, Hebbian readout weights are set independently for each task, after which the compression weights are updated in the direction that decreases the loss (binary cross-entropy) computed on noise-corrupted test patterns. The update step was performed using the Adam optimizer [43], with a learning rate $\eta = 10^{-4}$. To facilitate learning by gradient descent, we replaced threshold nonlinearities in the expansion layer with ReLU nonlinearities. Adaptive thresholds were set, as in the rest of the paper, to obtain the desired coding level f. We used the same setup to test the performance in the presence of nonlinearities in the compression layer. We introduced ReLU nonlinearities in the compression layer in the same way as we did for the expansion layer, with a coding level $f_c = 0.3$. The other parameters were $p = 1$, $\sigma = 0.1$, $\sigma_M = 0.1$, $f = 0.1$, $N = 500$, $N_c = 250$, $M = 1000$ and $K = 4$.

For Fig. 4b, the compression weights were adjusted to maximize dimension and minimize noise at the compression layer, i.e. to maximize $\mathrm{SNR} = \dim(\boldsymbol{c})\,(1-\Delta_c)^2$. At every epoch, excitatory (inhibitory) weights were constrained to be non-negative (non-positive). Isotropic Gaussian noise was added of strength $\sigma = 0.2$ was added to the input. Other parameters: $N_c = D = 10$, $p = 1$.

6 Derivation of optimal compression

Our main theoretical results is that in the bottleneck network architecture, when expansion weights are random, optimal compression is obtained when compression layer units are tuned to the input principal components (PCs). This is equivalent to say that the weight vectors, which contain all the weights onto a compression layer neuron, should be aligned with the input PCs. Additionally, the amount of whitening that maximizes the performance depends on the noise strength.

To formally justify this result, we start from Eq. 9, which gives a proxy of the classification performance on a random classification task. To maximize the SNR, we would ideally maximize $\dim(\boldsymbol{m})$ while minimizing the noise Δ_m. We find that aligning the weights to the PCs favors both objectives, while performing whitening on top increases dimension but also noise. Below we present the set of evidence that leads to this conclusion. In most cases, looking at the effect of linear compression is sufficient, since 1) noise at the expansion layer is a monotonic function of noise at the compression layer (see below),and 2) dimension of the expansion layer depends on the dimension of the compression layer. However, dimension can also depend on the fine structure of the compression layer representation, in particular when the expansion connectivity is sparse (see below). Below we first motivate these statements and then study how the compression layer dimension and noise depend on the compression weights.

7 Dimension of the expansion layer

We define $\boldsymbol{h} := J\boldsymbol{c}$ i.e. the input currents to the expansion layer units. The $\boldsymbol{h}$ representation has covariance matrix C^h. However, to compute the dimension of the expansion layer it is not sufficient to compute the dimension of $\boldsymbol{h}$, because we need to take into account that expansion layer units have nonlinear responses. Analogously to Eq. 25, we can express the dimension of $\boldsymbol{m}$ as

$$\dim(\boldsymbol{m}) = \frac{M\langle C^m_{ii}\rangle^2}{\langle (C^m_{ii})^2\rangle + (M-1)\langle (C^m_{ij})^2\rangle} = \frac{Mf^2(1-f)^2}{f^2(1-f)^2 + (M-1)\langle (C^m_{ij})^2\rangle} \quad , \tag{10}$$

where the second equality follows from the expansion layer having a fixed coding level. Eq. 10 highlights that dimension is a function of the distribution of pairwise covariance among units. For fixed coding level f and threshold nonlinearities the covariance between two mixed-layer units $C^m_{ij} = \langle (m_i - f)(m_j - f)\rangle$ depends only on f and on the Pearson's correlation between the respective input currents ρ^h_{ij}:

$$C^m_{ij} = \sum_{n=1}^{\infty} \frac{1}{n!} \frac{2^{-n+1}}{2\pi} H_n\left(\mathrm{erfc}^{-1}(2f)\right)^2 e^{-2\left(\mathrm{erfc}^{-1}(2f)\right)^2} (\rho^h_{ij})^n \quad , \tag{11}$$

where $\rho^h_{ij} = \frac{C^h_{ij}}{\sqrt{C^h_{ii} C^h_{jj}}}$, erfc^{-1} is the inverse of the complementary error function and H_n is the n-th (physicist) Hermite polynomial. Therefore the problem of computing $\langle (C^m_{ij})^2\rangle$ can be rephrased, for threshold nonlinearities, to the problem of computing all the moments of ρ^h_{ij}. For typical regimes encountered in this paper, a truncation of Eq. 11 to the first few terms provides a good approximation of C^m_{ij}, providing a much more efficient way to compute C^m_{ij} compared to its integral formulation (see for example [7], Eq. 43).

To compute $\langle (C^m_{ij})^2\rangle$, we need to estimate the moments of ρ^h_{ij} (across units). If all the weights in the network are sampled from mean-zero distributions, ρ^h_{ij} will also have mean zero and its distribution will be symmetric around zero. As a result, we only need to compute the even moments of ρ^h_{ij}. When other approaches are not feasible, one can directly estimate the required moments of ρ^h_{ij} (typically the second and fourth moments are sufficient) from their empirical distribution for a particular network realization. In some cases, however, it is possible to obtain good approximations of ρ^h_{ij} by considering the Gaussian approximation of the joint distribution of C^h_{ii}, C^h_{jj} and C^h_{ij}, as we discuss below.

7.1 Fully-connected random expansion

For a fully connected random expansion with $J_{ij} \sim \mathcal{N}(0, 1/N_c)$, we find that the joint distribution of the elements of the input current covariance matrix, can be approximated by

$$\Pr\left(\mathrm{C}^h_{ii}, \mathrm{C}^h_{jj}, \mathrm{C}^h_{ij}\right) = \mathrm{PDF}\left(\mathcal{N}\left(\left(\mathrm{Tr}(\mathrm{C}^c)/N_c, \mathrm{Tr}(\mathrm{C}^c)/N_c\right), \Sigma\right)\right)\left(\mathrm{C}^h_{ii}, \mathrm{C}^h_{jj}\right) \cdot \mathrm{PDF}\left(\mathcal{N}\left(0, \mathrm{Tr}\left((\mathrm{C}^c)^2\right)/N_c^2\right)\right)\left(\mathrm{C}^h_{ij}\right) \quad , \tag{12}$$

i.e. the probability density factorizes in a two-dimensional normal distribution in C^h_{ii} and C^h_{jj} and a one-dimensional normal distribution in C^h_{ij}. The covariance matrix of the two-dimensional Gaussian is given by

$$\Sigma = \frac{1}{N_c^2}\begin{pmatrix} 2\sum_k (\mathrm{C}^c_{kk})^2 & -2\sum_{k\neq l}(\mathrm{C}^c_{kl})^2 \\ -2\sum_{k\neq l}(\mathrm{C}^c_{kl})^2 & 2\sum_k (\mathrm{C}^c_{kk})^2 \end{pmatrix} \quad . \tag{13}$$

To find the first two terms in the expansion of C^m_{ij}, we will find approximations of ρ^h_{ij} which are valid in the large N_c regime, as long as as long as $\dim(\boldsymbol{c})$ is not too small. Since C^h_{ij} is independent of the diagonal elements, we write

$$\langle(\rho^h_{ij})^2\rangle \simeq \langle(\mathrm{C}^h_{ij})^2\rangle \left\langle \frac{1}{\mathrm{C}^h_{ii}\mathrm{C}^h_{jj}} \right\rangle \quad . \tag{14}$$

