nature neuroscience

Article

https://doi.org/10.1038/s41593-025-02183-y

Neural population geometry and optimal coding of tasks with shared latent structure

Received: 11 April 2024

Accepted: 18 November 2025

Published online: 04 February 2026

Check for updates

Albert J. Wakhloo[1,2], Will Slatton[2,3,4] & SueYeon Chung[2,3,4]

Animals can recognize latent structures in their environment and apply this information to efficiently navigate the world. Several works argue that the brain supports these abilities by forming neural representations from which behaviorally relevant variables can be read out across contexts and tasks. However, it is unclear which features of neural activity facilitate downstream readout. Here we analytically determine the geometric properties of neural activity that govern linear readout generalization on a set of tasks sharing a common latent structure. We show that four statistics summarizing the dimensionality, factorization and correlation structures of neural activity determine generalization. Early in learning, optimal neural representations are lower dimensional and exhibit higher correlations between single units and task variables than late in learning. We support these predictions through biological and artificial neural data analysis. Our results tie the linearly decodable information in neural population activity to its geometry.

Humans constantly solve different instances of similar problems. We brake at stop signs, stop at red lights and slow down in crowded streets. We do these things effortlessly and efficiently learn to use new sensory cues to regulate our behavior. This is possible because we are able to recognize overt symbols such as road signs as well as more abstract visual cues such as the crowdedness of a street. More generally, humans and other animals learn to recognize latent variables in their environment and use them to guide their behavior across contexts and tasks.

Recent experimental findings have described coding strategies that may underlie this ability. In particular, several studies have described cases where independent variables in the environment are represented along distinct directions of variation in the neuronal activity space[1–7]—for example, in orthogonal subspaces of the firing rates of a collection of neurons[2,7–9]. These independent factors have ranged from the contacts of distinct mouse whiskers[2] to more abstract latent variables, such as the values of different choices in a decision-making task[3]. Neural representations in which distinct environmental variables are represented along independent or orthogonal directions of variation are referred to as factorized or disentangled. Factorized representations were recently shown to emerge in artificial networks trained on multiple tasks[8] and are thought to support generalization to new contexts as well as efficient learning of new tasks that depend on shared latent variables[4,10].

In a related line of work, studies have argued that the brain makes widespread use of cognitive maps to solve these problems[11]. These are coding strategies in which environmental variables are represented in the population code in a way that preserves task-relevant relations between them. This idea is supported by a range of findings—for example, studies in which structurally similar tuning profiles emerge in a neural population for distinct types of environmental variables[12–17]. These variables have ranged from an animal's position to the frequency of an auditory stimulus[12] to more abstract quantities, such as the amount of evidence accumulated in a decision-making task[14]. Here and in the findings referenced above, neural populations represent latent structure in the environment in a way that supports a target behavior. However, defining measures for and understanding why certain neural activity patterns represent latent variables in task-efficient ways remains challenging[18].

A promising approach to tying neural activity patterns to computational goals is to study the geometry of neural responses[1]. Here, the overarching idea is to find which mesoscopic statistics of the population activity contribute to a macroscopic target computation or behavior. In this way, we can gain insight into neural computation without

[1]Zuckerman Institute, Department of Neuroscience, Columbia University, New York, NY, USA. [2]Center for Computational Neuroscience, Flatiron Institute, New York, NY, USA. [3]Center for Neural Science, New York University, New York, NY, USA. [4]Department of Physics and Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, USA. e-mail: sueyeonchung@g.harvard.edu

https://doi.org/10.1038/s41593-025-02183-y

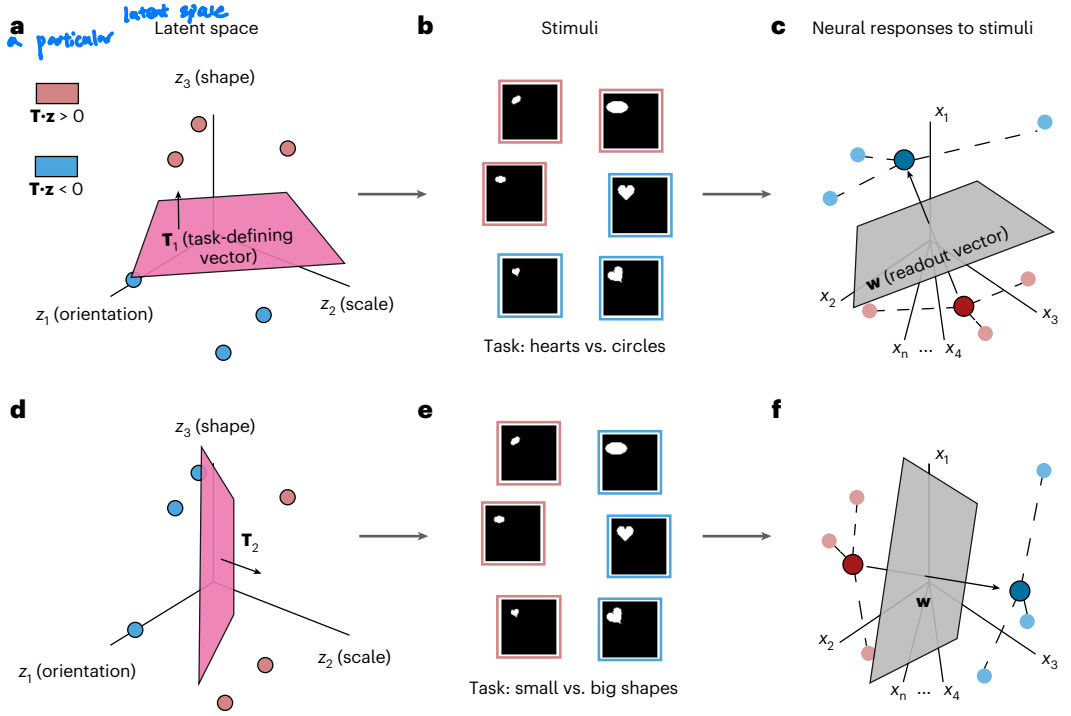

Fig. 1 | Schematic of the task and model setup using images from the dSprites dataset as an example. a,**b**, Stimuli in the dataset vary along a few latent dimensions, corresponding to the shape, orientation and position of the object. Thus, points in the latent space (**a**) can be directly mapped to visual stimuli (**b**). We form binary discrimination tasks, such as hearts versus circles, by linearly separating the latent space using a hyperplane with normal $\mathbf{T}_1$. **c**, Each stimulus elicits a neural activity pattern, visualized as points in an activity space. We form a linear readout of the neural activity by considering the difference between the mean activity pattern for circles (dark red) and hearts (dark blue). This readout corresponds to the activity of an idealized downstream unit with synaptic weights set by a supervised Hebbian learning rule (Methods)[40]. **d**,**e**, A new binary discrimination task (small versus big shapes) can be formed by separating the same set of stimuli using a different hyperplane with normal $\mathbf{T}_2$. **f**, For this task, the same supervised Hebbian rule leads to a new neural readout.

having to give a detailed account of microscopic single-unit activity. For example, in the domain of invariant object recognition, recent works analytically tied coding efficiency[20–23] and few-shot generalization performance[24] to measurable statistics of the population activity. In a related line of work, several studies tied the structure of neural correlations and population-level statistics to the information content of neural codes[25–29]. Recent studies have also investigated a wide range of motor[30,31], sensory[2,6,15,32–35] and decision-making[3,4,13,14] computations and behaviors by analyzing the geometry of neural population responses.

In the present study, we develop an analytical theory for learning binary decision-making tasks depending on a common latent structure that directly ties the statistics of neural population responses to generalization performance. Although several authors have used linear probes or heuristic geometric measures to analyze population activity in such tasks (for example, refs. 2,4), to our knowledge there is no cohesive theory that directly ties mesoscopic statistical features of neural activity to task performance in these settings. To fill this gap, we analytically calculate how neural population geometry shapes the generalization error of an agent learning multiple tasks that depend on a common latent structure.

Results

Geometric measures govern generalization of linear readouts

We study the ability of a neural population to support downstream learning of tasks that depend on a common latent structure. To do this, we start by considering a set of stimuli that vary along a few latent dimensions. For example, in the dSprites dataset depicted in Fig. 1a,b, different images can be described in terms of the shape, position and orientation of the shape in each image[36]. Within our modeling framework, this corresponds to the assumption that each stimulus can be mapped to a vector $\mathbf{z}$ in an underlying d-dimensional latent space. In response to each stimulus, we consider a vector $\mathbf{x}$ of neural responses—for example, the firing rates of a population of neurons to each image in the dataset (Fig. 1c). Thus, our modeling framework considers datasets comprising latent vectors describing individual stimuli, together with neural activity patterns associated with each stimulus.

Within this setting, we determine how well neural responses can be used to solve binary tasks involving the latent variables. The task labels for a given binary classification task are formed by linearly separating the latent space into two pieces using a hyperplane with normal vector $\mathbf{T}$[37–40]. In the example shown in Fig. 1, different choices of $\mathbf{T}$ generate shape categorization (Fig. 1a,b) or size categorization (Fig. 1d,e) tasks. Note that, although the separation is linear in the latent space, it may be highly nonlinear in the stimulus space.

We quantify how well these neural activity patterns support downstream classification by considering the performance of a simple linear readout mechanism. Specifically, we calculate the generalization error of an idealized downstream unit that receives inputs from all neurons in the population and adjusts its synaptic weights according to a supervised Hebbian plasticity rule (Methods). The activity of this downstream unit is then determined by a weighted summation of its inputs, followed by a threshold operation. This unit's activity acts as a difference of means classifier when the number of labels for each class is equal (Fig. 1c–f)[40]. We calculate the generalization error of this readout after being trained on a dataset of p-many stimuli both for a single, fixed task (for example, the hearts versus circles task in Fig. 1b) as well as the average generalization error across all possible tasks. In this way, we connect the statistical properties of the neural code to the multitask learning problem.

Using this simplified model, we derive a formula for the average generalization error. As described in the Supplementary Information, we prove that, under a certain Gaussian approximation, the error E_g is given by

https://doi.org/10.1038/s41593-025-02183-y

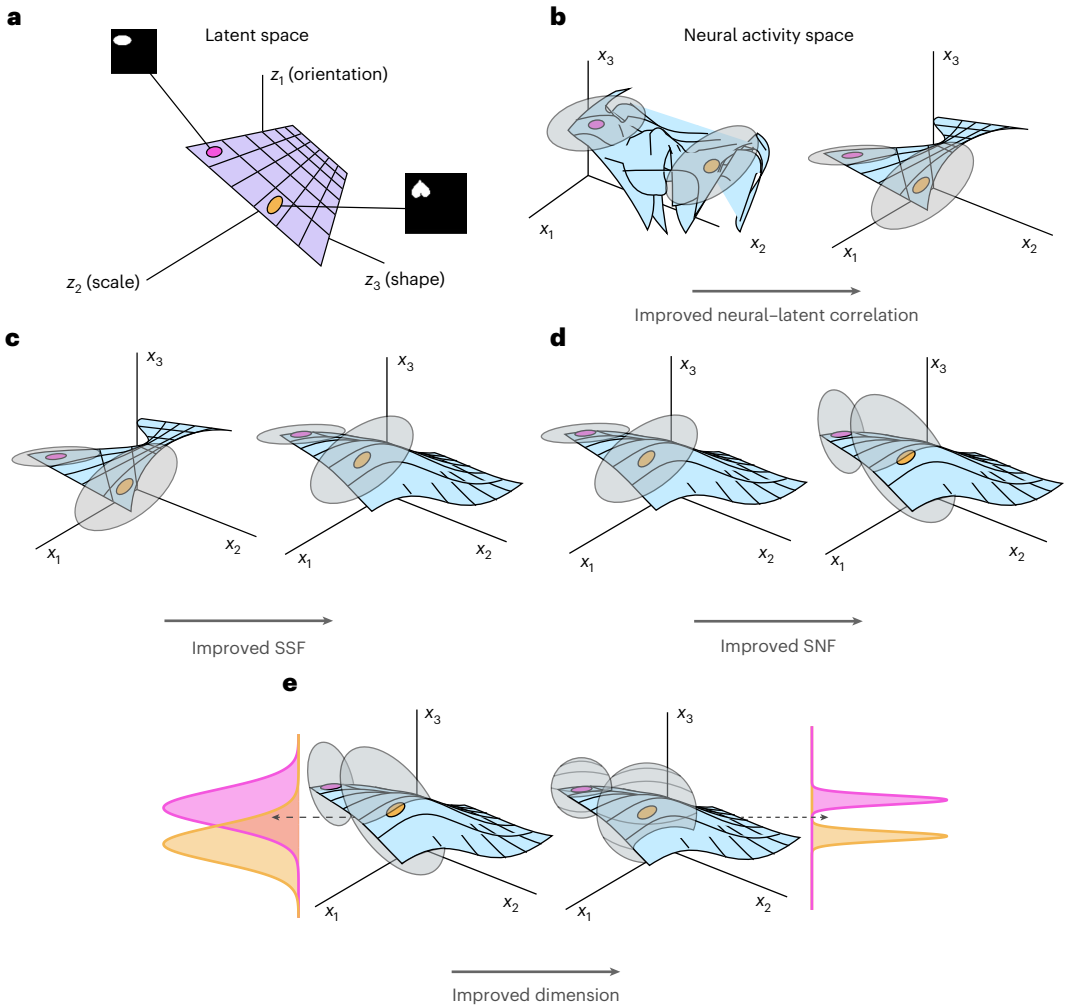

Fig. 2 | Schematic of the geometric terms. We visualize different possible neuronal activity patterns elicited by the same set of stimuli. **a**, A small slice of the latent space from which stimuli are generated. **b**, Visualization of neural activity patterns responding to this set of stimuli with low (left) and high (right) total correlation. When the correlation is high, the relative distances between points in the latent space are approximately preserved in the neural state space. **c**, SSF. When the SSF is low, different latent variables are represented along overlapping directions; when it is high, independent latent variables are represented along uncorrelated directions of variability in the neural state space. **d**, SNF. When the SNF is low, the noise distribution (gray ellipses) around a point in the firing rate space falls along the coding directions; when it is high, the noise distribution is uncorrelated with these directions. **e**, Neural dimension. In higher-dimensional representations, the neural activity and associated noise distribution occupy more directions in the state space, shown here as two-dimensional (left) versus three-dimensional (right) noise distributions. As the dimension increases, the projection of a sample of neural activity onto a given direction becomes increasingly concentrated, supporting generalization performance[24].

Ispiration
Visulize the geometrial patterns.

$$E_g = \frac{1}{\pi}\tan^{-1}\left(\sqrt{\frac{\pi}{2pc^2\mathrm{PR}(\mathbf{\Psi})} + \frac{1}{f} + \frac{1}{s} - 1}\right), \quad (1)$$

where we have written the generalization error as a strictly decreasing function of four geometric terms as well as the number of samples in the training dataset, p. In Fig. 2, we visualize different geometries that could arise in response to a fixed set of stimuli (Fig. 2a). We now discuss each of these terms one by one (see Methods for precise definitions and Supplementary Information section 3 for details):

- **Neural–latent correlation:** c. This term is a normalized sum of squared covariances between neurons and latent variables. As such, it measures the overall correlation between single-unit responses and latent variables. At the population level, the neural–latent correlation measures how sensitive the population activity is to variations in the latent space (Fig. 2b).
- **Signal–signal factorization (SSF):** f. This term measures the alignment between the coding directions of distinct latent variables. The SSF term favors neural coding schemes that represent independent latent variables along uncorrelated directions in the neural state space. Moreover, this term encourages these representations to devote equal variance to each independent latent factor—that is, to form a whitened representation of the latents (Fig. 2c).
- **Signal–noise factorization (SNF):** s. This term measures the magnitude of the noise that lies along the coding directions of the latent variables. Ideally, any noise present in the neural responses should lie in directions that are uncorrelated with the directions representing the latent variables[25,27] (Fig. 2d).
- **Neural dimension:** $\mathrm{PR}(\mathbf{\Psi})$. The participation ratio of the neural responses measures the effective number of dimensions that the population activity spans. When all else is equal, higher-dimensional responses are preferred, as neuronal noise is less correlated from trial to trial[24] (Fig. 2e).