We now expand the reciprocal above up to second order in $\dim(\boldsymbol{c})$, and get

$$\begin{aligned} \langle(\rho^h_{ij})^2\rangle &\simeq \frac{\langle(\mathrm{C}^h_{ij})^2\rangle}{\langle\mathrm{C}^h_{ii}\rangle^2}\left(1 - \frac{\Sigma_{12}}{\langle\mathrm{C}^h_{ii}\rangle^2}\right) \\ &= \frac{1}{\dim(\boldsymbol{c})}\left(1 + \frac{2}{\dim(\boldsymbol{c})} - \frac{2\sum_k(\mathrm{C}^c_{kk})^2}{\mathrm{Tr}^2\left((\mathrm{C}^c)^2\right)}\right) \end{aligned} \tag{15}$$

One can perform a similar calculation to compute $\langle(\rho^h_{ij})^4\rangle$, and find that, at the same order,

$$\langle(\rho^h_{ij})^4\rangle \simeq \frac{3}{\dim^2(\boldsymbol{c})} \quad . \tag{16}$$

Using these results and the first two coefficients of the series in Eq. 11, one finds

$$\begin{aligned} \langle(\mathrm{C}^m_{ij})^2\rangle =& \frac{1}{(2\pi)^2} H_1\left(\mathrm{erfc}^{-1}(2f)\right)^4 e^{-4\left(\mathrm{erfc}^{-1}(2f)\right)^2} \frac{1}{\dim(\boldsymbol{c})}\left(1 + \frac{2}{\dim(\boldsymbol{c})} - \frac{2\sum_k(\mathrm{C}^c_{kk})^2}{\mathrm{Tr}^2\left((\mathrm{C}^c)^2\right)}\right) \\ &+ \frac{1}{(8\pi)^2} H_2\left(\mathrm{erfc}^{-1}(2f)\right)^4 e^{-4\left(\mathrm{erfc}^{-1}(2f)\right)^2} \frac{3}{\dim^2(\boldsymbol{c})} \quad , \end{aligned} \tag{17}$$

from which one easily obtains the dimension. Notice that, at first order, $\langle(\mathrm{C}^m_{ij})^2\rangle$ depends only on $\dim(\boldsymbol{c})$ and as a result $\dim(\boldsymbol{m})$ scales approximately linearly with $\dim(\boldsymbol{c})$. Even though the above is an approximation, we find empirically that in most cases $\dim(\boldsymbol{c})$ is one of the strongest factors affecting $\dim(\boldsymbol{m})$.

7.2 Sparsely-connected random expansion

When considering a sparsely connected expansion, the probability density in Eq. 12 has the same form, but the statistics instead of depending on the full matrix C^c depend on the input covariance matrix "seen" by units i and j of the expansion layer, i.e. on C^c sub-sampled to the pre-synaptic partners of units i and j. To get an approximation of $\langle(\mathrm{C}^m_{ij})^2\rangle$, one has to compute the averages numerically in a Monte Carlo fashion. We find that for very sparse connectivity the degree at which the compression weights are aligned with PCs strongly affects the dimension of the expansion layer (Fig. 3). Since expansion connectivity in cerebellum-like structures is very sparse, we included the alignment with PCs as a feature of optimal compression.

8 Noise at the expansion layer

Here we show that the noise strength at the expansion layer is a monotonic function of noise at the compression layer (excluding additional noise sources at the expansion layer). Following previous work [6], the noise strength at the expansion layer Δ_m can be written, for threshold nonlinearities, as

$$\Delta_m = \frac{\sum_{i=1}^{M} \left[\Pr\left(h_i > \theta_i, \bar{h}_i < \theta_i\right) + \Pr\left(h_i < \theta_i, \bar{h}_i > \theta_i\right)\right]}{2Mf(1-f)} \quad , \tag{18}$$

where h and $\bar{h}$ are the noisy and noiseless input current to the expansion layer unit i, respectively. We treat h_i and $\bar{h}_i$ as correlated Gaussian variables with mean zero, variance σ_i and $\bar{\sigma}_i$ respectively and correlation coefficient ρ_i. We have that

$$\Pr\left(h_i < \theta_i, \bar{h}_i > \theta_i\right) = \int_{H^{-1}(f)}^{\infty} \mathcal{D}y H\left(\frac{\rho_i y - \frac{\bar{\sigma}_i}{\sigma_i} H^{-1}(f)}{\sqrt{1-\rho_i^2}}\right) \quad , \tag{19}$$

where $\mathcal{D}y$ indicates the standard Gaussian measure and $H(f)$ is the Gaussian tail probability, i.e. $H(f) := \int_f^{\infty} \mathcal{D}z$. Because of symmetry, we have that $\Pr\left(h_i < \theta_i, \bar{h}_i > \theta_i\right) = \Pr\left(h_i > \theta_i, \bar{h}_i < \theta_i\right)$, and

$$\Delta_m = \frac{1}{Mf(1-f)} \sum_i \int_{H^{-1}(f)}^{\infty} \mathcal{D}y H\left(\frac{\rho_i y - \frac{\bar{\sigma}_i}{\sigma_i} H^{-1}(f)}{\sqrt{1-\rho_i^2}}\right) \quad . \tag{20}$$

Under the approximation that all the expansion units are statistically equivalent, the sum over i can be simplified with the factor M at the denominator. Furthermore, the correlation coefficient ρ can be approximated by

$$\rho \simeq \frac{\sigma^2 + \bar{\sigma}^2}{2} - \Delta_c \quad . \tag{21}$$

In conclusion, noise strength at the expansion layer Δ_m can be approximated as a monotonic function of the noise strength at the compression layer Δ_c.

9 Dimension of the compression layer

Here we present analytical results on the dimension of the compression layer in the case of linear compression. Dimension is given by:

$$\dim(\boldsymbol{c}) = \frac{\mathrm{Tr}(\mathrm{C}^c)^2}{\mathrm{Tr}((\mathrm{C}^c)^2)} \tag{22}$$

In this case we can write $\boldsymbol{c} = \mathrm{G}\boldsymbol{x}$. Additionally, we can write the covariance matrix of $\boldsymbol{x}$ as a function of the covariance matrix of the task subspace representation $\mathrm{C}^x = \mathrm{A}\mathrm{C}^z\mathrm{A}^T$, and we can define $\tilde{\mathrm{G}} := \mathrm{G}\mathrm{A}$. Therefore, we need to compute

$$\mathrm{Tr}(\mathrm{C}^c) = \mathrm{Tr}(\mathrm{G}\mathrm{C}^x\mathrm{G}^T) = \mathrm{Tr}(\tilde{\mathrm{G}}\mathrm{C}^z\tilde{\mathrm{G}}^T) \tag{23}$$

$$\mathrm{Tr}((\mathrm{C}^c)^2) = \mathrm{Tr}(\mathrm{G}\mathrm{C}^x\mathrm{G}^T\mathrm{G}\mathrm{C}^x\mathrm{G}^T) = \mathrm{Tr}(\tilde{\mathrm{G}}\mathrm{C}^z\tilde{\mathrm{G}}^T\tilde{\mathrm{G}}\mathrm{C}^z\tilde{\mathrm{G}}^T) \quad . \tag{24}$$

9.1 Random compression

To obtain the *expected* dimension of $\boldsymbol{c}$ when the elements of G are chosen randomly as $G_{ij} \sim \mathcal{N}\left(0, \frac{1}{N}\right)$, we use the following expression for the dimension:

$$\dim(\boldsymbol{c}) = \frac{N_c \langle \mathrm{C}^c_{ii} \rangle^2}{\langle (\mathrm{C}^c_{ii})^2 \rangle + (N_c - 1)\langle (\mathrm{C}^c_{ij})^2 \rangle} \quad , \tag{25}$$

where the average is intended over i and j. Since the columns of A are orthonormal, the elements of $\tilde{\mathrm{G}}$ are also normally distributed and independent, with mean zero and variance $1/N$. We have that $\mathrm{C}^c_{ij} = \sum_{k=1}^{D} \tilde{G}_{ik}\tilde{G}_{jk}\lambda_k$.

Approximating the average over i and j with the average over the distribution of $\tilde{\mathrm{G}}$, we obtain the dimension of $\boldsymbol{c}$:

$$\begin{aligned}\dim(\boldsymbol{c}) &= \frac{\frac{N_c}{N^2}\mathrm{Tr}^2(\mathrm{C}^z)}{\frac{2}{N^2}\mathrm{Tr}((\mathrm{C}^z)^2)+\frac{1}{N^2}\mathrm{Tr}^2(\mathrm{C}^z)+\frac{N_c-1}{N^2}\mathrm{Tr}((\mathrm{C}^z)^2)} \\ &= \frac{\dim(\boldsymbol{z})}{1+\frac{\dim(\boldsymbol{z})+1}{N_c}} \quad .\end{aligned} \tag{26}$$

Notice that $\dim(\boldsymbol{x}) = \dim(\boldsymbol{z})$ since we assume an orthonormal embedding of the task subspace.

9.2 Orthonormal compression

Thanks to the cyclic property of the trace, computations of the traces in Eqs. 23, 24 above boils down to the computation of $\tilde{\mathrm{G}}^T\tilde{\mathrm{G}}$. In particular, if $\tilde{\mathrm{G}}^T\tilde{\mathrm{G}} = \mathrm{I}$, the traces will be unaffected by $\tilde{\mathrm{G}}$ and $\dim(\boldsymbol{c}) = \dim(\boldsymbol{z})$. One situation in which this happens when G satisfies two conditions: 1) the rows of G are orthonormal and 2) the columns of A are all contained in the span of the rows of G (see below for the proof). We call such compression "orthonormal" compression. Intuitively, the orthogonality of the rows of G avoids any distortion of the representation, while the columns of A needs to be in the span of the rows of G to avoid that parts of the task subspace are filtered out during compression.

Proof. We can express the columns of A as $\boldsymbol{a}_i = \sum_{k=1}^{N_c}\tilde{\mathrm{G}}_{ki}\boldsymbol{G}_k + \boldsymbol{a}_i^\perp$. If the rows of G are orthonormal, the orthonormality condition of the columns of A becomes:

$$\begin{aligned}\boldsymbol{a}_i^T\cdot\boldsymbol{a}_j &= \sum_{k=0}^{N_c}\tilde{\mathrm{G}}_{ki}\tilde{\mathrm{G}}_{kj} + (\boldsymbol{a}_i^\perp)^T\cdot\boldsymbol{a}_j^\perp = \delta_{ij} \\ \Rightarrow \tilde{\boldsymbol{g}}_i^T\cdot\tilde{\boldsymbol{g}}_j &= \delta_{ij} - (\boldsymbol{a}_i^\perp)^T\cdot\boldsymbol{a}_j^\perp \quad ,\end{aligned} \tag{27}$$

i.e. the columns of $\tilde{\mathrm{G}}$ are not always orthonormal, and therefore $\tilde{\mathrm{G}}^T\tilde{\mathrm{G}} \neq \mathrm{I}$ in general. A sufficient condition for $\tilde{\mathrm{G}}^T\tilde{\mathrm{G}} = \mathrm{I}$ is to have $\boldsymbol{a}_i^\perp = \boldsymbol{0}$ for all $i = 1,\dots,D$, i.e. to have that $\boldsymbol{a}_i \in \mathrm{span}(\boldsymbol{G}_1,\dots,\boldsymbol{G}_{N_c})$, for all $i = 1,\dots,D$. This concludes the proof.

While not relevant for the proof above, it is worth noticing that the rows of $\tilde{\mathrm{G}}$ are also not orthonormal in general. We separate the rows $\boldsymbol{G}_i \in \mathbb{R}^N$ of G in their component that lies in the task subspace and the one that is orthogonal to it

$$\boldsymbol{G}_i = \boldsymbol{G}_i^\parallel + \boldsymbol{G}_i^\perp = \sum_{k=1}^{D}(\boldsymbol{G}_i^T\cdot\boldsymbol{a}_k)\boldsymbol{a}_k + \boldsymbol{G}_i^\perp = \sum_{k=1}^{D}\tilde{\mathrm{G}}_{ik}\boldsymbol{a}_k + \boldsymbol{G}_i^\perp \quad , \tag{28}$$

where $\boldsymbol{a}_k$ is the k-th column of the matrix A and the last equality follows from the definition of $\tilde{\mathrm{G}}$. Since the rows of G are orthonormal, then we have that

$$\begin{aligned}\boldsymbol{G}_i^T\cdot\boldsymbol{G}_j &= \sum_{k=1}^{D}\tilde{\mathrm{G}}_{ik}\tilde{\mathrm{G}}_{jk} + (\boldsymbol{G}_i^\perp)^T\cdot(\boldsymbol{G}_j^\perp) = \delta_{ij} \\ \Rightarrow \tilde{\boldsymbol{G}}_i^T\cdot\tilde{\boldsymbol{G}}_j &= \delta_{ij} - (\boldsymbol{G}_i^\perp)^T\cdot\boldsymbol{G}_j^\perp \quad ,\end{aligned} \tag{29}$$

from which we conclude that the rows of $\tilde{\mathrm{G}}$ are also not orthonormal in general.

9.3 Whitening compression

We define the whitening compression matrix G as the compression matrix that yields $\mathrm{C}^c = \mathrm{I}$. Since $\mathrm{C}^c = \tilde{\mathrm{G}}\mathrm{C}^z\tilde{\mathrm{G}}^T$ and C^z is diagonal, $\tilde{\mathrm{G}}$ should also be diagonal, with diagonal elements equal to $\tilde{G}_{ii} = \lambda_i^{-1/2}$, where λ_i are the diagonal elements of C^z. The whitening compression matrix is therefore given by

$$\mathrm{G}_W = \mathrm{diag}(\lambda_1^{-1/2},\dots,\lambda_D^{-1/2})\mathrm{A}^T \quad . \tag{30}$$

Notice that G_W is defined up to a rotation. Indeed, if G_W is a whitening compression matrix, then $\mathrm{B}\mathrm{G}_W$ is also a whitening compression matrix if B is an orthonormal matrix.