Finally, we note that the contribution to the error of the neural dimension and neural–latent correlation decays to zero with the number of

https://doi.org/10.1038/s41593-025-02183-y

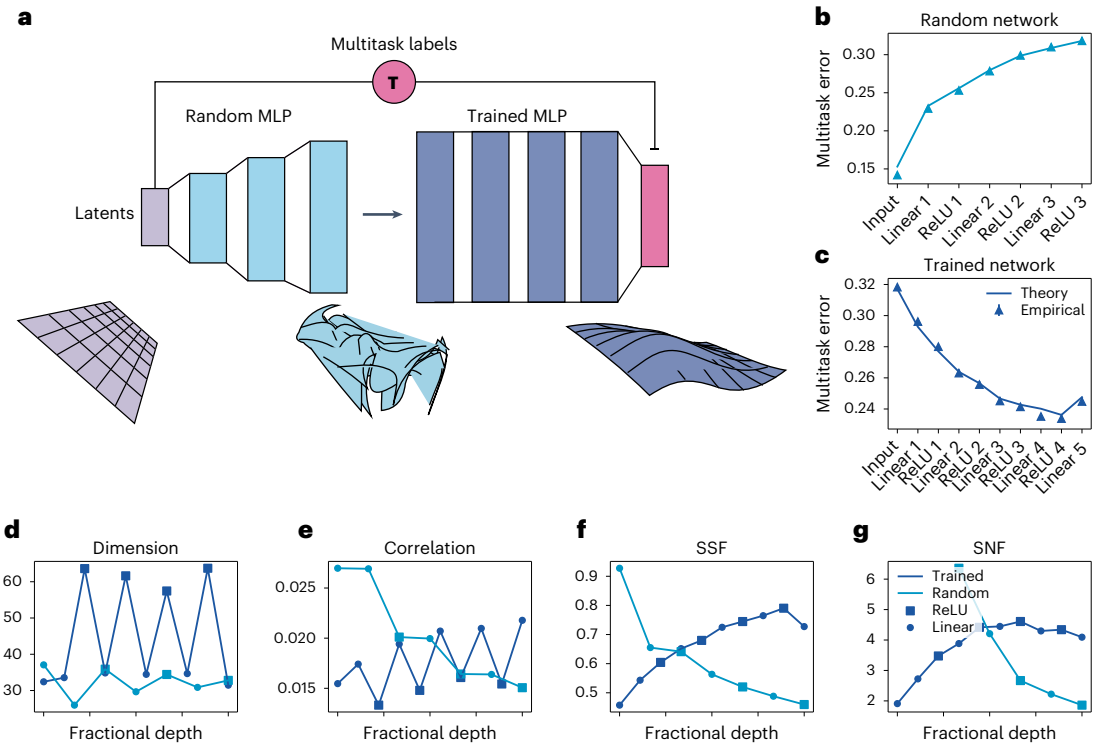

Fig. 3 | Theory predicts generalization error of the Hebbian rule in trained and random MLPs. a, Schematic of the simulation. Latent variables **z** are randomly shattered to generate task labels. These latents are passed through a random MLP (light blue) and are then used as inputs to train a three-hidden-layer MLP (dark blue) on the multitask binary classification problem using stochastic gradient descent. **b**,**c**, After training, we sample a new set of latents and teacher vectors and calculate the generalization error of the Hebbian rule on each layer of the random (**b**) and trained (**c**) network. Theoretical predictions closely track empirical errors, and the trained network achieves a lower error in later layers. **d**–**g**, Geometric terms across layers for the random (light blue line) and trained (dark blue line) networks. Linear layers are marked by circles and ReLU layers by squares. Interestingly, the error only slightly changes across linear and ReLU layers of the same model stage, in spite of sharp changes in the geometry. In the trained network, the application of ReLU consistently causes increases in the dimension and SSF as well as decreases in the correlation.

training samples (Methods). On the other hand, the error stemming from the SSF and SNF represents an 'irreducible error' that does not depend on the number of training samples. Intuitively, in the few-shot regime, the main concern is maximizing the amount of total signal in the neural code while minimizing the impact of noise. These two aspects are primarily controlled by the total correlation and dimension, respectively. On the other hand, in the many-shot regime, the main concern becomes keeping the representations of distinct latent variables separate from one another and separate from noise directions. The two factorization terms control these features of the code. These considerations suggest that certain geometries, which perform well early in learning, when p is small, may be suboptimal late in learning. We make these intuitions precise in 'Optimal representation of latent variables' below.

Geometry of multitask learning in multilayer perceptrons

In practice, neural data are complex and may be non-Gaussian. To test our theory in this regime, we now apply our formula for the generalization error (equation (1)) and its geometric decomposition to both random and trained nonlinear multilayer perceptrons (MLPs). As shown in Fig. 3a, we first sample a set of zero mean Gaussian latent variables **z**. We then feed these latent variables through a random MLP with increasing hidden layer sizes. After generating this set of stimuli, we sample 500 random **T** vectors and use these to generate a set of task labels (Methods). These labels and data points from the random MLP are then used to train a downstream three-hidden-layer MLP, which predicts the label for each task using a task-specific linear readout from the shared penultimate layer. This is a multilayer version of the hidden manifold modeling framework from ref. 38. Finally, we sample a new set of latent variables along with a new set of tasks and calculate the generalization error of the Hebbian rule when applied to the representations at different layers of the trained and untrained MLPs (Methods). This setup allows us to validate our theory on nonlinear transformations of Gaussian latent variables.

As shown in Fig. 3b,c, we find good agreement between our formula, equation (1), and the empirical generalization error. Turning to our geometric decomposition of the error, we find several interesting trends across layers and nonlinearities (Fig. 3d–g). Most notably, in the trained network, we find that linear and rectified linear unit (ReLU) layers orchestrate a tradeoff between the geometric terms that together lead to an overall decrease in multitask generalization error. As shown in Fig. 3d–g, the neural dimension spikes each time the nonlinearity is applied, at the cost of the total correlation. Conversely, at each linear layer, the correlation sharply rises while the dimension falls. On the other hand, we find that both the SSF and SNF terms monotonically increase through the penultimate layer of the network. Notably, this pattern is not present in the random network. Thus, the trained MLP learns to use the nonlinearity to increase the dimension of the representation as well as the factorization of the latent variables while simultaneously squashing particularly harmful noisy directions of variability in the input. We visualize the training dynamics that lead to these representations in Supplementary Fig. 4. Although these are sharp changes in the geometry, the overall generalization error across layers exhibits only minor fluctuations between most linear and ReLU layers, highlighting the limitations of a generalization error-based approach to analyzing network activity without studying the underlying geometry.

Disentangling in deep pose estimation networks

To test the validity of our theory on more complex natural data, we study the representations of animal pose in a deep convolutional

https://doi.org/10.1038/s41593-025-02183-y

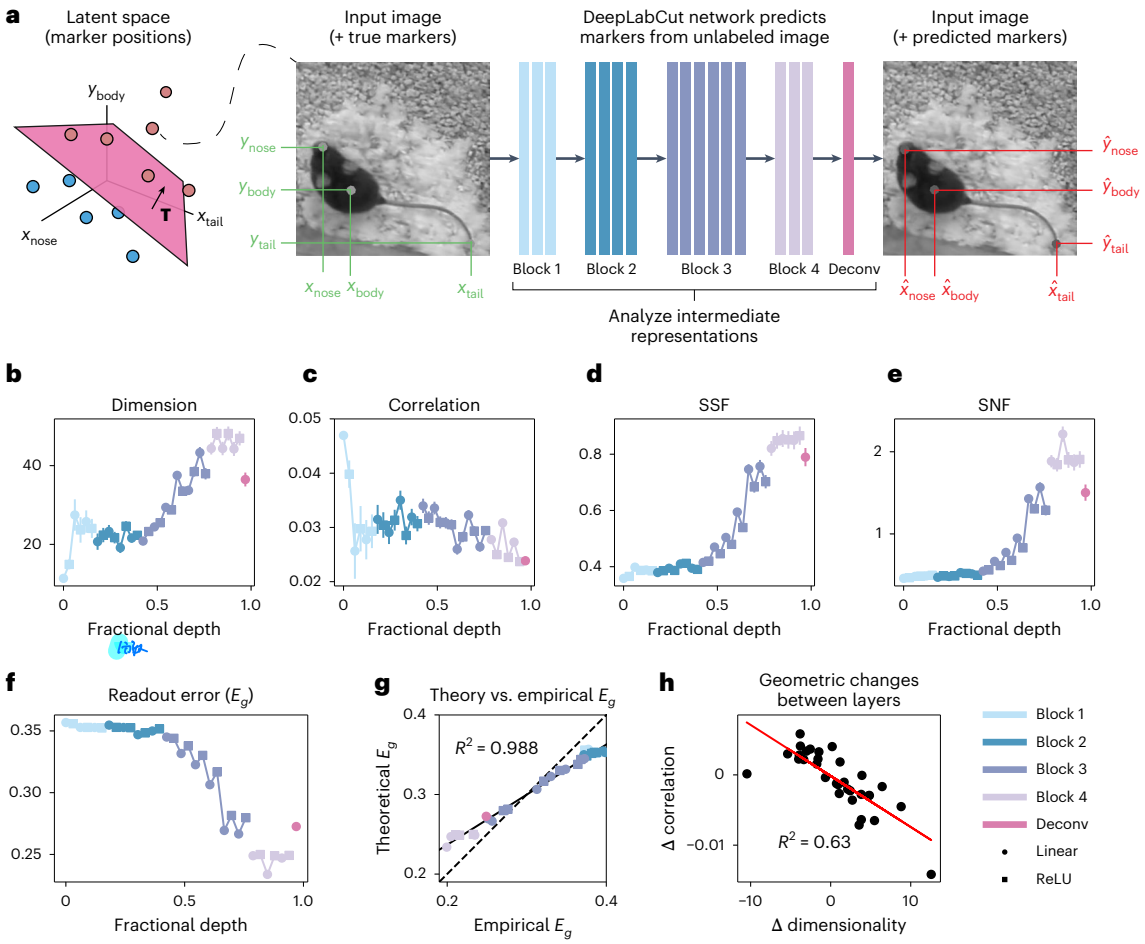

Fig. 4 | Disentangling of animal pose parameters by a deep neural network. Data are reported as means ± s.e.m., where statistics are over 20 random projections of the network representations down to a fixed dimensionality of $n = 100$. **a**, A DCNN trained with the DeepLabCut framework[41] predicts the position of 12 markers (only three are illustrated here) in unlabeled images of an adult mouse ($N = 1$) (Methods and ref. 42). We treat the $d = 24$ total (x, y) coordinates of these marker positions as latent variables and study how the DCNN disentangles them from the image inputs. **b**, The total dimensionality of the DCNN's intermediate representations increases across layers. **c**, The total correlation between the latents and the network's internal representations decays across layers, representing a tradeoff with dimensionality. **d**,**e**, Both SSF and SNF improve monotonically across layers. **f**, Multitask error decreases across layers, indicating successful and gradual disentangling of the latent variables across layers of the DCNN. Here, $p = 150$ training samples are used, and we calculate the error using equation (1). **g**, Our theoretical expression for multitask error in terms of our four geometric terms is not quantitatively exact for this complex natural dataset but, nonetheless, captures essentially all variance in the true multitask error. **h**, For each pair of adjacent layers in the network, we plot the change in the neural–latent correlation and the dimensionality term.

neural network (DCNN) trained to estimate pose parameters from natural images (Fig. 4a)[41,42]. In this setup, the $d = 24$ dimensional latent variables $\mathbf{z} \in \mathbb{R}^{24}$ consist of (x, y) coordinates for 12 different 'marker' locations on a mouse's body (for example, the left paw or the tip of the tail). We calculate the average multitask error generated by randomly shattering this latent space (Fig. 4a) both empirically and using our formula, equation (1). The distribution of the latent variables is highly non-Gaussian and provides a challenge for our theory. The pose estimation network consists of a fine-tuned ResNet50 and a final trained deconvolutional layer with 12 output channels, one for each marker location. This can be viewed as a multitask architecture in which the output of the fine-tuned ResNet50 produces a shared representation that is read out by 12 different deconvolution kernels to estimate the position of each independent marker.

The task-averaged generalization error E_g improves across the layers of the network (Fig. 4f). Interestingly, the geometric changes driving this improvement in E_g are qualitatively different than in the random MLP case. We observe a similar tradeoff between the dimension and correlation terms at each layer (Fig. 4h, compare to Fig. 3d,e), but this network prioritizes expanding the data dimensionality (Fig. 4b) while sacrificing correlation (Fig. 4c). In this more challenging setting with highly non-Gaussian latents and complex natural inputs, our analytical expression for E_g is slightly biased (Fig. 4g). However, the analytical prediction for E_g still linearly explains practically all of the variance in the empirically computed multitask error ($R^2 = 0.988$). Thus, our geometric terms capture the representational changes that drive the observed trends in E_g.

Predicting readout performance of macaque visual representations

Having considered nonlinear artificial neural networks, we now apply our theory to biological neural data. To do this, we draw from preexisting multi-unit recordings from macaque V4 and inferior temporal cortex (IT) taken while two monkeys viewed visual stimuli (Methods and Fig. 5a–c)[43]. The stimuli used in these experiments included images of 64 objects taken from eight categories and were generated by modifying $d = 6$ continuous latent variables that included the size, position and angle of the object. This allows us to form binary classification task labels by shattering this six-dimensional latent space on subsets of the data corresponding to individual object categories; for example,

https://doi.org/10.1038/s41593-025-02183-y

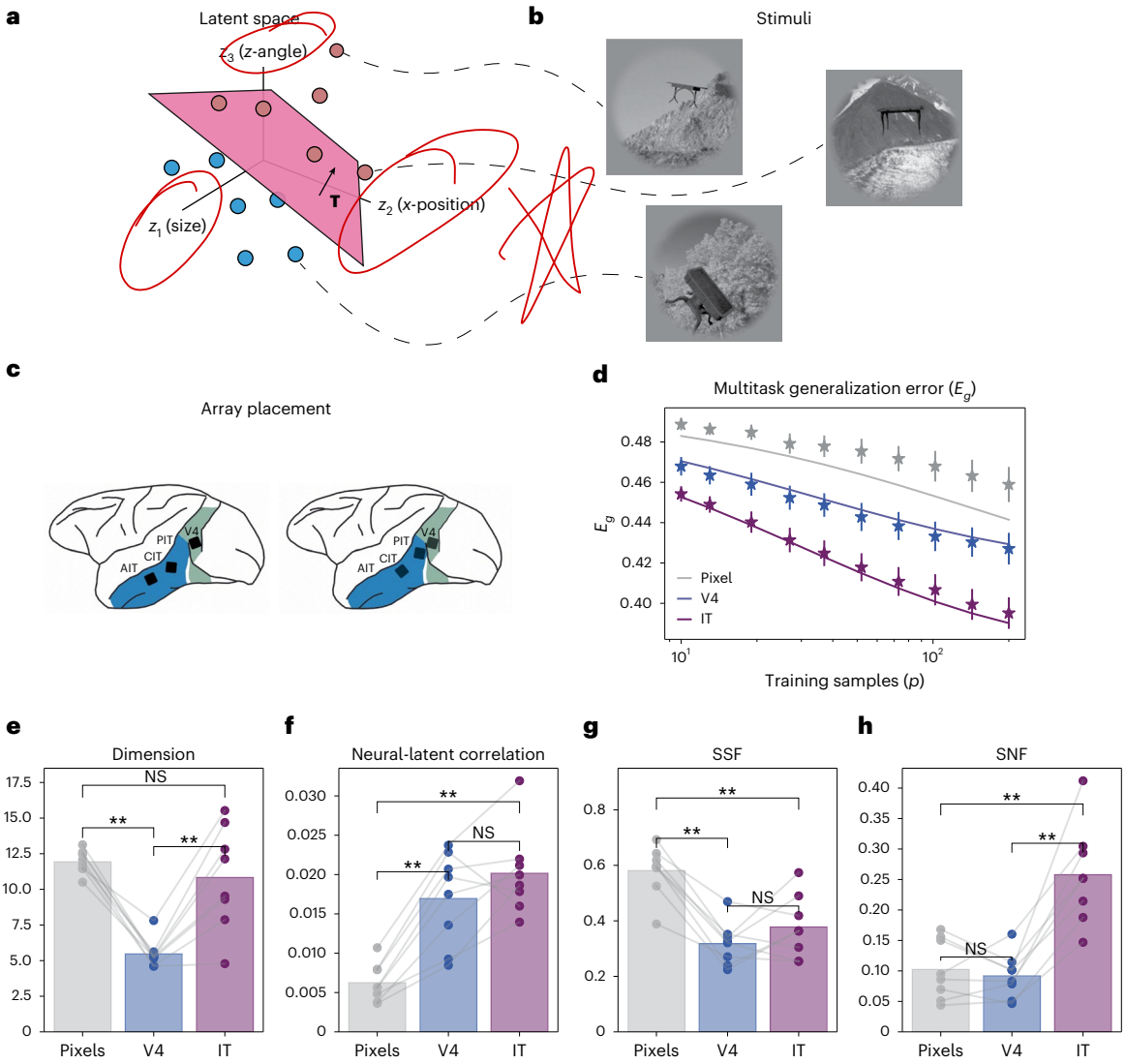

Fig. 5 | Theory predicts multitask error in macaque V4 and IT data. **a**,**b**, Example stimuli and tasks[43]. Visual stimuli included 64 objects grouped into eight categories and were generated by modifying $d = 6$ continuous latent variables that included object size, position and angle. We form binary classification tasks on subsets of the data coming from the same category—for example, all images from the 'Tables' category. **c**, Array placement. Figure was adapted with permission from ref. 43. **d**, Generalization error across pixels and brain regions calculated empirically (markers) and using our formula for the generalization error (solid line). Error bars denote the standard error across categories. **e**–**h**, Geometric terms across pixels and neural responses for each category. Bars denote the mean, and asterisks denote a significant difference ($P < 0.01$, two-sided paired t-test, $N = 8$ object categories; Supplementary Table 1). NS, not significant.

a given task could involve separating images of chairs on the left side of the screen from those on the right. We test the validity of our generalization error equation, equation (1), by calculating the error of Hebbian readouts applied to V4 and IT neural responses as well as the raw pixel values. We find that our formula accurately predicts the generalization error, particularly for the neural response readouts (Fig. 5d). Furthermore, we find that generalization error improves through the ventral stream, in line with previous results[44].