10 Effect of recurrent connectivity in the compression layer

We showed above that recurrent connectivity in the compression layer can be treated, at stationarity, as an effective feedforward layer with effective compression matrix

$$\mathrm{G}_{\mathrm{eff}} = (\mathrm{I} - \mathrm{W}_c)^{-1}\mathrm{G} \quad . \tag{31}$$

Since we are typically interested in recurrent inhibition for which all entries of W_c are non-positive, we can also write $\mathrm{G}_{\mathrm{eff}} = (\mathrm{I} + |\mathrm{W}_c|)^{-1}\mathrm{G}$, where the absolute value is taken element-wise. The fixed point of the dynamics is stable only if all the eigenvalues of the Jacobian are negative. Since

$$\mathrm{Jac}(c) = \mathrm{W}_c - \mathrm{I} \quad , \tag{32}$$

the condition on the eigenvalues λ^c of W_c is

$$\lambda_i^c < 1 \quad , \quad \forall i \leq N_c \quad . \tag{33}$$

Using recurrent weights to adjust the input eigenvalue spectrum. We call C_0^c the covariance matrix of the feedforward input, i.e. $\mathrm{C}_0^c = \mathrm{G}\mathrm{C}^x\mathrm{G}^T$. Since C_0^c is a covariance matrix, it is diagonalized by a set of orthonormal eigenvectors, i.e. $\mathrm{C}_0^c = \mathrm{U}\Lambda_0\mathrm{U}^T$, where the columns of U are the eigenvectors. Suppose we want to find W_c such that the covariance of the compressed representation has some desired eigenvalue spectrum $\{\bar{\lambda}_i\}$. If we do not aim to change the eigenvectors, we can write the desired covariance matrix as $\bar{\mathrm{C}} = \mathrm{U}\bar{\Lambda}\mathrm{U}^T$ and we can make the ansatz $\mathrm{W}_c = \mathrm{U}\Lambda_c\mathrm{U}^T$, so that we need to solve

$$\mathrm{U}\Lambda_c\bar{\Lambda}\Lambda_c\mathrm{U}^T - 2\mathrm{U}\Lambda_c\bar{\Lambda}\mathrm{U}^T + \mathrm{U}\bar{\Lambda}\mathrm{U}^T - \mathrm{U}\Lambda_0\mathrm{U}^T = 0 \quad . \tag{34}$$

Since all the matrices share the same eigenvectors, we need to only solve the following equation

$$\Lambda_c\bar{\Lambda}\Lambda_c - 2\Lambda_c\bar{\Lambda} + \bar{\Lambda} - \Lambda_0 = 0 \quad , \tag{35}$$

which can be solved for each eigenvalue of W_c separately, yielding

$$\lambda_i^c = 1 \pm \sqrt{\frac{\lambda_i^0}{\bar{\lambda}_i}} \quad . \tag{36}$$

Notice that of these two solutions only $\lambda_i^c = 1 - \sqrt{\frac{\lambda_i^0}{\bar{\lambda}_i}}$ corresponds to a stable fixed point.

Recurrent inhibition. In the previous section we assumed that the recurrent matrix is symmetric and that it shares the same eigenvectors as the input covariance matrix. This choice implies that recurrent connections might be both positive and negative, i.e. both excitatory and inhibitory. Here, we study the biologically relevant case of purely inhibitory recurrent connectivity.

We start by considering global inhibition, characterized by a rank-one connectivity matrix that can be written as

$$W_c = \frac{g_I}{N}\mathbf{1}\mathbf{1}^T \quad , \tag{37}$$

where $\mathbf{1}$ is the vector of all ones. In this case, the inverse of $\mathrm{I} - \mathrm{W}_c$ can be computed explicitly using the Sherman–Morrison formula:

$$\left(\mathrm{I} + \frac{g_I}{N}\mathrm{W}_c\right)^{-1} = \mathrm{I} - \frac{g_I}{N}\frac{\mathbf{1}\mathbf{1}^T}{1+g_I} \quad . \tag{38}$$

Plugging in this expression in the definition of C^c, we get

$$\mathrm{C}^c = \mathrm{C}_0^c - \frac{g_I}{N}\frac{1}{1+g_I}\left(\mathbf{1}\mathbf{1}^T\mathrm{C}_0^c + \mathrm{C}_0^c\mathbf{1}\mathbf{1}^T\right) + \frac{g_I^2}{N^2}\frac{1}{(1+g_I)^2}\mathbf{1}\mathbf{1}^T\mathrm{C}_0^c\mathbf{1}\mathbf{1}^T \quad . \tag{39}$$

We now define $\boldsymbol{u}^1 := \mathrm{U}^T\mathbf{1}$, i.e. the vector of the projections of the eigenvectors of C_0^c on the constant mode $\mathbf{1}$. We get

$$\mathrm{C}^c = \mathrm{C}_0^c - \frac{g_I}{N}\frac{1}{1+g_I}\left(\mathbf{1}(\boldsymbol{u}^1)^T\Lambda_0\mathrm{U}^T + \mathrm{U}\Lambda_0\boldsymbol{u}^1\mathbf{1}^T\right) + \frac{g_I^2}{N^2}\frac{1}{(1+g_I)^2}\mathbf{1}(\boldsymbol{u}^1)^T\Lambda_0\boldsymbol{u}^1\mathbf{1}^T \quad . \tag{40}$$

If $\mathbf{1}$ is one of the eigenvectors of C_0^c, then $\boldsymbol{u}^1$ only has one nonzero entry. In this case, global inhibition controls the strength of the uniform mode in C^c, and can be used to set it to zero. More generally, Eq. 40 shows that global inhibition acts on the projection of input modes on the constant mode. This is equivalent to say that the effect of global inhibition on a certain mode depends on the mode mean, i.e. the average of the eigenvector entries. It is straightforward to generalize this derivation to the case of multiple inhibitory neurons which act on non-overlapping groups of neurons.

11 Noise at the compression layer

To quantify noise strength, we consider the Euclidean distance between noisy patterns and noiseless patterns, e.g. for the input layer $d(\boldsymbol{x}, \bar{\boldsymbol{x}}) = \sum_{i=1}^N (x_i - \bar{x}_i)^2$. To obtain a meaningful metric, this distance should be averaged over the noise distribution. Additionally, we normalize the average distance by the average distance among two noiseless input patterns following the input distribution. We obtain the following metric:

$$\Delta_x = \frac{\langle d(\boldsymbol{x}^\mu, \bar{\boldsymbol{x}}^\mu)\rangle_{\mu,\xi}}{\langle d(\bar{\boldsymbol{x}}^\mu, \bar{\boldsymbol{x}}^\nu)\rangle_\mu} \quad , \tag{41}$$

where by the subscript μ we denote the average over the input distribution and by the subscript ξ the average over the noise distribution. With this normalization, $\Delta_x = 1$ if noisy patterns are on average as distant from their noiseless version as two different input patterns are with respect to each other.

11.1 Effect of compression on additive noise

Consider the scenario the input layer representation is corrupted by additive noise, i.e.

$$x_i = \bar{x}_i + \xi_i \quad , \tag{42}$$

where $\boldsymbol{\xi}$ is a random vector with mean zero and covariance matrix C^ξ. For this representation, the numerator in Eq. 41 is given by $\langle d(\boldsymbol{x}^\mu, \bar{\boldsymbol{x}}^\mu)\rangle_{\mu,\xi} = \mathrm{Tr}(\mathrm{C}^\xi)$. To compute the denominator, we notice that

$$\langle d(\bar{\boldsymbol{x}}^\mu, \bar{\boldsymbol{x}}^\nu)\rangle_\mu = 2\sum_{i=1}^N \langle (\bar{x}_i^\mu)^2\rangle_\mu = 2\mathrm{Tr}\left(\mathrm{C}^{\bar{x}}\right) \tag{43}$$

by definition of the covariance matrix. Using Eq. 43, together with the cyclic permutation invariance of the trace, we obtain that

$$\Delta_x = \frac{D}{2N}\frac{\mathrm{Tr}(\mathrm{C}^\xi)}{\mathrm{Tr}\left(\mathrm{C}^z\right)} \quad . \tag{44}$$

For isotropic noise, $\mathrm{C}^\xi = \sigma^2\mathrm{I}$, therefore

$$\Delta_x = \frac{D\sigma^2}{2\mathrm{Tr}\left(\mathrm{C}^z\right)} \quad . \tag{45}$$

Using the same approach we can compute the noise strength at the compression layer representation $\boldsymbol{c}$. By direct calculation, one can show that

$$\langle d(\boldsymbol{c}^\mu, \bar{\boldsymbol{c}}^\mu)\rangle_{\mu,\xi} = \mathrm{Tr}\left(\mathrm{G}\mathrm{C}^\xi\mathrm{G}^T\right) \quad . \tag{46}$$

Similarly, one can show that

$$\langle d(\bar{\boldsymbol{c}}^\mu, \bar{\boldsymbol{c}}^\nu)\rangle_\mu = 2\frac{N}{D}\mathrm{Tr}\left(\mathrm{C}^c\right) = 2\frac{N}{D}\mathrm{Tr}\left(\mathrm{G}\mathrm{A}\mathrm{C}^z\mathrm{A}^T\mathrm{G}^T\right) \quad . \tag{47}$$

Notice that the factor $\frac{N}{D}$ results from the fact that we want order 1 input layer activity. Introducing the overlap matrix $\mu := \mathrm{GA}$, we can write

$$\Delta_c = \frac{D\mathrm{Tr}\left(\mathrm{G}\mathrm{C}^{\xi}\mathrm{G}^T\right)}{2N\mathrm{Tr}\left(\mu\mathrm{C}^z\mu^T\right)} \quad . \tag{48}$$

If noise is isotropic, we have

$$\Delta_c = \frac{\sigma^2 D\mathrm{Tr}\left(\mathrm{G}\mathrm{G}^T\right)}{2N\mathrm{Tr}\left(\mu\mathrm{C}^z\mu^T\right)} \quad . \tag{49}$$

PC-aligned compression Since we use an orthonormal embedding of the task subspace into the input space, when compression weights are aligned to the input principal components (PCs), we obtain again the latent representation $\boldsymbol{z}$ in the compression layer. The compression weights are set as $\mathrm{G}^{\mathrm{PCA}} = \sqrt{\frac{D}{N}}\mathrm{A}^T$. In presence of isotropic input noise, the noise strength at the compression layer corresponding to this kind of compression is

$$\Delta_c^{\mathrm{PCA}} = \frac{D}{N}\Delta_x \quad . \tag{50}$$

Notice that this result is left unchanged if $\mathrm{G}^{\mathrm{PCA}}$ is left-multiplied by an orthonormal matrix.

Whitening compression Whitening compression results in a isotropic compression layer representation, i.e. $\mathrm{C}^c = \mathrm{I}$. To achieve this, the compression weights are set as $\mathrm{G}^W = \sqrt{\frac{D}{N}}\mathrm{diag}\left(\lambda_1^{-1/2}, \ldots, \lambda_D^{-1/2}\right)\mathrm{A}^T$. In presence of isotropic input noise, the noise strength at the compression layer corresponding to this kind of compression is

$$\Delta_c^W = \frac{\sigma^2\mathrm{Tr}\left(\left(\mathrm{C}^z\right)^{-1}\right)}{2N} = \frac{\mathrm{Tr}\left(\mathrm{C}^z\right)\mathrm{Tr}\left(\left(\mathrm{C}^z\right)^{-1}\right)}{D^2}\Delta_c^{\mathrm{PCA}} \quad . \tag{51}$$

This can also be rewritten as

$$\Delta_c^W = \frac{\sigma^2}{2N}\sum_{i=1}^{D}\lambda_i^{-1} \quad , \tag{52}$$

When C^z has a decaying eigenvalue spectrum, $\Delta_c^W > \Delta_c^{\mathrm{PCA}}$, i.e. whitening leads to stronger noise than standard PCA compression.

12 Hebbian and anti-Hebbian rules for principal component extraction

In Figs. 4, 6, and 7, we used Oja's rule [26] to set the compression weights for the "Hebbian" compression strategy. According to Oja's rule, weights follow the following differential equation:

$$\dot{G}_{ij}(t) = \eta c_i(t)\left(x_j(t) - G_{ij}(t)c_i(t)\right) \quad , \tag{53}$$

where η is the learning rate. If η is decays to zero over the course of training, c_i converges to the leading principal component (PC) of the input representation $\boldsymbol{x}$ [26]. We assume that the conditions under which this is true are verified, and set each row of the compression matrix G to extract the leading PC of the input reaching each compression layer unit.

In Fig. S4e,f, we used a Hebbian / anti-Hebbian learning rule proposed in [28] to learn the compression weights. This learning scheme updates both the feedforward (excitatory/inhibitory) and the recurrent (inhibitory only) weights to introduce competition among compression layer units, enabling the extraction of sub-leading PCs. The update rules for feedforward and recurrent weights are:

$$D_i^c(k+1) \leftarrow D_i^c(k) + c_i^2(k) \tag{54}$$

$$G_{ij}(k+1) \leftarrow G_{ij}(k) + \frac{1}{D_i^c(k+1)}\left(c_i(k)x_j(k) - c_i^2(k)G_{ij}(k)\right) \tag{55}$$

$$W_{c,ij}(k+1) \leftarrow W_{c,ij}(k) + \frac{1}{D_i^c(k+1)}\left((1+\gamma)c_i(k)c_j(k) - c_i^2(k)W_{c,ij}(k)\right) \quad , \tag{56}$$

where γ is a hyperparameter that controls the strength of competition among compression compression layer units.

12.1 Supervisory input from DCN to the pontine nuclei

To model supervisory input from DCN affecting plasticity at cortico-pontine synapses, we assume that the DCN output is close to the target output, i.e. $DCN(t) \simeq y(t)$. We change Oja's rule to

$$\dot{G}_{ij}(t) = \eta \left[c_i(t) + F_i y(t)\right] \left[x_j(t) - G_{ij}(t) \left(c_i(t) + F_i y(t)\right)\right] \quad , \tag{57}$$

where F_i is a constant determining the strength of the supervisory input. This form of plasticity can be interpreted as adding DCN input to the compression layer representation *only for plasticity purposes*, without changing the network dynamics. For simplicity, let's consider the effect on a single compressed-layer unit $c := c_i$, assuming that the target is a scalar, i.e. $y \in \mathbb{R}$. Assuming that the weight evolution is slow, so that we can average over $\boldsymbol{x}$, we obtain

$$\dot{\boldsymbol{G}} = \eta \left(\mathrm{C}^x \boldsymbol{G} - \left(\boldsymbol{G}^T \mathrm{C}^x \boldsymbol{G} - F^2 \langle y^2 \rangle \right) \boldsymbol{G} + F \boldsymbol{R}^{xy} \right) \quad , \tag{58}$$

where $\boldsymbol{R}^{xy} := \langle \boldsymbol{x} y \rangle_x$ is the vector of input output correlations, with one component for each input dimension, and $\boldsymbol{G}$ denotes the vector of incoming weights onto c.. We can express both $\boldsymbol{G}$ and $\boldsymbol{R}$ in the basis formed by the PCA eigenvector of $\boldsymbol{x}$, i.e. $\boldsymbol{G}(t) = \sum_{i=1}^{N} \mu_i(t) \boldsymbol{a}^{(i)}$ and $\boldsymbol{R}^{xy} = \sum_{i=1}^{N} r_i \boldsymbol{a}^{(i)}$. We can then rewrite Eq. 58 as

$$\dot{\mu}_i = \eta \left(\lambda_i \mu_i - \left(\sum_{j=1}^{N} \lambda_j \mu_j^2 - F^2 \langle y^2 \rangle \right) \mu_i + F r_i \right) \quad . \tag{59}$$

This equation shows that the overlap μ_i of the weight vector with a certain input PC $\boldsymbol{a}^{(i)}$ will be driven by how correlated that PC is with the target y (last term on the RHS). Therefore, if F is large enough, even leading PCs which are uncorrelated with the target will not be extracted by compression layer neurons.