From equation (1), we can see that there are many different geometries that yield the same generalization error. Thus, it is a priori unclear how these measures ought to change from the pixel space through the ventral stream. Applying our geometric decomposition of the generalization error to these data, we find that the dimension is lower in V4 than in either IT or the pixel data[24] (Fig. 5e) and that the correlation increases from the pixels to the visual areas. The transformation from pixels to V4 is reminiscent of the tension we found between correlation and dimension in the preceding sections, whereby transformations that improve generalization error by increasing neural–latent correlation can sometimes do so at the price of a slightly lower dimension (and vice versa). Turning to the factorization measures, the most dramatic trend is the increase in SNF in IT. This suggests that, in IT, latent-unrelated variability overlaps less with the coding directions than in V4 or the pixels (Fig. 5e–h). Taken together, these results show that our formula for the generalization error (equation (1)) captures empirical readout performance and demonstrates the applicability of our metrics as a tool for tying the geometry of neuronal population responses to the computational objective of multitask learning.

Optimal representation of latent variables

Our theory provides an ideal framework to pose normative questions regarding multitask learning. In our model, latent variables that have less variance contain, on average, less information about the task labels (Fig. 6a and Supplementary Information section 4). For the data analyses presented in the previous section, we normalize the latent variables so that they are all weighted equally. However, this aspect of our framework allows us to study how the structure of the latent variables

https://doi.org/10.1038/s41593-025-02183-y

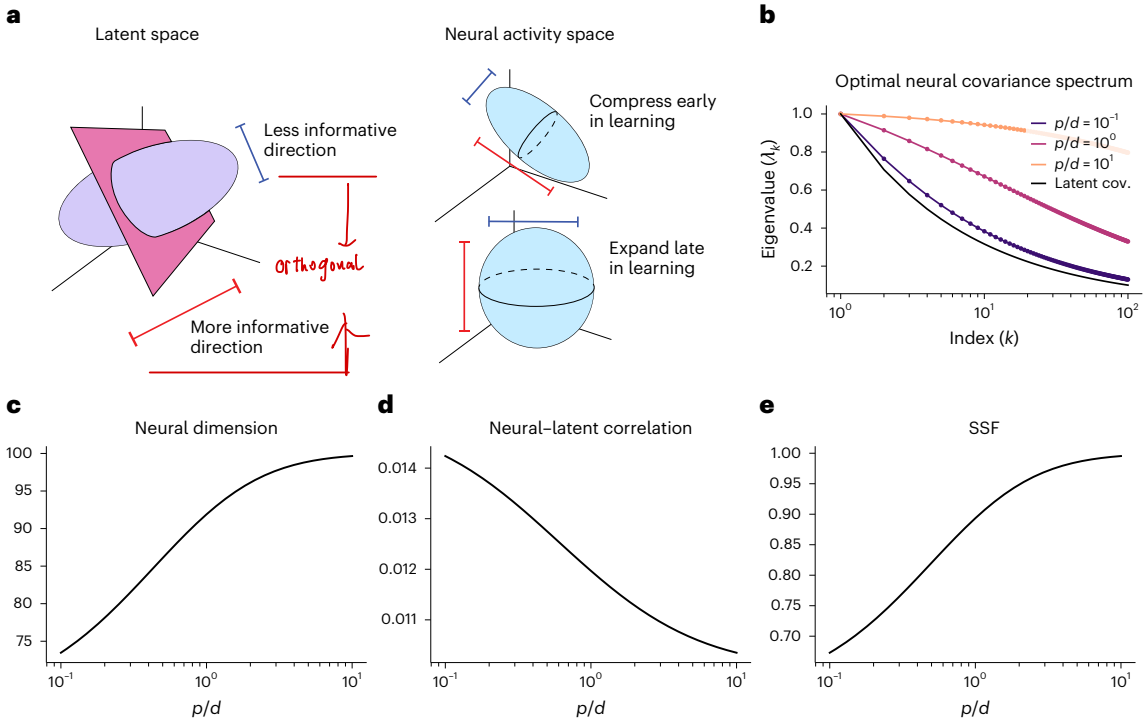

Fig. 6 | Optimal representational geometry over learning. a, In our task setup, directions in the latent space that have little variance are, on average, less informative of the task labels (Methods and Supplementary Information section 4). **b**, Eigenvalues of the optimal neural covariance as a function of the number of samples per latent dimension, p / d. We show the eigenvalues of the latent variables' covariance in black. Markers correspond to results obtained by optimizing our formula for the generalization error numerically, and solid lines correspond to our formula for the optimal code's spectrum (Methods). As training proceeds (that is, as p increases), the spectrum of the optimal neural code becomes increasingly flat. **c**–**e**, Optimal representational geometry at different stages of training. Early in training, optimal representations have a higher neural–latent correlation and a lower dimension and SSF than they do late in training. (Note that we do not plot the SNF, as optimal representations in our framework are noiseless or have noise that is uncorrelated with the signal directions. In this regime, the SNF diverges for all p).

and the number of available training samples determine which neural representations are optimal.

We show analytically that the optimal representation disentangles latent variables (Supplementary Information section 4). More precisely, we show that when the latent variables are not correlated with one another, neural responses represent distinct latent variables along orthogonal directions. When the latent variables are correlated with one another, we show that the principal components of the latent variables directly map onto the (mutually orthogonal) principal components of the neural activity (Supplementary Information section 4). Thus, disentangled representations are optimal in this framework.

How does the structure of the latent variables and the degree of learning affect the geometry of optimal neural representations? In addition to being disentangled, we find that the optimal code compresses less informative variables early in learning (that is, when p is small) and expands these variables late in learning. For an optimal code, independent latent variables are represented along uncorrelated directions; however, the variance along each direction is not equal. As shown in Fig. 6a, early in training, less informative latent variables correspond to directions in the neuronal activity space that have small variance. As learning proceeds (that is, as p increases), the amount of variance in neural space dedicated to these less informative latent variables grows (Methods and Supplementary Information section 4). Intuitively, the optimal code only starts paying attention to the less informative latent variables when there are enough samples to learn their relevance to a given task.

We trace these features of the optimal neural code back to the eigenvalues of the neuronal covariance matrix and our geometric terms. As shown in Fig. 6b, the eigenspectrum of the optimal code becomes increasingly flat over the course of learning. This reflects the fact that more and more variance is being dedicated to the less informative directions in the latent space (that is, directions with smaller variance). Arranging the eigenvalues of the latent variables' covariance matrix in descending order, $\omega_1 \geq \omega_2 \geq \ldots \geq \omega_d$, we show that the optimal code's covariance matrix has at least d non-zero eigenvalues that are given by:

$$\psi_i = C\frac{\omega_i}{2p\omega_i + \pi\sum_k \omega_k}, \qquad (2)$$

where C is an arbitrary constant. We can see that, as p grows, the spectrum becomes increasingly flat, reflecting the expansion strategy (Methods and Supplementary Information section 4).

Turning to our geometric measures, we find that, early in training, optimal neural codes are lower dimensional with higher neural–latent correlations than they are late in training (Fig. 6c–e). A key normative prediction from our theory is, therefore, that the task-related variability—that is, the signal variance—of neural responses becomes increasingly high dimensional as an agent learns to perform tasks that depend on a complex latent structure. Moreover, these results suggest that the strength of the correlations between single units and task variables may decrease in the late stages of learning. We present evidence for these predictions in the following section.

Geometry of spatial representations in prefrontal cortex and CA1 during learning

We now analyze the geometry of navigational representations in prefrontal cortex (PFC) and CA1 as rats learn to perform a continuous alternation task over the course of eight sessions (Fig. 7a,b)[45]. To do this, we analyze the degree to which an animal's position and velocity can be

https://doi.org/10.1038/s41593-025-02183-y

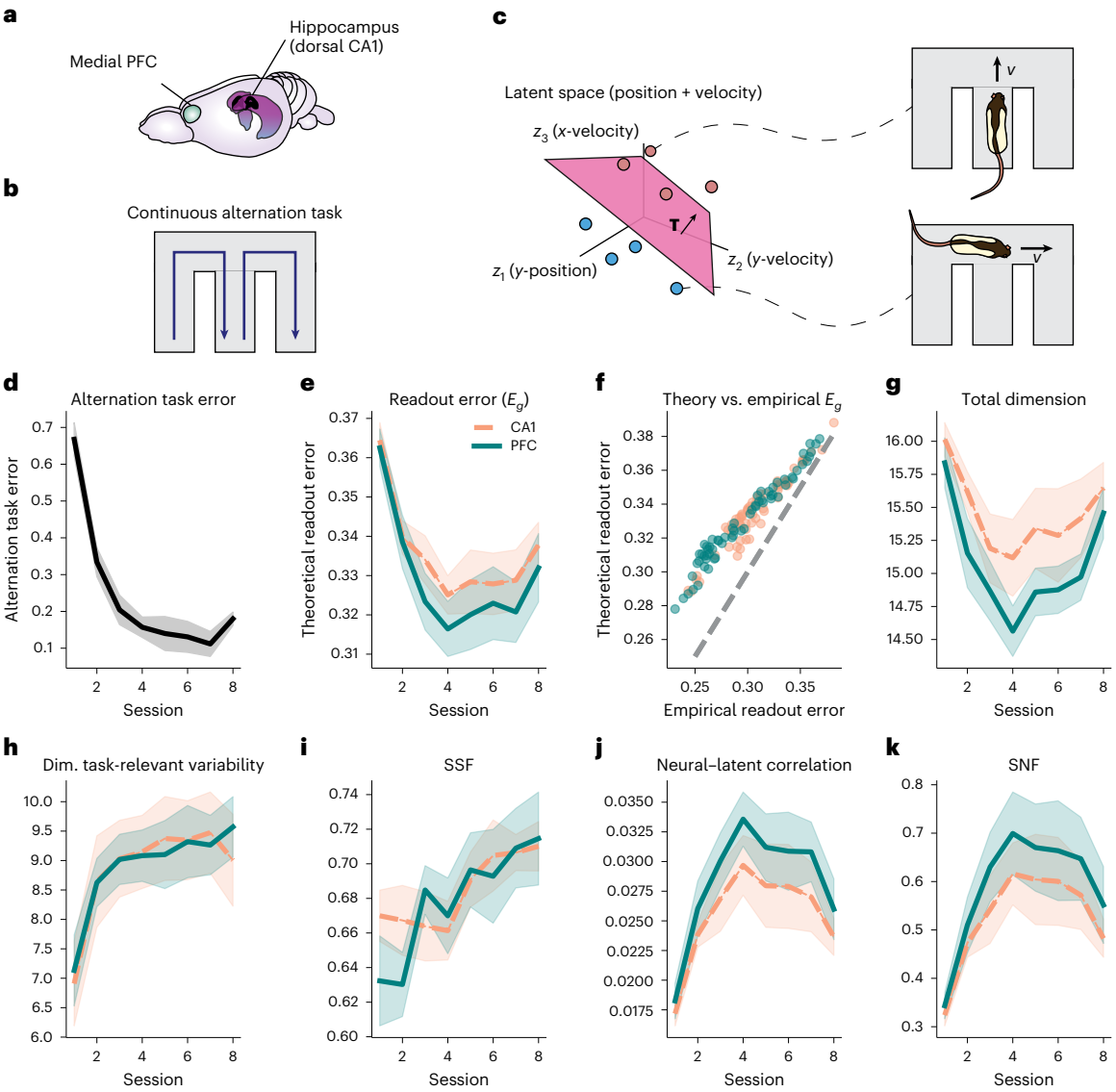

Fig. 7 | Changes in the geometry of spatial representations during learning in PFC (solid) and CA1 (dashed). a, Multi-unit recordings were obtained from medial PFC and dorsal CA1 (ref. 45). Figure is adapted from ref. 45. **b**, Rats performed a continuous alternation task in a W-maze. **c**, For our multitask readout analysis, we treated the *x* and *y* positions and velocity values as latent variables from which binary classification tasks were formed. **d**,**e**, Behavioral error on the continuous alternation task and neural readout error quickly decrease over the first few sessions, leading to a large correlation between the behavioral and readout errors on single sessions ($R^2 = 0.49$ (CA1) and $R^2 = 0.51$ (PFC); Supplementary Fig. 12c). **f**, Empirical generalization error versus our estimate from equation (1). Although there is a small bias for low error values, we find a good agreement overall. **g**–**k**, Geometric terms over the course of learning. Lines denote the mean within a region, and error bars denote the standard error ($N = 8$ rats).

decoded from population firing rates. Specifically, we treat the animal's *x* and *y* position and velocity as latent variables from which we form binary classification tasks (Fig. 7c). We then calculate the task-averaged readout error, E_g, and geometric measures from PFC and CA1 firing rate data for each session ($n \geq 19$ units; Methods). The session-by-session neural readout errors E_g thus quantify the ability of a linear readout to decode position and velocity information through the course of learning. These readout errors decrease over learning (Fig. 7e,f) and are correlated with single-session behavioral task performance ($R^2 = 0.48$ (CA1) and $R^2 = 0.51$ (PFC); Supplementary Fig. 12c).

We now analyze the trends in the readout error and geometry during learning. Over the first four sessions, both behavioral and neural readout error rapidly improve (Fig. 7d,e). During this time, nearly every geometric measure increases (Fig. 7h–k). Interestingly, we find that although the total neural dimension drops over these initial sessions (Fig. 7g), the dimension of the subspace containing task-relevant variability increases (Fig. 7h) (Methods). This suggests that the initial drop in the dimension is driven by a compression of navigation-irrelevant activity. After session four, both behavioral and readout error largely plateau. At this point, several of these initial trends reverse; in particular, the total dimension begins to increase, whereas the correlation drops.

To test the significance of these non-monotonic effects of learning on the geometry, we fit mixed-effects quadratic regression models for each geometric term (Methods and Supplementary Fig. 12). We found strong, significant effects for both the linear and quadratic terms in PFC and CA1 for the total dimension, correlation and SNF ($P < 10^{-5}$; Supplementary Table 2). In each case, the sign of the quadratic term was the opposite of the linear term, demonstrating the non-monotonic effects of learning on these quantities. We did not find evidence for a non-monotonic trend in the SSF in either region ($P > 0.05$). Although we did find weaker significant effects of the quadratic model for the

task-relevant dimension ($P < 0.01$), the model was a comparatively poorer fit to this quantity in PFC (Supplementary Fig. 12). As such, we carried out a follow-up analysis using mixed-effect linear models, which demonstrated significant, monotonic effects of learning on the SSF and task-relevant dimension across both regions ($P < 10^{-3}$ SSF and $P < 10^{-5}$ task dimension; Supplementary Table 2).

How do these trends in the geometry align with our normative predictions? In the previous section, we analyzed the behavior of optimal representations over the course of learning. However, the initial state of the network is likely far from any kind of optimality, as suggested by the very large drops in behavioral error (Fig. 7d), readout error (Fig. 7e) and increases in nearly every geometric measure (Fig. 7h–k) over the first four sessions. (Indeed, we find similar geometric trends for the training dynamics of the multi-task MLPs around initialization (Supplementary Fig. 4).) Once the readout and behavioral errors begin to plateau at session four, we find that the ensuing trends are consistent with our predictions for optimal representations: the dimensionality measures and SSF increase over learning, whereas the neural–latent correlation decreases. In other words, once the network has formed a representation from which task variables can be efficiently decoded, the trends in the geometry align with the behavior we found for optimal representations over learning.

Discussion

In this work, we analyzed a model of learning multiple tasks that share a common latent structure. We showed that the generalization error in this model decomposes into four terms summarizing the correlation structures, factorization and dimension of the neural activity. Our work complements previous studies of multitask learning[46–48] and neural correlations[25–27,29,49–52] by analytically describing how noise correlation structures interact with signal correlations, dimensionality and signal factorization to collectively determine readout generalization performance.

We leveraged these analytical results to determine which geometries minimize the generalization error, given a fixed latent structure and number of available training samples. In addition to being disentangled, we found that the task-relevant subspace of optimal neural codes shifts from being low to high dimensional as the number of available samples increases. This reflects a strategy of compressing less useful information when data are scarce and expanding it when data are abundant. Note that, in a given dataset, low-variance latent variables may be of greater interest than those with high variance. To apply our theory to such situations, one can bias the linear shatterings to emphasize low-variance latent variables. As discussed at the end of Supplementary Information section 1, this is equivalent to rescaling individual latent variables.

It is interesting to compare these findings to previous work reporting that higher-dimensional representations of individual object classes are preferred for invariant object recognition in the few-shot learning setting[24]. Although the contribution to the error of the neural dimension scales as $1/p$ in our formula just as in this work, the additional contribution to the error of the SSF and SNF terms, together with the constraint of the covariance matrix remaining positive semi-definite, leads to optimal representations with lower dimension in the few-shot regime. These results highlight the fact that different geometric terms may compete with each other in ways that cannot be directly read off from generalization error equations when there are additional constraints imposed on the system.

These optimal coding trends may hold for other forms of readout, beyond the Hebbian learning rule considered here. In particular, theories of linear least squares estimation suggest that similar phenomena may occur for regression tasks with more complex readouts[53–56]. Furthermore, we compared our theoretically calculated Hebbian readout error to the empirical generalization error of a linear support vector classifier (SVC) trained on the same tasks. Across every analysis reported in this paper, we found a strong correlation between SVC and Hebbian errors, particularly in the brain data, suggesting that the geometric terms described here are also informative of the generalization errors of other linear readouts (Supplementary Fig. 3). A related line of work has considered the role of neural dimension in random pattern separation[57–59]. In these settings, the optimal dimension can decrease when neural noise increases, and future work can examine whether a similar relation holds in the multitask learning setting considered here.

To test our theory, we first carried out a series of analyses on artificial networks. Across both networks, we found a tension between the neural dimension and correlation: layers that yielded a sharp increase in the dimension typically did so at the cost of the correlation and vice versa. These results illustrate a tradeoff between a population code's dimensionality and the amount of correlated variability between single units and task variables. In light of previous work demonstrating the interaction between the baseline firing rate of a population and neuronal correlations, it would be interesting to analyze how population sparsity and single-unit response reliability interact with the neural dimension and our neural–latent correlation statistic in recordings or models of tuned sensory neurons[28,29].

We next applied our theory to electrophysiological recordings from macaque V4 and IT[43] as well as navigational representations in rat CA1 and PFC[45]. Both of these analyses were carried out using multi-unit recordings, and, thus, these trends should be interpreted with caution. Geometric measures calculated from these recordings, such as the neural dimension, may not necessarily be indicative of the dimension of the entire population within a region. For example, the local similarity of tuning properties within certain visual regions may artificially decrease dimension, and systematic differences in recording quality between regions could artificially inflate the SNF estimates. Although these caveats should be kept in mind, we found a sharp increase in the SNF from the pixels and V4 to IT, suggesting that latent-unrelated variability may become increasingly orthogonal to the coding directions through the ventral stream. Indeed, orthogonalization of neural variability was previously observed in early visual areas, and it would be interesting to study the evolution of this phenomenon across the cortical hierarchy[25,34].

Turning to the rat CA1 and PFC analyses, we found that the SSF and the dimension of the subspace containing task-relevant navigational information increase over the course of learning. Furthermore, we found that the correlation first increases and then decreases. These trends are consistent with the predictions that we derived by considering optimal neural representations. Together, these results demonstrate the applicability of our theory to probing neural representations across brain regions and through learning.

In this work, we analyzed linear readout applied to tasks that came from shattering a continuous latent space. Determining which geometric measures are relevant for common nonlinear decoders remains an important open problem. Furthermore, our modeling framework makes several distributional assumptions, and future work could extend this theory to consider distributions of latents and tasks that more closely mirror common experimental settings. One could repeat our general calculation while restricting the distribution of task vectors to a handful of relevant directions or restricting the distribution of latent variables to be fixed to discrete values. This would be particularly interesting in settings where only particular groups of tasks or latent variable values are relevant for an animal to consider. As a further extension, it would be interesting to extend this work to consider the generalization error on a test set with distinct statistical properties. We plan to pursue these lines of research in subsequent work.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-025-02183-y.

References

1. Courellis, H. S. et al. Abstract representations emerge in human hippocampal neurons during inference. *Nature* **632**, 841–849 (2024).
2. Nogueira, R., Rodgers, C. C., Bruno, R. M. & Fusi, S. The geometry of cortical representations of touch in rodents. *Nat. Neurosci.* **26**, 239–250 (2023).
3. Johnston, W. J., Fine, J. M., Yoo, S. B. M., Ebitz, R. B. & Hayden, B. Y. Semi-orthogonal subspaces for value mediate a tradeoff between binding and generalization. *Nat. Neurosci.* **27**, 2218–2230 (2024).
4. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967 (2020).
5. Higgins, I. et al. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat. Commun.* **12**, 6456 (2021).
6. Chang, L. & Tsao, D. Y. The code for facial identity in the primate brain. *Cell* **169**, 1013–1028 (2017).
7. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A. & Summerfield, C. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron* **110**, 1258–1270 (2022).
8. Johnston, W. J. & Fusi, S. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nat. Commun.* **14**, 1040 (2023).
9. Srinath, R., Ni, A. M., Marucci, C., Cohen, M. R. & Brainard, D. H. Orthogonal neural representations support perceptual judgements of natural stimuli. *Sci. Rep.* **15**, 5316 (2025).
10. Higgins, I. et al. Towards a definition of disentangled representations. Preprint at https://doi.org/10.48550/arXiv.1812.02230 (2018).
11. Behrens, T. E. et al. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
12. Aronov, D., Nevers, R. & Tank, D. W. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature* **543**, 719–722 (2017).
13. Knudsen, E. B. & Wallis, J. D. Hippocampal neurons construct a map of an abstract value space. *Cell* **184**, 4640–4650 (2021).
14. Nieh, E. H. et al. Geometry of abstract learned knowledge in the hippocampus. *Nature* **595**, 80–84 (2021).
15. Bao, X. et al. Grid-like neural representations support olfactory navigation of a two-dimensional odor space. *Neuron* **102**, 1066–1075 (2019).
16. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
17. Muhle-Karbe, P. S. et al. Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron* **111**, 3885–3899 (2023).
18. Urai, A. E., Doiron, B., Leifer, A. M. & Churchland, A. K. Large-scale neural recordings call for new insights to link brain and behavior. *Nat. Neurosci.* **25**, 11–19 (2022).
19. Chung, S. & Abbott, L. Neural population geometry: an approach for understanding biological and artificial neural networks. *Curr. Opin. Neurobiol.* **70**, 137–144 (2021).
20. Chung, S., Lee, D. D. & Sompolinsky, H. Classification and geometry of general perceptual manifolds. *Phys. Rev. X* **8**, 031003 (2018).
21. Cohen, U., Chung, S., Lee, D. D. & Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* **11**, 746 (2020).
22. Wakhloo, A. J., Sussman, T. J. & Chung, S. Linear classification of neural manifolds with correlated variability. *Phys. Rev. Lett.* **131**, 027301 (2023).
23. Chou, C. N. et al. Geometry linked to untangling efficiency reveals structure and computation in neural populations. Preprint at *bioRxiv* https://doi.org/10.1101/2024.02.26.582157 (2025).
24. Sorscher, B., Ganguli, S. & Sompolinsky, H. Neural representational geometry underlies few-shot concept learning. *Proc. Natl Acad. Sci. USA* **119**, e2200800119 (2022).
25. Rumyantsev, O. I. et al. Fundamental bounds on the fidelity of sensory cortical coding. *Nature* **580**, 100–105 (2020).
26. Moreno-Bote, R. et al. Information-limiting correlations. *Nat. Neurosci.* **17**, 1410–1417 (2014).
27. Panzeri, S., Moroni, M., Safaai, H. & Harvey, C. D. The structures and functions of correlations in neural population codes. *Nat. Rev. Neurosci.* **23**, 551–567 (2022).
28. Cohen, M. R. & Kohn, A. Measuring and interpreting neuronal correlations. *Nat. Neurosci.* **14**, 811–819 (2011).
29. Montijn, J. S. et al. Strong information-limiting correlations in early visual areas. Preprint at *bioRxiv* https://doi.org/10.1101/842724 (2019).
30. Russo, A. A. et al. Motor cortex embeds muscle-like commands in an untangled population response. *Neuron* **97**, 953–966 (2018).
31. Perich, M. G., Gallego, J. A. & Miller, L. E. A neural population mechanism for rapid learning. *Neuron* **100**, 964–976 (2018).
32. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).
33. She, L., Benna, M. K., Shi, Y., Fusi, S. & Tsao, D. Y. T. Temporal multiplexing of perception and memory codes in IT cortex. *Nature* **629**, 861–868 (2024).
34. Zhong, L. et al. Unsupervised pretraining in biological neural networks. *Nature* **644**, 741–748 (2025).
35. Froudarakis, E. et al. Object manifold geometry across the mouse cortical visual hierarchy. Preprint at *bioRxiv* https://doi.org/10.1101/2020.08.20.258798 (2020).
36. Matthey, L., Higgins, I., Hassabis, D. & Lerchner, A. dSprites: Disentanglement testing Sprites dataset. *GitHub* https://github.com/deepmind/dsprites-dataset/ (2017).
37. Gardner, E. & Derrida, B. Three unfinished works on the optimal storage capacity of networks. *J. Phys. A Math. Gen.* **22**, 1983 (1989).
38. Goldt, S., Mézard, M., Krzakala, F. & Zdeborová, L. Modeling the influence of data structure on learning in neural networks: the hidden manifold model. *Phys. Rev. X* **10**, 041044 (2020).
39. Loureiro, B. et al. Learning curves of generic features maps for realistic datasets with a teacher-student model. In *NIPS'21: Proc. 35th International Conference on Neural Information Processing Systems* 18137–18151 (Curran Associates, 2021).
40. Engel, A. & Van den Broeck, C. *Statistical Mechanics of Learning* (Cambridge Univ. Press, 2005).
41. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
42. Lauer, J. et al. Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nat. Methods* **19**, 496–504 (2022).
43. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.* **35**, 13402–13418 (2015).
44. Hong, H., Yamins, D. L., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).
45. Tang, W., Shin, J. D. & Jadhav, S. P. Geometric transformation of cognitive maps for generalization across hippocampal-prefrontal circuits. *Cell Rep.* **42**, 112246 (2023).
46. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).

47. Driscoll, L., Shenoy, K. & Sussillo, D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nat. Neurosci.* **27**, 1349–1363 (2024).
48. Sagiv, Y., Musslick, S., Niv, Y. & Cohen, J. Efficiency of learning vs. processing: towards a normative theory of multitasking. In *Proc. 40th Annual Meeting of the Cognitive Science Society* (Cognitive Science Society, 2018).
49. Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).
50. Bartolo, R., Saunders, R. C., Mitz, A. R. & Averbeck, B. B. Information-limiting correlations in large neural populations. *J. Neurosci.* **40**, 1668–1678 (2020).
51. Cohen, M. R. & Maunsell, J. H. Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.* **12**, 1594–1600 (2009).
52. Ni, A. M., Ruff, D. A., Alberts, J. J., Symmonds, J. & Cohen, M. R. Learning and attention reveal a general relationship between population activity and behavior. *Science* **359**, 463–465 (2018).
53. Bordelon, B., Canatar, A. & Pehlevan, C. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning* 1024–1034 (PMLR, 2020).
54. Canatar, A., Bordelon, B. & Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nat. Commun.* **12**, 2914 (2021).
55. Canatar, A., Feather, J., Wakhloo, A. J. & Chung, S. A spectral theory of neural prediction and alignment. In *Thirty-Seventh Conference on Neural Information Processing Systems* 47052–47080 (Curran Associates, 2023).
56. Saxe, A. M., McClelland, J. L. & Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proc. Natl Acad. Sci. USA* **116**, 11537–11546 (2019).
57. Babadi, B. & Sompolinsky, H. Sparseness and expansion in sensory representations. *Neuron* **83**, 1213–1226 (2014).
58. Litwin-Kumar, A., Harris, K. D., Axel, R., Sompolinsky, H. & Abbott, L. Optimal degrees of synaptic connectivity. *Neuron* **93**, 1153–1164 (2017).
59. Muscinelli, S. P., Wagner, M. J. & Litwin-Kumar, A. Optimal routing to cerebellum-like structures. *Nat. Neurosci.* **26**, 1630–1641 (2023).

Methods

Model of multitask learning

We model a setting in which an agent learns a set of binary classification tasks by performing linear readout on a set of neural activity vectors. Formally, we assume that each stimulus is associated with a d-dimensional latent vector $\mathbf{z}_\mu \in \mathbb{R}^d$, with $1 \le \mu \le p$ denoting the sample index. The labels for a specific task are formed by shattering the latent space using a hyperplane with a normal vector, $\mathbf{T}$. Thus, the binary classification task labels, y_μ, satisfy the relation, $y_\mu = \text{sign}(\mathbf{T} \cdot \mathbf{z}_\mu)$, so that each $\mathbf{T}$ vector defines a specific classification task. Associated with each latent, we also consider n-dimensional neural activity patterns, $\mathbf{x}_\mu \in \mathbb{R}^n$. From these data, the agent then forms predictions of new data points using a supervised Hebbian readout rule[40]. More precisely, given a new stimulus associated with latent variables $\mathbf{z}_+$ and firing rates $\mathbf{x}_+$, the agent forms a prediction $\hat{y}_+$ using the rule

$$\hat{y}_+ = \text{sign}(\mathbf{w} \cdot \mathbf{x}_+), \quad \mathbf{w} = \frac{1}{p}\sum_\mu y_\mu \mathbf{x}_\mu. \tag{3}$$

When the labels are balanced, this corresponds to using the difference in the mean activities for positively and negatively labeled examples (Fig. 1f; see Supplementary Fig. 3 for results on SVC classifiers). We evaluate how well the neural code can support downstream classification by calculating the generalization error of these predictions, averaged across different tasks (that is, different $\mathbf{T}$ vectors), although we also calculate the error for a fixed $\mathbf{T}$ along the way (Supplementary Information section 1).

Although the generalization error for arbitrary distributions of neuronal activities and latent variables is analytically intractable, we show that, in many cases, the error only depends on a few key statistics. To do this, we draw from work in deep learning theory describing a Gaussian equivalence principle (GEP), which states that the generalization error of linear readouts trained on complex distributions can, in many cases, be well approximated by studying simpler Gaussian models[38,39,60,61]. We analytically calculate the generalization error using a Gaussian model and show empirically that our theory accurately predicts the generalization error of the linear readout rule when applied to more complex settings. Formally, we assume that each pair of neural responses and latent variables ($\mathbf{x}_\mu$, $\mathbf{z}_\mu$) are jointly zero mean Gaussians with covariance matrices:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{\Psi}, \quad \mathbb{E}[\mathbf{x}\mathbf{z}^\top] = \mathbf{\Phi}, \quad \mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \mathbf{\Omega}. \tag{4}$$

We can see that $\mathbf{\Psi}$ describes the neuron–neuron covariances; $\mathbf{\Phi}$ contains the covariances between single-unit responses and the latent variables; and $\mathbf{\Omega}$ describes covariances between latent variables in the dataset. This model corresponds to a variant of the popular student–teacher model[37,39]. Insofar as the GEP holds, the neuronal and latent (cross-)covariances fully specify the generalization error of the linear readout.

In our analyses below, we enforce the zero mean condition by centering the $\mathbf{x}$ and $\mathbf{z}$ data matrices. Note that because the $\mathbf{T}$ vectors are chosen randomly from a Gaussian distribution, the latent covariance $\mathbf{\Omega}$ determines which directions in the latent space are, on average, most informative of the task labels. Directions in the latent space that have a significant amount of variance are, on average, more informative of the task labels in this setup (Supplementary Information section 4; note that one can simply whiten latent variables when applying this framework to data to circumvent this feature).