13 Forward model learning

When learning a forward model, the network should learn to predict the sensory consequences of motor commands. We assume that the dynamics of a motor plant, such as an arm or a leg, can be summarized by a set of differential equations

$$\dot{\boldsymbol{s}}(t) = f(\boldsymbol{s}, \boldsymbol{u}) \quad , \tag{60}$$

where $\boldsymbol{s} \in \mathbb{R}^{N_s}$ describes the sensory state associated with the plant (such as proprioceptive or visual feedback), $\boldsymbol{u} \in \mathbb{R}^{N_u}$ is the motor command, and f is a smooth, vector-valued nonlinear function that summarizes the dynamics of the plant. The forward model task is then to predict $\boldsymbol{s}(t + \Delta)$, given $\boldsymbol{s}(t)$ and $\boldsymbol{u}(t)$. If the time interval Δ is small to compared to the speed of the plant dynamics, we can approximate

$$\boldsymbol{s}(t + \Delta) \simeq \boldsymbol{s}(t) + \Delta f(\boldsymbol{s}(t), \boldsymbol{u}(t)) \quad . \tag{61}$$

We assume that the cerebral cortex sends to the cerebellum information about both $\boldsymbol{s}(t)$ and $\boldsymbol{u}(t)$. To implement a forward model, the cerebellum should relay the information received by the cortex (first term in Eq. 61), and add to it the nonlinear function $f(\boldsymbol{s}, \boldsymbol{u})$. We assume that the relay operation is carried out by the mossy fiber to DCN pathway, while the Purkinje cell compute the negative of the nonlinear term and feed it to the DCN. The target of learning at Purkinje cells is then given by $\boldsymbol{y}(t) = -f(\boldsymbol{s}(t), \boldsymbol{u}(t))$.

In the model, we concatenate $\boldsymbol{s}(t)$ and $\boldsymbol{u}(t)$ in a single vector $\boldsymbol{z}(t) \in \mathbb{R}^{N_s + N_u}$ and embed it in the input representation in a distributed fashion, using an orthogonal matrix embedding A. The target $\boldsymbol{y}$ of forward model learning is a vector, with an entry for each degree of freedom of the motor plant. In simulations, we only consider one target entry at the time, i.e. we assume that different target components are learned by separate sets of Purkinje cells.

13.1 Random target functions from Gaussian process priors

In Fig. S5, we considered an ensemble of target functions sampled from a Gaussian process (GP) prior. We set the mean function of the GP to be zero and its covariance function to be given by

$$K(\boldsymbol{z}, \boldsymbol{z}') = e^{-\frac{||\boldsymbol{z} - \boldsymbol{z}'||^2}{2\sqrt{D}\lambda^2}} \quad , \tag{62}$$

where λ is a parameter that controls how quickly-varying the sampled functions are.

13.2 Planar two-link arm target

We consider dynamics of a two-joint arm in the absence of gravity (planar) [44]. The arm consists of two bars of length l and mass m, and its state is defined by the two joint angles θ_1 and θ_2, and by the corresponding angular velocities. The dynamics equations are written, in matrix form, as

$$\mathrm{M}(\boldsymbol{\theta})\ddot{\boldsymbol{\theta}} + \mathrm{B}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})\dot{\boldsymbol{\theta}} = \boldsymbol{u} \quad , \tag{63}$$

where $\boldsymbol{u}$ contains the two torques and M is a two-by-two matrix that contains the inertial terms

$$\mathrm{M}(\boldsymbol{\theta}) = \begin{pmatrix} I_1 + I_2 + m_2 l_1^2 + m_2 l_1 l_2 \cos(\theta_2) & I_2 + \frac{m_2 l_1^2 l_2}{2} \cos(\theta_2) \\ I_2 + \frac{m_2 l_1^2 l_2}{2} \cos(\theta_2) & I_2 \end{pmatrix} \quad , \tag{64}$$

where I_1 and I_2 are the moments of inertia of the two links. The matrix C is given by

$$\mathrm{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) = \frac{m_2 l_1 l_2}{2} \sin(\theta_2) \begin{pmatrix} -2\dot{\theta}_2 & -\dot{\theta}_2 \\ \dot{\theta}_1 & 0 \end{pmatrix} - \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \quad , \tag{65}$$

where D_1 and D_2 control the damping strength.

14 Data analysis

14.1 Summary of experimental setup

The recordings analyzed in Fig. 7 were performed using the experimental setup described in detail in [13]. In brief, *Ai93/ztTA/Math1-Cre/Rbp4-Cre* quadruple transgenic mice were head-fixed and performed pushed a handle to perform L-shaped trajectories, either to the left or to the right. We only considered data from the first session after a mouse was considered expert on the task. Furthermore, we only retained "pure" turn trials, i.e. trials in which mice did not push the handle in the incorrect lateral direction by more than 500 μm at any point during either the forward or lateral motion segments. During the task, neural activity from layer-5 pyramidal neurons in premotor cortex and cerebellar granule cells was monitored simultaneously using two-photon microscopy, with a 30Hz sampling rate. More precisely, granule cells were imaged through a cranial window on top of lobules VI, simplex, and crus I.

14.2 Model and input representation

To estimate task-relevant and task-irrelevant activity from the cortical recordings, we regress the cortical activity using a set of basis functions aligned to the turn point in the behavioral trajectories. More precisely, we used two boxcar functions covering at most one second before the turn and two boxcar functions covering at most one second after the turn. Furthermore, we used separate basis functions for right and left turns. In total we then used eight boxcar basis functions, whose length was adapted to each trial trajectory. We considered task-relevant the activity that could be predicted using a linear model with such basis functions as predictors. All the residual (unpredicted) activity was deemed task-irrelevant.

The unobserved population followed the same update equations as the observed one. However, instead of having the measured cortical activity as the input, we generated synthetic data based on the measured task-relevant and task-irrelevant statistics. Synthetic task-relevant activity was generated using the linear regression model described above as a generative model. To sample task-irrelevant activity, we measured the sample covariance matrix of task-irrelevant activity separately for each session, and use it to generate new task-irrelevant activity for the unobserved population, assuming Gaussian statistics. Importantly, task-relevant and task-irrelevant activity was respectively weighted by two parameters, σ_s^{uo} and σ_n^{uo}.

14.3 Measures of correlation and selectivity

The correlations in Fig. 7c were measured by computing the Pearson correlation coefficient among all neuron pairs. These correlation coefficient had mean zero across neuron pairs, therefore we took their standard deviation across neuron pairs as the measure of correlation strength for a single session, and then averaged across sessions. The same procedure was applied to correlations among granule cells and between granule cells and layer-5 cells, both in data and in the random and Hebbian model.

Our measure of cell selectivity to left/right turns is analogous to the one used in [13]. In particular, we devise an encoding model using four boxcar basis functions corresponding to before/after left/right turns. Each boxcar function was at most 300 ms long. After fitting the linear regression model with these basis functions, we quantified the number of coefficient significantly different from zero, independently for each of the 4 basis functions (criterion: $p < 0.01$), and normalized it by the number of neurons.