Decomposition of generalization error across tasks

We prove a formula for the generalization error of the linear readout, averaged across different binary classification tasks (that is, different $\mathbf{T}$ vectors). To do so, we consider the limiting case in which the number of neurons, latent dimensions and training samples are all large and of similar size. Furthermore, we assume that the covariance matrices given above satisfy certain spectral properties (Supplementary Information section 1). Under these assumptions, we show that the average generalization error, E_g, given a training dataset of size p, is a decreasing function of the four geometric terms. Formally, we have

$$E_g = \frac{1}{\pi}\tan^{-1}\left(\sqrt{\frac{\pi}{2pc^2\text{PR}(\mathbf{\Psi})} + \frac{1}{f} + \frac{1}{s} - 1}\right), \tag{5}$$

where we have introduced the total neural–latent correlation c, the SSF f, the SNF s and the dimension of the population responses as measured by the participation ratio, PR($\mathbf{\Psi}$). These terms correspond to statistics of the covariance matrices specified above:

$$c = \frac{\text{Tr}(\mathbf{\Phi}\mathbf{\Phi}^\top)}{\text{Tr}(\mathbf{\Psi})\text{Tr}(\mathbf{\Omega})} \tag{6}$$

$$\text{PR}(\mathbf{\Psi}) = \frac{\text{Tr}(\mathbf{\Psi})^2}{\text{Tr}(\mathbf{\Psi}^2)} \tag{7}$$

$$f = \frac{\text{Tr}(\mathbf{\Phi}\mathbf{\Phi}^\top)^2}{\text{Tr}(\mathbf{\Omega})\text{Tr}(\mathbf{\Phi}^\top\mathbf{\Phi}\mathbf{\Omega}^{-1}\mathbf{\Phi}^\top\mathbf{\Phi})} \tag{8}$$

$$s = \frac{\text{Tr}(\mathbf{\Phi}\mathbf{\Phi}^\top)^2}{\text{Tr}(\mathbf{\Omega})\text{Tr}(\mathbf{\Phi}^\top(\mathbf{\Psi} - \mathbf{\Phi}\mathbf{\Omega}^{-1}\mathbf{\Phi}^\top)\mathbf{\Phi})} \tag{9}$$

Because the function $F(w) = \frac{1}{\pi}\tan^{-1}\left(\sqrt{w-1}\right)$ is strictly increasing, we can see that the error is a decreasing function of each of these geometric terms. Note that, in writing equation (5), we have separated a term that depends on the number of training samples, $1/[c^2\text{PR}(\mathbf{\Psi})]$, from a term that is independent of the sample size, $1/f + 1/s$. Thus, we can see that the correlation and dimension terms become less important as p grows.

We give a detailed discussion motivating the mathematical definitions of these terms in Supplementary Information section 3. To summarize, the definition of c corresponds to a normalized sum of squared covariances between all single units and latent variables and is, thus, a multidimensional generalization of the Pearson correlation coefficient between neural responses and latent variables. The participation ratio is a standard measure of the dimension of a neural representation (see, for example, refs. 57,58). Turning to the factorization measures, the term f measures the overall degree to which independent latent variables are represented along uncorrelated coding directions (Supplementary Fig. 2), and the term s measures the amount of noise that lies along the signal directions. To see this, first note that the neural noise is described by the noise covariance matrix, $\text{cov}(\mathbf{x}|\mathbf{z}) = \mathbf{\Psi} - \mathbf{\Phi}\mathbf{\Omega}^{-1}\mathbf{\Phi}^\top$, which appears in the definition of s above. Under the Gaussian model, this matrix measures the covariability of signal-unrelated, trial-to-trial fluctuations in pairs of units. On the other hand, the matrix $\text{cov}(\mathbf{x}) - \text{cov}(\mathbf{x}|\mathbf{z}) = \mathbf{\Phi}\mathbf{\Omega}^{-1}\mathbf{\Phi}^\top$ measures the signal-related covariability of pairs of units, motivating its appearance in the definition of f[27,28]. The terms in the denominators of s and f give the projections of neural coding directions onto signal and noise subspaces, whereas the term $\text{Tr}(\mathbf{\Phi}\mathbf{\Phi}^\top)^2$ acts as a normalization. Given that 'neural noise' likely includes variability that is related to variables that are not measured experimentally, we additionally describe, in the Supplementary Information, how these two factorization terms can be collapsed into a single factorization term.

Gaussian simulations

To test our theory on data points that violate the assumptions of our theory, we began by sampling from a finite Gaussian model. Specifically,

we drew latent vectors, $\mathbf{z}_\mu$, from a multivariate Gaussian distribution whose covariance matrix had eigenvalues that decayed as a power law with rate α. Specifically, we set: $\omega_i = 5i^{-\alpha}$, where α is the power law of the spectrum, and ω_i is the i-th eigenvalue of the latent covariance. The neural responses were then given by the formula $\mathbf{x}_\mu = \mathbf{A}\mathbf{z}_\mu$ for a random $n \times d$ Gaussian matrix $\mathbf{A}$ with i.i.d. elements or by applying a whitening transform, $\mathbf{x}_\mu = \mathbf{\Omega}^{-1/2}\mathbf{z}_\mu$. We set $n = 80$ and $d = 40$ for these simulations. For a fixed training set, we sampled $N_{\text{task}} = 300$ task $\mathbf{T}$ vectors and calculated the generalization error across all tasks using a set of new latent variables. Finally, we averaged over this entire procedure 30 times to generate the markers in Supplementary Fig. 1.

Optimal codes

We derive which neuronal codes achieve the lowest generalization error, given a fixed number of samples to train on and a fixed latent structure. Specifically, we calculate which neuron–neuron and neuron–latent covariance matrices $\mathbf{\Psi}$ and $\mathbf{\Phi}$ achieve the lowest multitask generalization error, given a fixed latent covariance $\mathbf{\Omega}$ and training set size p (Supplementary Information). Note that, because we generate binary classification tasks by shattering the latent space uniformly, latent variables with more variance are, on average, more informative of the task labels than variables with low variance (Supplementary Information section 4). We perform this calculation by optimizing equation (5) with respect to $\mathbf{\Psi}$ and $\mathbf{\Phi}$ subject to the constraint that the entire covariance matrix between neurons and latents be positive semi-definite, a necessary condition for the code to be realizable. Using this approach, we find that the left and right singular vectors of $\mathbf{\Phi}$ are the eigenvectors of $\mathbf{\Psi}$ and $\mathbf{\Omega}$, respectively, for the optimal code. This shows that independent latent variables map directly onto uncorrelated directions in the firing rate space and that the principal components of the latent variables directly map onto the principal components of the neural activity (Supplementary Information). Furthermore, we obtain the following simple formula for the eigenvalues of the optimal neural code. Denoting the eigenvalues of $\mathbf{\Psi}$ and $\mathbf{\Omega}$ as ψ_i and ω_i, respectively, we have up to a permutation symmetry:

$$\psi_i = C\frac{\omega_i}{2p\omega_i + \pi\sum_k \omega_k}, \tag{10}$$

where C is an arbitrary constant. As p grows, we can see that the spectrum becomes flatter, reflecting the expansion strategy, whereas, as p shrinks, the spectrum decays faster and faster, reflecting the fact that less informative directions are compressed in the state space (Fig. 6c).

To validate our calculation, we numerically calculated the optimal code by optimizing on the space of positive semi-definite matrices. Because the full covariance matrix is positive semi-definite, there must exist matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ such that

$$\mathbf{L} = \begin{pmatrix} \mathbf{\Omega}^{1/2} & 0 \\ \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix}, \quad \mathbf{L}\mathbf{L}^T = \begin{pmatrix} \mathbf{\Omega} & \mathbf{\Phi}^T \\ \mathbf{\Phi} & \mathbf{\Psi} \end{pmatrix} \tag{11}$$

The space of possible $\mathbf{X}$ matrices is unconstrained, so we simply optimize equation (5) with respect to the $\mathbf{X}_i$ matrices and calculate $\mathbf{\Psi}$ and $\mathbf{\Phi}$ after the fact.

MLP experiments

We used random and trained MLPs to test several predictions from our theory using explicitly non-Gaussian artificial neural response data. To generate these data, we first sampled a set of $d = 40$ dimensional latent variables from a multivariate Gaussian distribution with eigenvalues $\omega_k = k^{-0.2}$. For these experiments, we sampled 5×10^5 latent variables independently from this distribution as a training dataset. Using these latent variables, we generated a set of $N_t = 500$ tasks by randomly shattering the latent space. Latent variables were then passed through a three-layer perceptron with randomly initialized weights. We chose the size of each intermediate layer to be twice the size of the previous one to ensure that the dimension of the representation did not decrease. Each layer was composed of a linear transform, followed by a ReLU nonlinearity (see Supplementary Figs. 5 and 6 for results using a tanh nonlinearity and Supplementary Fig. 7 for results using trained networks with an expanding layer structure). After sampling the task labels and passing the latents through the random MLP, we trained a three-hidden-layer MLP to predict task labels from the random MLP responses. The trained network was composed of linear-batchnorm-relu blocks, and we used the Adam optimizer to train the network[62] through a single epoch. This setup corresponds to a multitask version of the hidden manifold modeling framework studied in deep learning theory[38].

We evaluated the generalization error through layers and training using a new set of latents and tasks. Specifically, we sampled 10^3 latent variables from the same distribution as well as a new set of 300 randomly selected tasks. To calculate the theoretical generalization error, we then used the network representations of these new latent variables to calculate the geometric metrics and evaluate the theoretical generalization errors using equation (5) with $p = 300$. To evaluate the empirical generalization error, we randomly split the new set of latents into a set of 300 training points and 700 test points. This process was repeated to calculate the average test error of the Hebbian rule across each layer and timepoint in training. Finally, we averaged over train/test splits to generate the markers shown in Fig. 3.

Analysis of pose estimation network

To study the applicability of our theory to deep neural networks trained on complex data, we studied pose estimation networks trained using the DeepLabCut framework (version 2.3.10)[41,63]. We trained a pose estimation network based on a ResNet50 backbone[64] using the default parameters of the DeepLabCut package for a total of 30,000 iterations. We trained this network on the 'parenting mouse' benchmark dataset[42], which consists of 542 labeled frames (70% used for training the network and 30% for computing generalization error) containing an adult mouse with 12 body markers and two pups with five body markers each. We found that the test error was: 11.0 pixels, train: 3.1 pixels (image size varied but was a minimum of 704×480).

For simplicity, for all further analysis we examined only the 12 body markers of the adult mouse ($N = 1$) and only the 259 frames of the benchmark dataset in which all 12 of these markers were visible. Our $d = 24$ dimensional latent variables $\mathbf{z} \in \mathbb{R}^{24}$ consisted of the (x, y) coordinates (in pixels, normalized to an image size of 563×384) of the 12 body markers of the adult mouse. We also shifted the latents to have zero mean but did not rescale them. We extracted internal representations from the pose estimation network immediately before and immediately after each ReLU nonlinearity, resulting in a total of 33 different internal representations. For each of these 33 internal representations, we performed 20 random projections down to a fixed dimensionality of $n = 100$ to aid comparisons between layers of different sizes. All reported multitask errors and geometric measures were averaged over these 20 random projections. The internal representations $\mathbf{x} \in \mathbb{R}^{100}$ were also shifted to have zero mean but were not rescaled.

Macaque analyses

We drew from a publicly available dataset containing multi-unit recordings from V4 and IT taken from two monkeys[43] via the Brain-Score package[65,66]. These recordings were taken as the monkeys viewed visual stimuli as described in ref. 43 and contained 88 V4 and 168 IT neural sites, although we reproduce all results projecting IT and pixel responses down to 88 randomly chosen dimensions (Supplementary Fig. 8). The stimuli for this task were drawn from a generative model in which a total of 64 objects coming from eight categories were displayed against varying backgrounds. These images were generated by varying $d = 6$ continuous latent variables that controlled the object size,

angle and position. Latent variables were drawn randomly from a uniform distribution.

We tested the linear decodability of these latent variables from the neural firing rates using the scheme described in Methods 'Model of multi-task learning'. For the raw pixels, we first carried out a Gaussian random projection onto a 500-dimensional space before applying this scheme. Specifically, for each of the eight object category types, we formed task labels for $N_t = 300$ tasks by randomly shattering the latent space. There were 320 examples per category type. We z-scored the latent variables to ensure that there was no especially informative variable. We then formed predictions using the supervised Hebbian readout described in Methods 'Model of multi-task learning' after mean centering the neuronal firing rates. This procedure was repeated over 15 different train/test data splits. To generate the results in Fig. 5, we then averaged the generalization error across the N_t classification tasks, data splits and image categories. To test the significance of the differences across groups, we used paired-sample t-tests across each of the eight category types and adjusted using a false discovery rate (FDR) criterion using the statsmodels package (Supplementary Table 1). In Supplementary Information, we also show the geometry and generalization errors for each of the categories individually (Supplementary Figs. 9 and 10) as well as the error obtained by pooling all categories together (Supplementary Fig. 11).

Spatial representations in rat PFC and CA1

Rat data processing. We drew from a dataset reported in ref. 45, which is publicly available[67]. In brief, rats learned to perform a continuous alternation task in a W-maze that required them to travel to the center of the maze and then to the opposite end from the one they began from (Fig. 7b). Learning occurred over eight sessions lasting approximately 20 minutes each. The position of each animal was measured at 30 fps using an LED light attached to the animal's head. Neural recordings were collected from multi-tetrodes implanted in dorsal CA1 and medial PFC, and we used the preprocessed spike-sorted data provided in ref. 67.

For the analyses reported here, we ran several additional preprocessing steps. First, the activity of putative single units was binned at 500 ms. To be included in the analyses for a given session, a unit had to achieve a mean firing rate of 0.1 Hz during that session. To align the behavioral position data to spike bins, we used Nadaraya–Watson kernel smoothing with a Gaussian kernel to interpolate the reported position information. We similarly calculated x and y velocities using the displacements in the animal's position, followed by Nadaraya–Watson interpolation. For all analyses here, we only considered timepoints when the animal had an overall speed of at least 4 cm s^{-1} (ref. 45).

Generalized linear model analysis. In Fig. 7g, we showed that the total neural dimensionality decreases over the initial four sessions. This initial decrease in dimension can come from reductions in navigation-unrelated directions, signal directions or a mix of the two. In the main text, we argued that this initial drop in the total dimension comes from a compression of navigation-unrelated directions. To support this interpretation, we showed that the dimension of the subspace containing task-relevant information (that is, the dimension of the 'signal subspace') increases monotonically over learning, even during those initial four sessions (Fig. 7h). This suggests that any drop in the total dimensionality comes from a compression of navigation-unrelated directions. The monotonic increase of the task-relevant dimensionality through learning is in line with our predictions regarding optimal representations. We begin by giving a brief summary of this analysis, before describing the details.

To estimate the dimension of the subspace containing task-relevant information, we require a method to average out those components of neuronal activity that are unrelated to the task—in this case, unrelated to the navigational state of the animal. Formally, one requires

$$\mathbf{v}(\mathbf{z}) := \mathbb{E}_{\mathbf{x}}[\mathbf{x}|\mathbf{z}], \tag{12}$$

where $\mathbf{z}$ is the vector of latents describing the navigational state of the animal and x are the firing rates. To estimate this conditional expectation, we used Poisson generalized linear models (GLMs). The dimension of the task-related subspace follows from the participation ratio of $\mathbf{v}(\mathbf{z})$:

$$\text{dim. task subspace} = \text{PR}(\text{cov}(\mathbf{v})). \tag{13}$$

This is the quantity we plot in Fig. 7h, after z-scoring $\mathbf{v}$.

We used Poisson GLMs to predict single-unit spike counts from position and velocity behavioral variables. To do this, we used isotropic Gaussian basis functions to tile the two-dimensional maze coordinates and one-dimensional Gaussian basis functions to tile the x and y velocity information of the animal. We then fit coefficients that predicted spike count probabilities using a linear combination of the values of these basis functions that was then fed through an exponential link function. To accommodate idiosyncrasies in each animal's movement preferences, we adaptively chose the locations of each basis function separately for each animal. Basis function centers were determined by binning the distribution of x and y coordinates for each animal across all sessions and using the center of bins that contained data from more than 20 time bins. To accommodate for the possibility of trajectory-specific place maps, we fit separate coefficients for these basis functions for each trajectory type (inbound versus outbound). We formed basis functions for the velocity variables by separately tiling each one-dimensional space of x and y velocities using a similar adaptive binning procedure. All GLM analyses were carried out using the PoissonRegressor scikitlearn class.