Supplemental figures

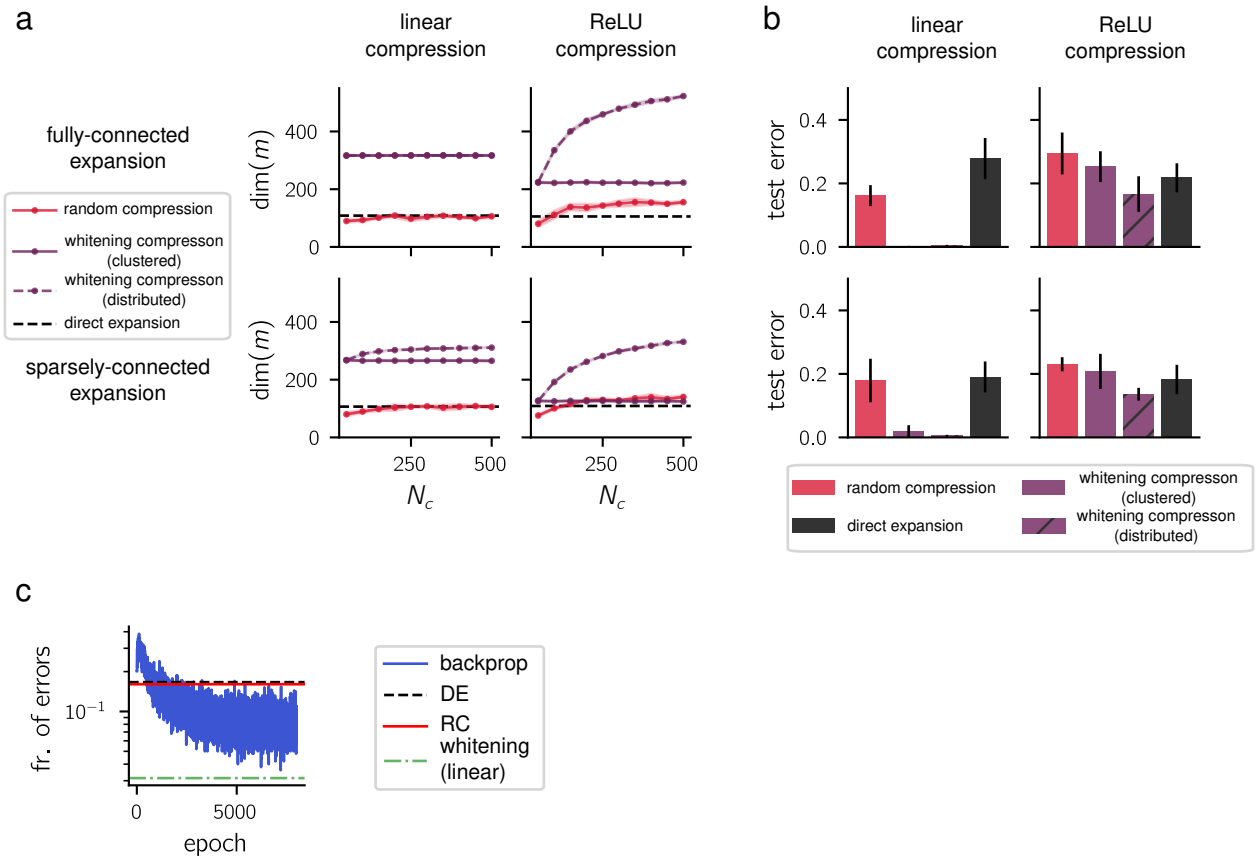

Figure S1: **Effect of nonlinearities at the compression layer.** **a:** Dimensionality of the expansion layer representation for fully-connected versus sparsely connected expansion (first versus second row) and for linear versus nonlinear (ReLU) compression. Shading indicates mean ± standard deviation. **b:** Same as **a**, but showing the generalization error in random classification. Errorbars indicate standard deviation. For both panels, the coding level of the compression layer was set to $f_c = 0.3$ and the other parameters were $p = 1$, $D = 50$. **c:** Classification error as a function of training epoch when compression weights are trained using error backpropagation. Trained, direct expansion and random compression architectures all have ReLU nonlinearities. For comparison, we also plot the performance of whitening with a linear compression layer.

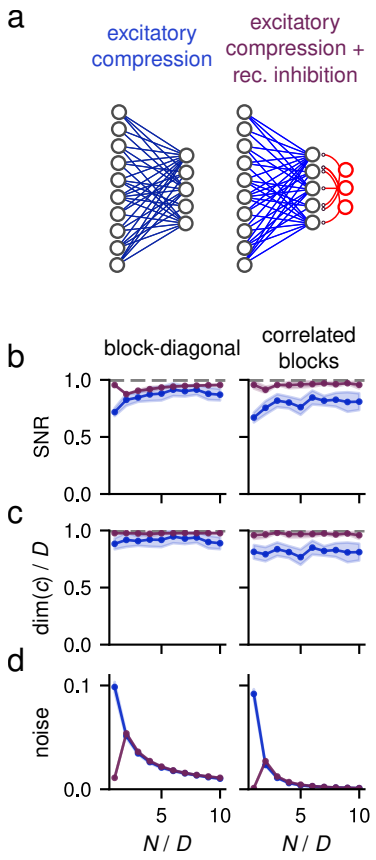

Figure S2: **Constrained compression. a:** Schematic of purely excitatory compression (left) and excitatory compression in presence of recurrent inhibition (right). For panels **b**, **c**, **d**, the left column shows results for gradient descent training for a clustered representation with block-diagonal covariance matrix, while the right columns for a clustered representation with correlations among clusters. The training objective is to maximize the SNR of a Hebbian classifier reading out from the compression layer, i.e. $\mathrm{SNR} = \dim(c)(1-\Delta_c)^2$. **b:** SNR (training objective) as a function of the number of input neurons N per task manifold dimension D. The blue curve is for purely excitatory compression, while the purple curve is for simultaneous training of excitatory compression weights and recurrent inhibitory weights in the compres- sion layer. Shading indicates plus-minus the standard deviation across input realizations and weight initializations. **c:** Same as **b**, but for dimension of the compression layer representation. **d:** Same as **b**, but for noise strength in the compression layer. Parameters: $D = 10$, $\sigma = 0.2$, $p = 1$, $N_c = 10$.

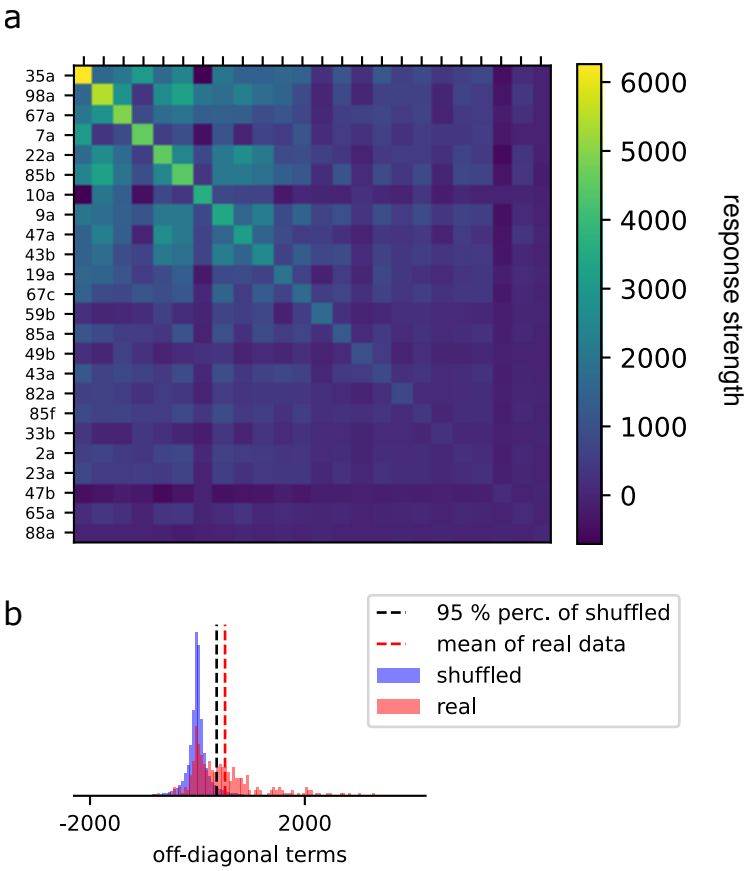

Figure S3: **Realstic properties of odor receptor responses. a:** Covariance of single odor receptors responses, computed from the Hallem-Carlson dataset [18], sorted according to the response variances. **b:** Histogram of off-diagonal terms in the covariance matrix in **a** (in red), compared to a shuffle distribution (blue) obtained by shuffling the responses to different odorant for a given odor receptor.

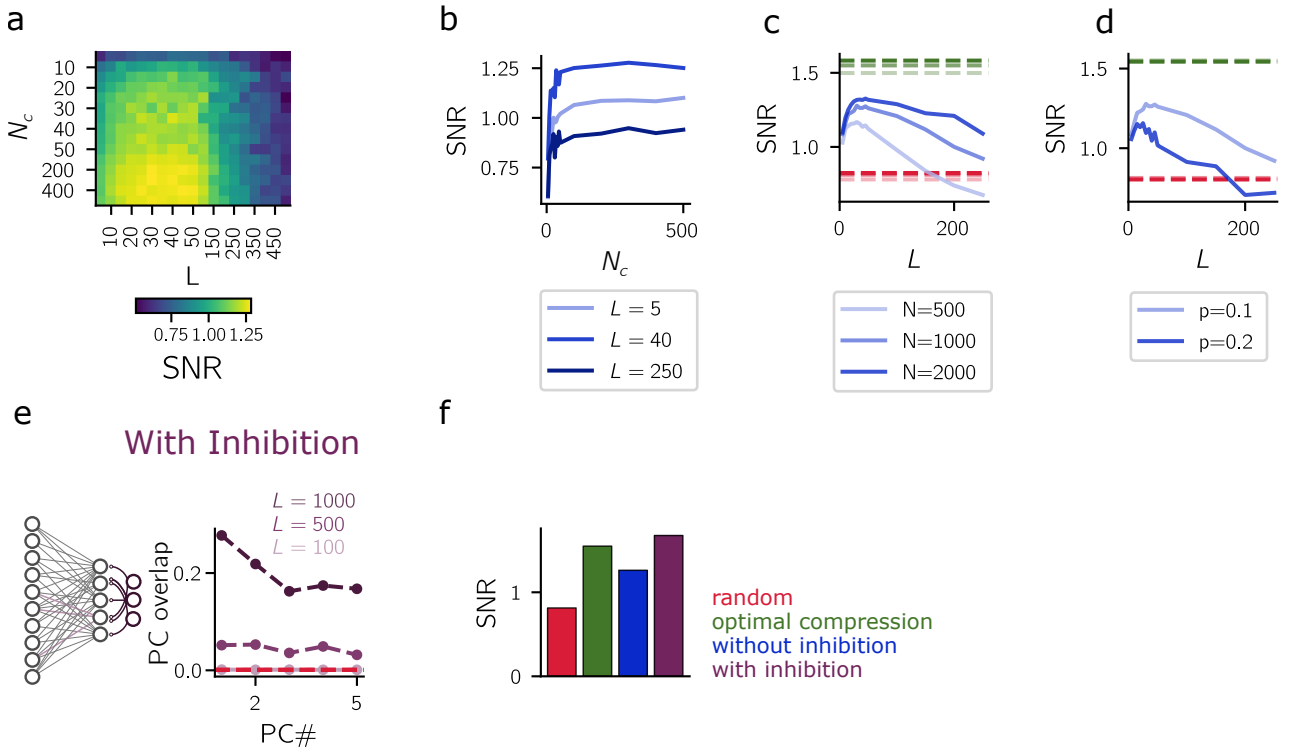

Figure S4: **Effect of architectural paramters on Hebbian plasticity efficacy a, b:** Dependence of SNR on N_c. The non-monotonic profile of the SNR with L is present for a large range of N_c. **a:** heatmap and **b:** vertical slices from L. **c:** The SNR depends weakly on the number of input units N. **d:** SNR versus L for different values of p. **e:** Left: schematics of the setup in which compression weights are learned with Hebbian and anti-Hebbian learning rules in the presence of recurrent inhibition. Right: resulting mean squared overlaps as a function of PC number as in Fig. 6a. For very large in-degree, several PCs are estimated considerably better than without recurrent inhibition. **f:** SNR at the mixed layer for different compression strategies. Biologically plausible learning achieves SNR which is statistically indistinguishable from the one of PCA compression.

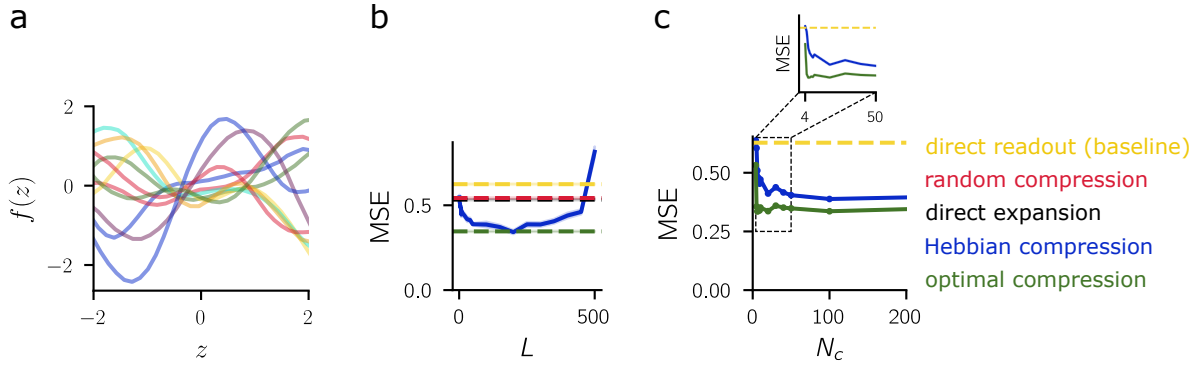

Figure S5: **Learning random target functions. a:** Examples of random target functions sampled from a Gaussian process prior. **b:** MSE of our model with Hebbian compression at the cortico-pontine synapses in learning a smooth Gaussian-process (GP) target, plotted against the pontine in-degree L. Similar to the results for classification in Fig. 6, the performance is non-monotonic with L. **c:** Same as **b**, but plotted against N_c. The inset magnifies the small-N_c region to highlight the faster decay of the MSE for optimal compression compared to Hebbian compression.

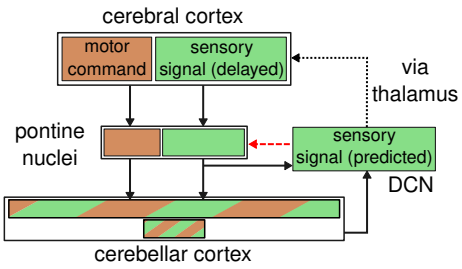

Figure S6: **Forward model.** Schematic illustrating information flow in the mammalian cortico-cerebellar pathway. We assume that the cortical representation can be segregated into motor and sensory related components. Sensory information the cortex is delayed due to sensory delays. Both representations are sent to the pontine nuclei and then mixed in the cerebellar cortex. The pontine nuclei also receive feedback (red dashed arrow) from the deep cerebellar nuclei (DCN), the output structure of the cerebellum.