All GLM model hyperparameters were set by cross-validation. Our models had an $\ell 2$ regularization hyperparameter, together with four hyperparameters governing the distribution of basis functions: the bin sizes for the velocity variables, the bin sizes for the position variables and the variances of the corresponding basis functions. We set $\ell 2$ regularization parameters separately for each unit using 10-fold cross-validation over 10 possible regularization parameter values. We chose a single set of basis function hyperparameters across rats. To choose the basis hyperparameters, we searched over a small grid and chose the combination that achieved the smallest cross-validated generalization error. The cross-validated error for each single-unit model was calculated using the default percentage of deviance explained. The average cross-validated population D^2 across all sessions and rats was 0.29 in PFC and 0.25 in CA1, and the standard deviation of the session-averaged population D^2 across rats was 0.022 in PFC and 0.1 in CA1. Single-unit models that performed worse than a constant firing rate null model ($D^2 < 0$) were discarded and replaced by the null model—that is, were treated as having an average firing rate that did not depend on the latent variables.

We used these GLMs to estimate the dimensionality of the component of neural activity related to navigational signals. Just as with previous analyses, we calculated this dimensionality using the participation ratio of the covariance matrix—in this case, the covariance of expected neuronal firing rates, $\mathbf{v}(\mathbf{z}) = \mathbb{E}_{\mathbf{x}}[\mathbf{x}|\mathbf{z}]$. The covariance matrix was given by $\mathbb{E}_{\mathbf{z}}\left[\mathbf{v}(\mathbf{z})\mathbf{v}(\mathbf{z})^{\top}\right] - (\mathbb{E}_{\mathbf{z}}\mathbf{v}(\mathbf{z}))(\mathbb{E}_{\mathbf{z}}\mathbf{v}(\mathbf{z}))^{\top}$, and the expectation over $\mathbf{z}$ was calculated using the empirical distribution of latent variables within a given session. (Note that we z-scored the $\mathbf{v}$ vector to match our treatment of the raw neural data as described below).

Readout analysis. Before carrying out the linear readout analysis of the spiking data, we normalized both behavioral variables and spike count data. We centered each unit using its mean from a session and divided by the unit's standard deviation within a session plus a small regularizer of 0.02 Hz. This regularizer was added to avoid massively overweighting very sparse units, and modifying its precise value did not significantly affect our results. Within the multitask learning

framework considered here, the importance of a given latent variable depends on its variance ('Optimal representation of latent variables' and Supplementary Information section 4). Hence, we z-scored the behavioral variables so that all position and velocity variables would be weighted equally when forming random binary classification tasks. Finally, because each rat and session had different numbers of trials and usable units, we calculated the geometric terms from random subsets of neurons and trials. To use all available rat data while balancing across rats and sessions, we used 19 PFC units and 24 CA1 units. For each random neuronal subset, we chose a random subset of 500 trials ($p = 300$ training points) to use for the readout and geometric analysis. All quantities reported in the main text and the Supplementary Information correspond to an average over 500 such subsets.

To test the significance of our trends, we fit quadratic regression models with mixed-effect intercepts to each geometric measure over time. That is, a given geometric quantity y during session t for rat i was modeled as $y_{it} = \beta_i + \gamma t + \eta t^2$. In the Supplementary Information, we plot model fits and report both the parameter values and P values of these quantities, which we calculated using the statsmodels Python package (Supplementary Fig. 12 and Supplementary Table 2).

Statistics and reproducibility

We provide publicly available code to reproduce our results. We did not collect any data for this study, so no statistics were used to determine sample size. We excluded neurons with extremely low firing rates from the analyses presented in the 'Geometry of spatial representations in PFC and CA1 during learning'. We additionally excluded frames in which body markers were invisible in 'Disentangling in deep pose estimation networks' Finally, for the analyses in 'Predicting readout performance of macaque visual representations', we excluded noisy units, following the recommendations in the Brain-Score package[65,66].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All macaque data are publicly available from Brain-Score[65,66] at https://github.com/brain-score/vision. Rat data are available from the DANDI Archive[67]. Mouse pose data are available from DeepLabCut[41,63] at github.com/DeepLabCut/DeepLabCut.

Code availability

All code and results used in this paper are publicly available at https://github.com/awakhloo/population_geometry_optimal_coding.

References

60. Goldt, S. et al. The Gaussian equivalence of generative models for learning with shallow neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (eds Bruna, J. et al.) 426–471 (IEEE, 2016).
61. Montanari, A., Ruan, F., Saeed, B. & Sohn, Y. Universality of max-margin classifiers. Preprint at https://doi.org/10.48550/arXiv.2310.00176 (2023).
62. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://doi.org/10.48550/arXiv.1412.6980 (2017).
63. Nath, T. et al. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protoc.* **14**, 2152–2176 (2019).
64. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (IEEE, 2016).
65. Schrimpf, M. et al. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* **108**, 413–423 (2020).
66. Schrimpf, M. et al. Brain-Score: which artificial neural network for object recognition is most brain-like? Preprint at *bioRxiv* https://doi.org/10.1101/407007 (2018).
67. Shin, J. & Jadhav, S. Single day W-track learning (Version 0.240511.0307) [Data set]. *DANDI* https://doi.org/10.48324/dandi.000978/0.240511.0307 (2024).

Acknowledgements

We thank J. Feather, T. Yerxa, C.-H. Chou, A. Raha, C. Windolf and members of the Aronov laboratory for helpful discussions and comments on an earlier version of this manuscript. This work was funded by the Center for Computational Neuroscience at the Flatiron Institute of the Simons Foundation. S.Y.C. is also partially supported by a Sloan Research Fellowship, a Klingenstein-Simons Award and the Samsung Advanced Institute of Technology (under the project 'Next Generation Deep Learning: From Pattern Recognition to AI'). All experiments were performed on the high-performance computing cluster at the Flatiron Institute. A.J.W. was additionally supported by the National Institute of Mental Health of the National Institutes of Health under award number T32MH126036 and by the Gatsby Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.J.W. and S.Y.C. conceived the study. A.J.W. and W.S. derived mathematical results and carried out data analysis. A.J.W., W.S. and S.Y.C. developed the theory, interpreted the analysis results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-025-02183-y.

Correspondence and requests for materials should be addressed to SueYeon Chung.

Peer review information *Nature Neuroscience* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): SueYeon Chung

Last updated by author(s): Apr 19, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Code for simulations was written in Python 3.
Data analysis	Custom code for data analysis was written in Python 3 and is provided in the following DOI-minted repository: https://github.com/awakhloo/population_geometry_optimal_coding

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All macaque electrophysiology data is publicly available from the brainscore repository (https://github.com/brain-score/vision), rat data is available from the DANDI archive (https://dandiarchive.org/dandiset/000978?search=single+day+W&pos=1), and behavioral pose tracking data from DeepLabCut (https://github.com/DeepLabCut/DeepLabCut)

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The datasets we drew from with neural recordings contained samples from n=2 monkeys and n=8 rats. The behavioral dataset contained tracking data from N=1 mouse and a pup.
Data exclusions	For the monkey data, the creators of this dataset recommend excluding noisy units, and we followed their guidelines while carrying out our analyses. For the rat data, we excluded moments when the animal was stationary or walking at a speed less than 4 cm/s. We also excluded cells with a baseline firing rate of less than .1 Hz. For the behavioral data, we excluded frames in which any of the adult markers were not visible.
Replication	In the supplemental material, we reproduce our main results with the macaque using alternative groupings of the data as well as a random projection methodology to demonstrate that our main results are insensitive to irrelevant features of the stimulus.
Randomization	Randomization was not needed for this task.
Blinding	Macaques were presented images in a random order and were thus blind to the sequence of stimuli.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A

nature neuroscience

Supplementary information

https://doi.org/10.1038/s41593-025-02183-y

Neural population geometry and optimal coding of tasks with shared latent structure

In the format provided by the authors and unedited

Contents

1 Proof of generalization error formula

Here we state and prove our formula for the generalization error.

Theorem 1. *Let $x_{n,\mu} \in \mathbb{R}^n, z_{n,\mu} \in \mathbb{R}^d$ with $1 \leq \mu \leq p$ be a sequence of random variables drawn independently from a Gaussian distribution with mean zero and covariance matrices:*

$$C_n = \begin{pmatrix} \Psi_n & \Phi_n \\ \Phi_n^\top & \Omega_n \end{pmatrix}, \tag{1}$$

where C_n is a positive-definite matrix with sub-matrices $\Psi_n \in \mathbb{R}^{n \times n}, \Omega_n \in \mathbb{R}^{d \times d}, \Phi_n \in \mathbb{R}^{n \times d}$, and the ratios $d/n = \alpha$ and $p/n = \beta$ are held fixed. In addition, let $x_{n,+}, z_{n,+}$ be a pair of samples drawn independently from the same distribution, and let r_n be a sequence of d-dimensional teacher vectors drawn uniformly from the surface of the sphere of radius $\sqrt{d}$. Assume that there exist positive constants c_1, c_2 such that

$$\frac{c_1}{n} \leq \psi_{n,i}, \omega_{n,i} \leq \frac{c_2}{n}, \tag{2}$$

$$0 \leq \phi_{n,i} \leq \frac{c_2}{n}, \tag{3}$$

where we use lower-case Greek letters to denote the eigenvalues of Ψ_n and Ω_n and the singular values of Φ_n. Define the Hebbian readout $w := \frac{1}{p} \sum_\mu \mathrm{sgn}(\langle z_{n,\mu}, r_n \rangle) x_{n,\mu}$, and let $\Theta(\cdot)$ denote the Heaviside step function. Under these assumptions the generalization error is given by

$$\begin{aligned} &\mathbb{E}_{x_n, z_n, r_n} \left\{ \Theta(-\mathrm{sgn}(\langle r_n, z_{n,+} \rangle) \langle w, x_{n,+} \rangle) \right\} \\ &= \frac{1}{\pi} \tan^{-1} \left(\sqrt{\frac{\mathrm{Tr}(\Omega_n) \left[\frac{\pi}{p} \mathrm{Tr}(\Psi_n^2) \mathrm{Tr}(\Omega_n) + 2\mathrm{Tr}(\Phi_n^\top \Psi_n \Phi_n) \right]}{2\mathrm{Tr}(\Phi_n^\top \Phi_n)^2} - 1} \right) + O(n^{-1/2}). \end{aligned} \tag{4}$$

Proof. We start by defining the random variables

$$\gamma_n^\mu = \mathrm{sgn}(\langle z_{n,\mu}, r_n\rangle)\langle x_{n,+}, x_{n,\mu}\rangle, \tag{5}$$

so that the generalization error may be written

$$\mathbb{E}_{r_n}\mathbb{E}_{x_{n,+},z_{n,+}}\mathbb{E}_{x_{n,\mu},z_{n,\mu}}\Theta\bigg(-\mathrm{sgn}(\langle z_{n,+}, r_n\rangle)\sum_\mu \gamma_n^\mu\bigg). \tag{6}$$

Note that we have used Fubini's theorem to separate the expectations over the training and test points, as well as the teacher vector r_n.

Our goal is now to show that the distribution of the random variable $\sum_\mu \gamma_n^\mu$ is approximately Gaussian. This allows us to carry out the inner expectation in Eq. 6. We do this using the Berry-Esseen theorem. We can obtain the relevant moments needed to apply this theorem using the identity

$$\mathbb{E}_{p,q\sim\mathcal{N}(0,\Sigma)}\mathrm{sgn}(q)p = \sqrt{\frac{2}{\pi\sigma}}\kappa, \quad \Sigma := \begin{pmatrix}\zeta & \kappa\\ \kappa & \sigma\end{pmatrix}, \tag{7}$$

together with

$$\mathbb{E}_{h\sim\mathcal{N}(0,v)}|h|^3 = \frac{2\sqrt{2}}{\sqrt{\pi}}v^{3/2}. \tag{8}$$

The mean and variance are simply

$$\mathbb{E}_{x_{n,\mu},z_{n,\mu}}\gamma_n^\mu = \sqrt{\frac{2}{\pi r_n^\top\Omega_n r_n}}x_{n,+}^\top\Phi_n r_n, \tag{9}$$

$$\mathbb{E}_{x_{n,\mu},z_{n,\mu}}(\gamma_n^\mu - \mathbb{E}_{x_{n,\mu},z_{n,\mu}}\gamma_n^\mu)^2 = x_{n,+}^\top\Psi_n x_{n,+} - \frac{2(x_{n,+}^\top\Phi_n r_n)^2}{\pi r_n^\top\Omega_n r_n}. \tag{10}$$

It suffices to bound the third central moment as

$$\mathbb{E}_{x_{n,\mu},z_{n,\mu}}|\gamma_n^\mu - \mathbb{E}_{x_{n,\mu},z_{n,\mu}}\gamma_n^\mu|^3 \le 4\bigg(\frac{2\sqrt{2}}{\sqrt{\pi}}\big(x_{n,+}^\top\Psi_n x_{n,+}\big)^{3/2} + |\mathbb{E}_{x_{n,\mu},z_{n,\mu}}\gamma_n^\mu|^3\bigg). \tag{11}$$

We now define the following random variables:

$$\sigma_n := r_n^\top\Omega_n r_n, \quad \kappa_n := x_{n,+}^\top\Phi_n r_n, \quad \zeta_n := x_{n,+}^\top\Psi_n x_{n,+}, \quad \eta_n := z_{n,+}^\top r_n. \tag{12}$$

If we now let $F_n(s)$ denote the cumulative distribution function of the standardized sum

$$s = \frac{\sum_\mu (\gamma_n^\mu - \mathbb{E}_{x_{n,\mu}} \gamma_n^\mu)}{\sqrt{p(\zeta_n - \frac{2\kappa_n^2}{\pi \sigma_n})}} \tag{13}$$

and denote by $N(s)$ the standard normal c.d.f., the Berry-Esseen theorem [3] gives

$$\left| F_n(s) - N(s) \right| \leq \frac{K}{\sqrt{p}\big(1 - \frac{2\kappa_n^2}{\pi \sigma_n \zeta_n}\big)^{3/2}} \big(1 + (x_{n,+}^\top \Psi_n x_{n,+})^{-3/2} |\mathbb{E}_{x_{n,\mu}, z_{n,\mu}} \gamma_n^\mu|^3\big), \tag{14}$$

where K is a constant. We can see that the rightmost term contributes at most a constant factor from

$$\frac{|\mathbb{E}_{x_{n,\mu}, z_{n,\mu}} \gamma_n^\mu|^3}{(x_{n,+}^\top \Psi_n x_{n,+})^{3/2}} \leq \frac{M |x_{n,+}^\top \Phi_n r_n|^3}{(r_n^\top \Omega_n r_n)^{3/2} (x_{n,+}^\top \Psi_n x_{n,+})^{3/2}} \leq M' \frac{\phi_{n,\max}^3}{\omega_{n,\min}^{3/2} \psi_{n,\min}^{3/2}} = O(1), \tag{15}$$

for some positive constants M, M'. Next we show that the entire right hand side of Eq. (14) is $O(n^{-1/2})$ using the following Lemma.

Lemma 2. *Let C be a positive definite matrix with submatrices*

$$C = \begin{pmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{pmatrix}, \tag{16}$$

with $\Psi \in \mathbb{R}^{n \times n}, \Omega \in \mathbb{R}^{d \times d}$, and $\Phi \in \mathbb{R}^{n \times d}$. Then for all $h \in \mathbb{R}^n$ and $r \in \mathbb{R}^d$

$$\frac{(h^\top \Phi r)^2}{h^\top \Psi h r^\top \Omega r} \leq 1. \tag{17}$$

Proof. Let (x, z) be mean-zero jointly Gaussian vectors with covariance matrix C. It follows that

$$\begin{aligned} (h^\top \Phi r)^2 &= \big(\mathbb{E}[\langle x, h \rangle \langle r, z \rangle]\big)^2 && (18) \\ &\leq \mathbb{E}[\langle x, h \rangle^2] \mathbb{E}[\langle z, r \rangle^2] && (19) \\ &= (h^\top \Psi h)(r^\top \Omega r). && (20) \end{aligned}$$

□

We therefore obtain

$$\frac{2\kappa_n^2}{\pi \sigma_n \zeta_n} \leq \frac{2}{\pi}, \tag{21}$$

so that the right hand side of Eq. (14) is $O(n^{-1/2})$. Keeping only the leading order terms, we obtain the following expression for the generalization error

$$\frac{1}{2}\mathbb{E}_{r_n}\mathbb{E}_{x_n,z_n}\mathrm{erfc}\left(\frac{\mathrm{sgn}(\eta_n)\kappa_n}{\sqrt{\frac{\pi}{p}\sigma_n\zeta_n(1-\epsilon)}}\right)+O(n^{-1/2}), \tag{22}$$

$$\epsilon := \frac{2\kappa_n^2}{\pi\sigma_n\zeta_n}. \tag{23}$$

Note that we have dropped the + subscript from the test point and latent pair. We can deal with the term $\epsilon \leq 2/\pi$ by noting that

$$\mathbb{E}_{x_n,r_n}\frac{2\kappa_n^2}{\pi\sigma_n\zeta_n} \leq \mathbb{E}_{s,r_n}\frac{2(s^\top\Psi^{1/2}\Phi_n r_n)^2}{\pi d\psi_{n,\min}\omega_{n,\min}} = O(n^{-1}), \tag{24}$$

where $\psi_{n,\min}, \omega_{n,\min}$ are the smallest eigenvalues of Ψ_n, Ω_n, and s is a random vector drawn uniformly from the surface of the unit sphere. From here we can simply expand to first order to see that the ϵ term contributes at most $O(n^{-1})$.

We now show that we may replace the quadratic form involving x_n with its expected value. To do this, we first note that we can focus on estimating the expectation over a region which in which $|\zeta_n - \mathrm{Tr}(\Psi_n^2)| < c\mathrm{Tr}(\Psi_n^2)$ for some $0 < c < 1$ by applying the Hanson-Wright inequality. This gives

$$\left|\mathbb{E}_{r_n,x_n,z_n}\left[\mathrm{erfc}\left(\frac{\mathrm{sgn}(\eta_n)\kappa_n}{\sqrt{\frac{\pi}{p}\sigma_n\mathrm{Tr}(\Psi_n^2)}}\right)-\mathrm{erfc}\left(\frac{\mathrm{sgn}(\eta_n)\kappa_n}{\sqrt{\frac{\pi}{p}\sigma_n\zeta_n}}\right)\right]\right| \tag{25}$$

$$\leq \mathbb{E}_{r_n,x_n,z_n}\mathbf{1}_{\{|\zeta_n-\mathrm{Tr}(\Psi_n^2)|<c\mathrm{Tr}(\Psi_n^2)\}}\left|\mathrm{erfc}\left(\frac{\mathrm{sgn}(\eta_n)\kappa_n}{\sqrt{\frac{\pi}{p}\sigma_n\mathrm{Tr}(\Psi_n^2)}}\right)-\mathrm{erfc}\left(\frac{\mathrm{sgn}(\eta_n)\kappa_n}{\sqrt{\frac{\pi}{p}\sigma_n\zeta_n}}\right)\right|+O(e^{-g\sqrt{n}}),$$

for some constant g that depends on the choice of c. We can then obtain the following bound for the integrand in the region where $|\zeta_n - \mathrm{Tr}(\Psi_n^2)| < c\mathrm{Tr}(\Psi_n^2)$ by the mean value theorem:

$$\mathbb{E}_{r_n,x_n,z_n}\mathbf{1}_{\{|\zeta_n-\mathrm{Tr}(\Psi_n^2)|<c\mathrm{Tr}(\Psi_n^2)\}}\left|\mathrm{erfc}\left(\frac{\mathrm{sgn}(\eta_n)\kappa_n}{\sqrt{\frac{\pi}{p}\sigma_n\mathrm{Tr}(\Psi_n^2)}}\right)-\mathrm{erfc}\left(\frac{\mathrm{sgn}(\eta_n)\kappa_n}{\sqrt{\frac{\pi}{p}\sigma_n\zeta_n}}\right)\right| \tag{26}$$

$$\leq \mathbb{E}\frac{\sqrt{p}C|\kappa_n|}{\mathrm{Tr}(\Psi_n^2)^{3/2}}\left|\zeta_n - \mathrm{Tr}(\Psi_n^2)\right| \leq \frac{\sqrt{p}C\sqrt{\mathbb{E}[\kappa_n^2]\mathbb{E}|\zeta_n - \mathrm{Tr}(\Psi_n^2)|^2}}{\mathrm{Tr}(\Psi_n^2)^{3/2}}. \tag{27}$$

for some constant C. We can now use the following Lemma to show that the right hand side is $O(n^{-1/2})$ in expectation:

Lemma 3. *Let $r \in \mathbb{R}^d$ be a vector distributed either uniformly on the surface of the sphere of radius $\sqrt{d}$ or according to a mean-zero Gaussian with covariance I_d. Then for any matrix $A \in \mathbb{R}^{d\times d}$ we have*

$$\mathbb{E}|r^\top Ar - \mathrm{Tr}(A)| = O(||A||_{\mathrm{op}}) + O(||A||_F), \tag{28}$$

$$\mathbb{E}|r^\top Ar - \mathrm{Tr}(A)|^2 = O(||A||_{\mathrm{op}}^2) + O(||A||_F^2) \tag{29}$$

where $||\cdot||_{\mathrm{op}}$ is the operator norm induced by the Euclidean norm.

Proof. Applying the Hanson-Wright inequality for random vectors satisfying the convex concentration property [1, 5] we obtain the first identity

$$\int p(dr)|r^\top Ar - \mathrm{Tr}(A)| = \int_0^\infty dt\mathbb{P}(|r^\top Ar - \mathrm{Tr}(A)| > t) \tag{30}$$

$$\leq \int_0^\infty dt \exp\left(-\frac{Ct^2}{||A||_F^2}\right) + \exp\left(-\frac{ct}{||A||_{\mathrm{op}}}\right) = O(||A||_F) + O(||A||_{\mathrm{op}}), \tag{31}$$

where C, c are positive constants. Similarly we have

$$\int p(dr)|r^\top Ar - \mathrm{Tr}(A)|^2 = 2\int_0^\infty dt\mathbb{P}(|r^\top Ar - \mathrm{Tr}(A)| > t)t = O(||A||_F^2) + O(||A||_{\mathrm{op}}^2), \tag{32}$$

yielding the second identity. □

Applying this estimate and using the fact that $\mathbb{E}\kappa_n^2 = \mathrm{Tr}(\Phi_n^\top \Psi_n \Phi_n) = O(n^{-2})$ then shows that the right hand side is $O(n^{-1/2})$. We are therefore left with

$$\frac{1}{2}\mathbb{E}_{r_n}\mathbb{E}_{\eta_n,\kappa_n}\mathrm{erfc}\left(\frac{\mathrm{sgn}(\eta_n)\kappa_n}{\sqrt{\frac{\pi}{p}\sigma_n \mathrm{Tr}(\Psi_n^2)}}\right) + O(n^{-1/2}). \tag{33}$$

From our definitions, the variables η_n, κ_n follow a bivariate Gaussian distribution:

$$\eta_n, \kappa_n \sim \mathcal{N}\left(0, \begin{pmatrix} r_n^\top \Omega_n r & r_n^\top \Phi_n^\top \Phi_n r_n \\ r_n^\top \Phi_n^\top \Phi_n r_n & r_n^\top \Phi_n^\top \Psi_n \Phi_n r_n \end{pmatrix}\right). \tag{34}$$

The inner expectation can now be carried out analytically. We start by integrating over the conditional distribution $\kappa_n|\eta_n$, which is Gaussian with mean $\eta_n r_n^\top \Phi_n^\top \Phi_n r/\sigma_n$ and variance $r_n^\top \Phi_n^\top \Psi_n \Phi_n r_n - (r_n^\top \Phi_n^\top \Phi_n r_n)^2/\sigma_n$. To do this, we need the Gaussian integral over the complementary error function

$$\frac{1}{2}\int \frac{dx}{\sqrt{2\pi v}} e^{-(x-m)^2/(2v)}\mathrm{erfc}(cx) = \frac{1}{2}\mathrm{erfc}\left(\frac{mc}{\sqrt{1+2vc^2}}\right). \tag{35}$$

This formula can be derived by considering the function

$$I(a,b) := \int Dx \mathrm{erfc}(ax+b), \tag{36}$$

where Dx is a standard Gaussian measure. This integral can be solved by differentiating under the integral with respect to b, leading to Eq. (35). Applying this identity to Eq. (33) we obtain

$$\frac{1}{2}\mathbb{E}_{r_n}\mathbb{E}_{\eta_n}\mathrm{erfc}\left(\frac{|\eta_n| r_n^\top \Phi_n^\top \Phi_n r_n}{r_n^\top \Omega_n r_n \sqrt{\frac{\pi}{p}\sigma_n \mathrm{Tr}(\Psi_n^2) + 2\left[r_n^\top \Phi_n^\top \Psi_n \Phi_n r_n - \frac{(r_n^\top \Phi_n^\top \Phi_n r_n)^2}{r_n^\top \Omega_n r_n}\right]}}\right). \tag{37}$$

We can now split the integral over the η_n into half Gaussian integrals over complementary error functions. This allows us to invoke the relation

$$\frac{1}{2}\int_0^\infty \frac{\sqrt{2}dx}{\sqrt{\pi}} e^{-x^2/2}\text{erfc}(cx) = \frac{1}{\pi}\tan^{-1}\left(\frac{1}{\sqrt{2}c}\right). \tag{38}$$

This formula can similarly be derived by considering the function

$$J(c) := \int_0^\infty Dx\text{erfc}(cx) \tag{39}$$

and differentiating with respect to c. This leaves

$$\frac{1}{\pi}\mathbb{E}_{r_n}\tan^{-1}\left(\sqrt{\frac{r_n^\top \Omega_n r_n\left[\frac{\pi}{p}\text{Tr}(\Psi_n^2) r_n^\top \Omega_n r_n + 2 r_n^\top \Phi_n^\top \Psi_n \Phi_n r_n\right]}{2(r_n^\top \Phi_n^\top \Phi_n r_n)^2} - 1}\right) + O(n^{-1/2}). \tag{40}$$

Note that *this gives the generalization error for a fixed task vector* r_n.

To carry out the final expectation, we expand these quadratic forms around their mean values. To do this, let us introduce the $O(1)$ variables

$$\sigma_1^n = r_n^\top \Omega_n r_n, \quad \sigma_2^n = n^2 r_n^\top \Phi_n^\top \Psi_n \Phi_n r_n, \quad \sigma_3^n = n r_n^\top \Phi_n^\top \Phi_n r_n, \tag{41}$$

and the function

$$G(\sigma^n) = \tan^{-1}\left(\sqrt{\frac{\sigma_1^n(c\sigma_1^n + 2n^{-2}\sigma_2^n)}{2n^{-2}(\sigma_3^n)^2} - 1}\right), \tag{42}$$

where $c := \frac{\pi}{p}\text{Tr}(\Psi_n^2) = O(n^{-2})$. We start by noting that we can focus on estimating the expectation over a region in which $||\sigma^n - \mathbb{E}\sigma^n|| < t$ for $t > 0$. To see this, note that by the Hanson-Wright inequality:

$$\mathbb{E}_{r_n} G(\sigma^n) = \mathbb{E}_{r_n} \mathbf{1}_{\{||\sigma^n - \mathbb{E}\sigma^n|| < t\}} G(\sigma^n) + O(e^{-c\sqrt{n}}). \tag{43}$$

We can now bound the remaining error incurred by replacing σ^n with its expected value as:

$$|G(\sigma^n) - G(\mathbb{E}\sigma^n)| \leq \sum_i \sup_{||\xi - \mathbb{E}\sigma^n|| < t} |\partial_i G(\xi)||\sigma_i^n - \mathbb{E}\sigma_i^n| \tag{44}$$

$$= \sum_i \sup_{||\xi - \mathbb{E}\sigma^n|| < t} \left|\left[\frac{\xi_1(c\xi_1 + 2n^{-2}\xi_2)}{2n^{-2}(\xi_3)^2} - 1\right]^{-1/2}\left[\frac{\xi_1(c\xi_1 + 2n^{-2}\xi_2)}{2n^{-2}(\xi_3)^2}\right]^{-1}\right. \tag{45}$$

$$\tag{46}$$

$$\left.\partial_i \frac{\xi_1(c\xi_1 + 2n^{-2}\xi_2)}{2n^{-2}(\xi_3)^2}\right| |\sigma_i^n - \mathbb{E}\sigma_i^n| \tag{47}$$

$$\leq M \sup_{||\xi - \mathbb{E}\sigma^n|| < t} \left|\frac{\xi_1(c\xi_1 + 2n^{-2}\xi_2)}{2n^{-2}(\xi_3)^2} - 1\right|^{-1/2} \sum_i |\sigma_i^n - \mathbb{E}\sigma_i^n|, \tag{48}$$

for some constant M. We can now use the following lemma to give a uniform bound on the remaining term:

Lemma 4. *Let C be a positive definite matrix with submatrices*

$$C = \begin{pmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{pmatrix}, \tag{49}$$

with $\Psi \in \mathbb{R}^{n\times n}, \Omega \in \mathbb{R}^{d\times d}$, and $\Phi \in \mathbb{R}^{n\times d}$. Then

$$\frac{\mathrm{Tr}(\Omega)\mathrm{Tr}(\Phi^\top \Psi \Phi)}{\mathrm{Tr}(\Phi^\top \Phi)^2} \geq 1. \tag{50}$$

Proof. The claim follows from an application of the Cauchy-Schwarz inequality:

$$\mathrm{Tr}(\Phi^\top \Phi)^2 = (\mathbb{E}\langle \Phi^\top x, z\rangle)^2 \leq \mathbb{E}||\Phi^\top x||^2 \mathbb{E}||z||^2 = \mathrm{Tr}(\Phi^\top \Psi \Phi)\mathrm{Tr}(\Omega), \tag{51}$$

where the expectation is taken over zero-mean, jointly Gaussian vectors (x, z) which have a covariance matrix C. □

From this lemma and our assumptions on the spectrum, it follows that there exists a constant $M' > 0$ that does not depend on n such that

$$\frac{\langle \sigma_1^n \rangle (c\langle \sigma_1^n \rangle + 2n^{-2}\langle \sigma_2^n \rangle)}{2n^{-2}\langle \sigma_3^n \rangle^2} \geq 1 + M', \tag{52}$$

where we use brackets to denote the expectation. Hence if we choose t sufficiently small, we obtain

$$|G(\sigma^n) - G(\mathbb{E}\sigma^n)| \leq M'' \sum_i |\sigma_i^n - \mathbb{E}\sigma_i^n|, \tag{53}$$

where M'' is a positive constant. By Lemma 3, the right hand side is $O(n^{-1/2})$.

Replacing these quadratic forms with their means, we obtain the final result:

$$\frac{1}{\pi}\tan^{-1}\left(\sqrt{\frac{\mathrm{Tr}(\Omega_n)\left[\frac{\pi}{p}\mathrm{Tr}(\Psi_n^2)\mathrm{Tr}(\Omega_n) + 2\mathrm{Tr}(\Phi_n^\top \Psi_n \Phi_n)\right]}{2\mathrm{Tr}(\Phi_n^\top \Phi_n)^2} - 1}\right) + O(n^{-1/2}). \tag{54}$$

□

Remark. *Note that if one was to consider a non-isotropic Gaussian distribution over teacher vectors r, this would lead to a non-uniform shattering of the latent space. Supposing r follows a multivariate Gaussian distribution with covariance matrix Σ, one can see from Eq.* (6) *that this setting is equivalent to changing the covariance of z via the transformation $\Omega \mapsto \Sigma^{1/2}\Omega\Sigma^{1/2}$. To give a concrete example, if one takes the covariance matrix of r to be diagonal with elements $(v_1, \ldots, v_d)$, then this is equivalent to rescaling each latent dimension by a factor of $\sqrt{v_i}$. Thus, this would be equivalent to considering a distribution of binary classification problems that emphasize those dimensions with larger v_i.*

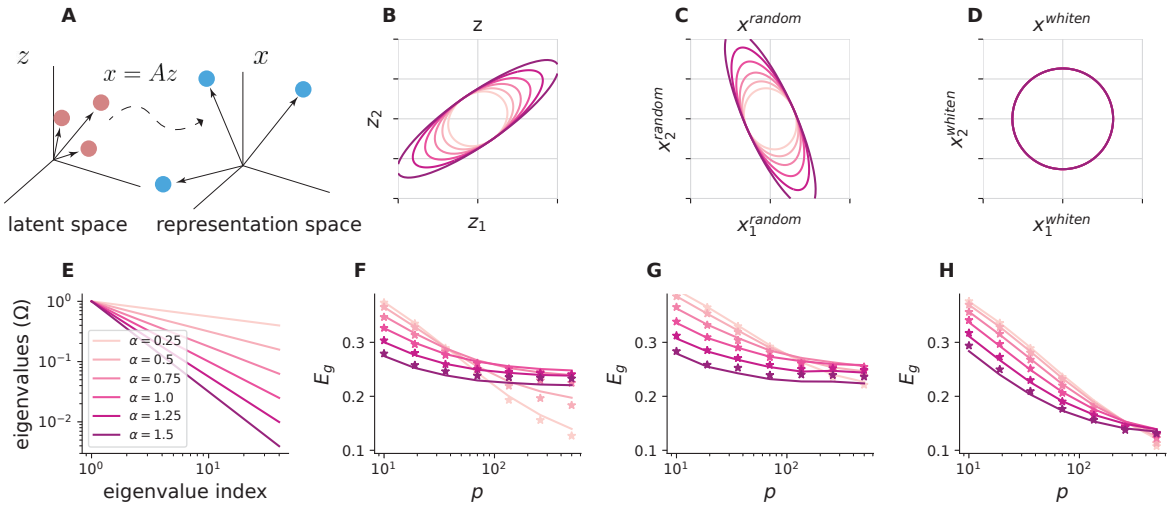

Figure 1: Theory predicts empirical generalization error in Gaussian model with power law covariance spectra. (a) Schematic illustrating simulation setup. Gaussian latent variables z are used to generate task labels, and predictions are formed using a linear transformation of the latent variables, $x = Az$. (b-d) Two typical units for the (b) latent variables, (c) random high dimensional projection, and (d) whitened transform for various values of the spectral decay exponent, α. (e) Eigenvalues of the latent covariance, Ω, for different decay rates, α. (f-h) Multi-task generalization error as a function of training samples, p, for (f) the latent variables themselves, (g) the random projection, and (h) the whitened transform.

2 Validation on Gaussian simulations

We validate our theory numerically on data drawn from a Gaussian model. Here, we find that our theory yields an excellent agreement with numerical simulations for a wide range of values (Methods). In these simulations, we sample the latents z_μ and neural responses x_μ from a multivariate Gaussian. We set the covariance matrix of the latents to have a spectrum that decays as a power law. Since the proof of our main theorem assumes that the spectra of the covariances decays slowly, this allows us to parametrically study violations of our main theorem's assumptions (Fig. 1a-e; SM). After sampling the latents, we form the neural responses either by taking a random high dimensional projection of the latent variables, or by applying a whitening transform to the latent variables (Fig. 1b-d; see Methods). We then calculate the empirical generalization error across a set of binary classification tasks which are formed by shattering the latent space as above.

As shown in Fig. 1f-h, our theory yields an excellent fit to numerical simulations. Importantly, the theoretical prediction holds all the way down to the few shot learning regime in which the number of training samples p is small, despite the theory being derived in the limit of large p. Furthermore, we find that the theory predicts the empirical generalization error well for a relatively small number of neurons n and latent dimensionality d, as well as latent variables whose covariance has an eigenspectrum that decays relatively quickly [4] (Methods). These results suggest that our theory can be applied to a wide range of datasets and is informative of the few shot learning regime.

3 Decomposition of generalization error into geometric terms

Having obtained a closed-form solution for generalization error in the thermodynamic limit, we now seek to rewrite this formula as a monotone function of interpretable geometric terms. We start by splitting Eq. (54) into a p-dependent and p-independent term:

$$\frac{1}{\pi}\tan^{-1}\left(\sqrt{\frac{\pi}{2p}\frac{\mathrm{Tr}(\Omega_n)\mathrm{Tr}(\Psi_n^2)\mathrm{Tr}(\Omega_n)}{(\mathrm{Tr}(\Phi_n\Phi_n^\top))^2}+\frac{\mathrm{Tr}(\Omega_n)\mathrm{Tr}(\Phi_n^\top\Psi_n\Phi_n)}{(\mathrm{Tr}(\Phi_n\Phi_n^\top))^2}-1}\right). \tag{55}$$

If we re-write the p-dependent term as

$$\frac{\pi}{2p}\left(\frac{\mathrm{Tr}(\Omega_n)\,\mathrm{Tr}(\Psi_n)}{\mathrm{Tr}(\Phi_n\Phi_n^\top)}\right)^2\cdot\frac{\mathrm{Tr}(\Psi_n^2)}{\mathrm{Tr}(\Psi_n)^2}, \tag{56}$$

we can notice the **participation ratio of the neural activity**

$$\mathrm{PR}(\Psi)=\frac{\mathrm{Tr}(\Psi_n)^2}{\mathrm{Tr}(\Psi_n^2)}, \tag{57}$$

which is a well-known measure of the dimensionality of neural activity. We collect the rest of the p-dependent term into our **correlation term**

$$c=\frac{\mathrm{Tr}(\Phi_n\Phi_n^\top)}{\mathrm{Tr}(\Omega_n)\,\mathrm{Tr}(\Psi_n)}, \tag{58}$$

and the generalization error now reduces to

$$\frac{1}{\pi}\tan^{-1}\left(\sqrt{\frac{\pi}{2pc^2\,\mathrm{PR}(\Psi)}+\frac{\mathrm{Tr}(\Omega_n)\mathrm{Tr}(\Phi_n^\top\Psi_n\Phi_n)}{(\mathrm{Tr}(\Phi_n\Phi_n^\top))^2}-1}\right). \tag{59}$$

To interpret the correlation term, note that the numerator $\mathrm{Tr}(\Phi_n\Phi_n^\top)$ can be expanded as

$$\mathrm{Tr}(\Phi_n\Phi_n^\top)=\sum_{i=1}^{d}\sum_{j=1}^{n}\mathbb{E}[z_i x_j]^2, \tag{60}$$

which is a sum-of-squares of the covariance between all pairs of latent variables z_i and all components x_j of neural activity. The denominator of c is just the total latent variance $\mathrm{Tr}(\Omega_n)$ times the total variance $\mathrm{Tr}(\Psi_n)$ of neural activity. The correlation term c can therefore be viewed as a generalization of the well-known Pearson correlation between two scalar variables to capture the total correlation strength between the d-dimensional latent variable z and the n-dimensional neural activity x. In particular, when $n, d = 1$, the total correlation c reduces to the square of the standard Pearson correlation between z and x.

Before moving on, we restrict ourselves to the case of diagonal latent covariance with

$$\Omega_n=\mathrm{diag}(\omega_1,\omega_2,\ldots,\omega_n).$$

Note that for *any* latent covariance, we can rotate the latent space to diagonalize Ω. Rotating the latent space does not affect our generalization error at all because we sample the task-defining vectors T from a rotationally-symmetric standard Gaussian distribution. The statements we make below about the latent variables $z_1, z_2, \ldots, z_d$ can more generally be interpreted as holding for the d *principal component directions* of the latent distribution when Ω_n is not diagonal, since these principal component directions will be rotated onto the d standard axes when we diagonalize Ω_n.

We call (the inverse of) the remaining p-independent term the **alignment term** a:

$$a = \frac{(\mathrm{Tr}(\Phi_n \Phi_n^\top))^2}{\mathrm{Tr}(\Omega_n)\,\mathrm{Tr}(\Phi_n^\top \Psi_n \Phi_n)}.$$

In the main text (and as described below), the alignment term is subdivided further, but we can also interpret a as a whole. We can expand the main term in the denominator as

$$\mathrm{Tr}(\Phi_n^\top \Psi_n \Phi_n) = \sum_{i=1}^{d} \phi_i^\top \Psi_n \phi_i = \sum_{i=1}^{d} \|\phi_i\|^2 \,\mathrm{Var}(\hat{\phi}_i \cdot x), \tag{61}$$

where ϕ_i is the i^{th} column of Φ_n, and $\mathrm{Var}(\hat{\phi}_i \cdot x)$ is the variance of the total neural response x projected onto the direction of ϕ_i. Using the fact that the average neural activity, conditioned on a specific value of the latents is given by $\mathbb{E}[x|z] = \Phi\Omega^{-1}z$, we can see that for a diagonal Ω, the column $\hat{\phi}_i$ corresponds to the *coding direction of the latent variable* z_i. Additionally, the norm $\|\phi_i\|$ reflects the *coding strength of the latent variable* z_i. The numerator of a can be written as

$$(\mathrm{Tr}(\Phi_n \Phi_n^\top))^2 = \left(\sum_{i=1}^{d} \|\phi_i\|^2 \right), \tag{62}$$

and therefore depends only on the norms of the coding directions ϕ_i but not on their directions. Holding the norms of each ϕ_i fixed, we can see that the alignment term a prefers that each coding direction $\hat{\phi}_i$ be positioned along a direction of minimal response variance in the neural state space. In other words, the only variance along a direction $\hat{\phi}_i$ should be caused by variations in the corresponding latent variable. Intuitively, the alignment term a encourages arranging the coding directions of each latent variable in such a way that the signal for these variables do not interfere with one another and so that all latents are coded along directions with low noise. We make this notion more formal below.

Now let us subdivide the alignment term a as in the main text. We partition the total neural variance Ψ_n into stimulus-driven variance and stimulus-independent variance. Recall that in our setup, latents z_n and neural responses x_n are drawn from a joint Gaussian distribution

$$(x_n, z_n) \sim \mathcal{N}\left(0, \begin{pmatrix} \Omega_n & \Phi_n^\top \\ \Phi_n & \Psi_n \end{pmatrix}\right). \tag{63}$$

This is equivalent to a model in which we first sample $z_n \sim \mathcal{N}(0, \Omega_n)$ and then sample x_n as $x_n = \Phi_n \Omega_n^{-1} z_n + \varepsilon_n$, where ε_n is stimulus-independent Gaussian noise with a covariance given by

$$\Psi_n - \Phi_n \Omega_n^{-1} \Phi_n^\top, \tag{64}$$

which is the covariance of the neural responses conditional on the latents, $\mathrm{cov}(x_n|z_n)$. This suggests a partitioning of the total neural variance Ψ_n into stimulus-independent variance, $H_n := \Psi_n - \Phi_n\Omega_n^{-1}\Phi_n^\top$, and stimulus-driven variance, $\mathrm{cov}(x) - H_n = \Phi_n\Omega_n^{-1}\Phi_n^\top$.

We can now decompose the p-independent term of our generalization error formula as

$$\frac{\mathrm{Tr}(\Omega_n)\,\mathrm{Tr}(\Phi_n^\top\Psi_n\Phi_n)}{(\mathrm{Tr}(\Phi_n\Phi_n^\top))^2} = \frac{\mathrm{Tr}(\Omega_n)\,\mathrm{Tr}(\Phi_n^\top H_n\Phi_n)}{(\mathrm{Tr}(\Phi_n\Phi_n^\top))^2} + \frac{\mathrm{Tr}(\Omega_n)\,\mathrm{Tr}(\Phi_n^\top(\Phi_n\Omega_n^{-1}\Phi_n^\top)\Phi_n)}{(\mathrm{Tr}(\Phi_n\Phi_n^\top))^2}. \tag{65}$$

We call the inverse of the first term **signal-noise factorization** s:

$$\frac{1}{s} = \frac{\mathrm{Tr}(\Omega_n)\,\mathrm{Tr}(\Phi_n^\top H_n\Phi_n)}{(\mathrm{Tr}(\Phi_n\Phi_n^\top))^2}. \tag{66}$$

This interpretation can be understood by noting that

$$\mathrm{Tr}(\Phi_n^\top H_n\Phi_n) = \sum_{i=1}^d \phi_i^\top H_n\phi_i = \sum_{i=1}^d \|\phi_i\|^2 \cdot \hat{\phi}_i^\top H_n\hat{\phi}_i. \tag{67}$$

We can see that $\hat{\phi}_i^\top H\hat{\phi}_i$ gives the projection of stimulus-independent noise along the coding direction $\hat{\phi}_i$ of the i^{th} latent variable, so $\mathrm{Tr}(\Phi_n^\top H\Phi_n)$ measures the amount of stimulus-independent noise corrupting the signal directions for each latent variable. In particular, note that s is maximized when there is no stimulus-independent noise in the neural dimensions used for coding the latent variables.

Finally, we call the inverse of the second part of the p-independent term the **signal-signal factorization**:

$$\frac{1}{f} = \frac{\mathrm{Tr}(\Omega_n)\,\mathrm{Tr}(\Phi_n^\top(\Phi_n\Omega_n^{-1}\Phi_n^\top)\Phi_n)}{(\mathrm{Tr}(\Phi_n\Phi_n^\top))^2}. \tag{68}$$

To understand this term, recall our assumption that

$$\Omega_n = \mathrm{diag}(\omega_1, \ldots, \omega_d), \tag{69}$$

which can be accomplished by making an orthogonal change of variables. Now observe that

$$\mathrm{Tr}(\Phi_n^\top\Phi_n\Omega_n^{-1}\Phi_n^\top\Phi_n) = \sum_{i=1}^d\sum_{j=1}^d \frac{1}{\omega_i}\langle\phi_i, \phi_j\rangle^2. \tag{70}$$

The numerator of $1/f$ is a weighted sum of the dot products between *all pairs* of coding vectors ϕ_i and ϕ_j for latents z_i and z_j. By contrast, the denominator $(\mathrm{Tr}(\Phi_n\Phi_n^\top))^2$ can be expanded to

$$(\mathrm{Tr}(\Phi_n\Phi_n^\top))^2 = \left(\sum_{i=1}^d \langle\phi_i, \phi_i\rangle\right)^2, \tag{71}$$

which only captures the signal strengths $\|\phi_i\|$ without regard for the angles between pairs of distinct signal directions. With the norms $\|\phi_i\|$ fixed, f is maximized by making all coding

directions orthogonal (and more generally for non-diagonal Ω, by mapping the PCs of the latents to orthogonal directions). That is, f is maximized by a factorized code.

The interpretation of our factorization term f is somewhat complicated by its dependence on the latent covariance matrix Ω, especially when Ω can change between conditions (e.g., the distribution of the rats' positions can change between sessions in our analysis of CA1 and PFC spatial codes) or when latents are strongly correlated (e.g., the positions of different marker locations in the pose estimation data are highly correlated). We therefore compare the signal-signal factorization to a simplified measure that removes the dependence on Ω completely:

$$\frac{1}{f_{\text{simplified}}} = \frac{\text{Tr}(\Phi_n^\top \Phi_n \Phi_n^\top \Phi_n)}{(\text{Tr}(\Phi_n \Phi_n^\top))^2}, \tag{72}$$

Letting θ_{ij} denote the angle between the coding directions $\hat{\phi}_i$ and $\hat{\phi}_j$ and defining weights $w_{ij} = \|\phi_i\|^2 \|\phi_j\|^2$, we can rewrite

$$\frac{1}{f_{\text{simplified}}} = \frac{\sum_{i,j=1}^D w_{ij} (\cos\theta_{ij})^2}{\sum_{i,j=1}^D w_{ij}} \tag{73}$$

so that $1/f_{\text{simplified}}$ is just a weighted average of the squared cosine angle between all pairs of coding directions and is manifestly optimized by a factorized code with $\cos\theta_{ij} = 0$ for $i \neq j$. The weighting by the signal norms $\|\phi_i\|$ is unavoidable. For example, if one coding direction ϕ_i has norm very near zero (i.e., latent z_i is not decodable at all), then the angle it makes with other coding directions is irrelevant to generalization error and should not enter into f, since our measures capture the geometry *relevant for generalization error*.

We demonstrate numerically that practically all of the variation we observe in f across all of our analyses is captured by $f_{\text{simplified}}$ and is therefore independent of any changes in the latent covariance Ω (SM Fig. 2).

4 Optimal geometry

Here we derive our formula for the spectrum of the optimal neural representation and show that it is disentangled. The argument is not fully rigorous, but it matches numerical results very well.

Let us begin by showing that directions in the latent space with more variance are, on average, more informative of the task labels than those with low variance. To do this, consider a latent variable z_i and a label for a given task: $y = \text{sgn}(z \cdot r)$, where as above r denotes the normal vector of the hyperplane defining the task and z is the vector of all latent variables. The correlation between this latent variable and the labels for this task is given by, $\rho := \mathbb{E}_z[z_i \text{sgn}(z \cdot r)] / \sqrt{\mathbb{E}_z[z_i^2] \mathbb{E}_z \text{sgn}(z \cdot r)^2}$. Thus, we can quantify how informative a latent direction is for the labels across all tasks by calculating $\mathbb{E}_r|\rho|$, where the absolute value is taken to break the symmetry between positive and negative ρ values as we average over r. Using Eq. (7), together with the concentration of $r \cdot \Omega r$ about its mean value, we find that for large d:

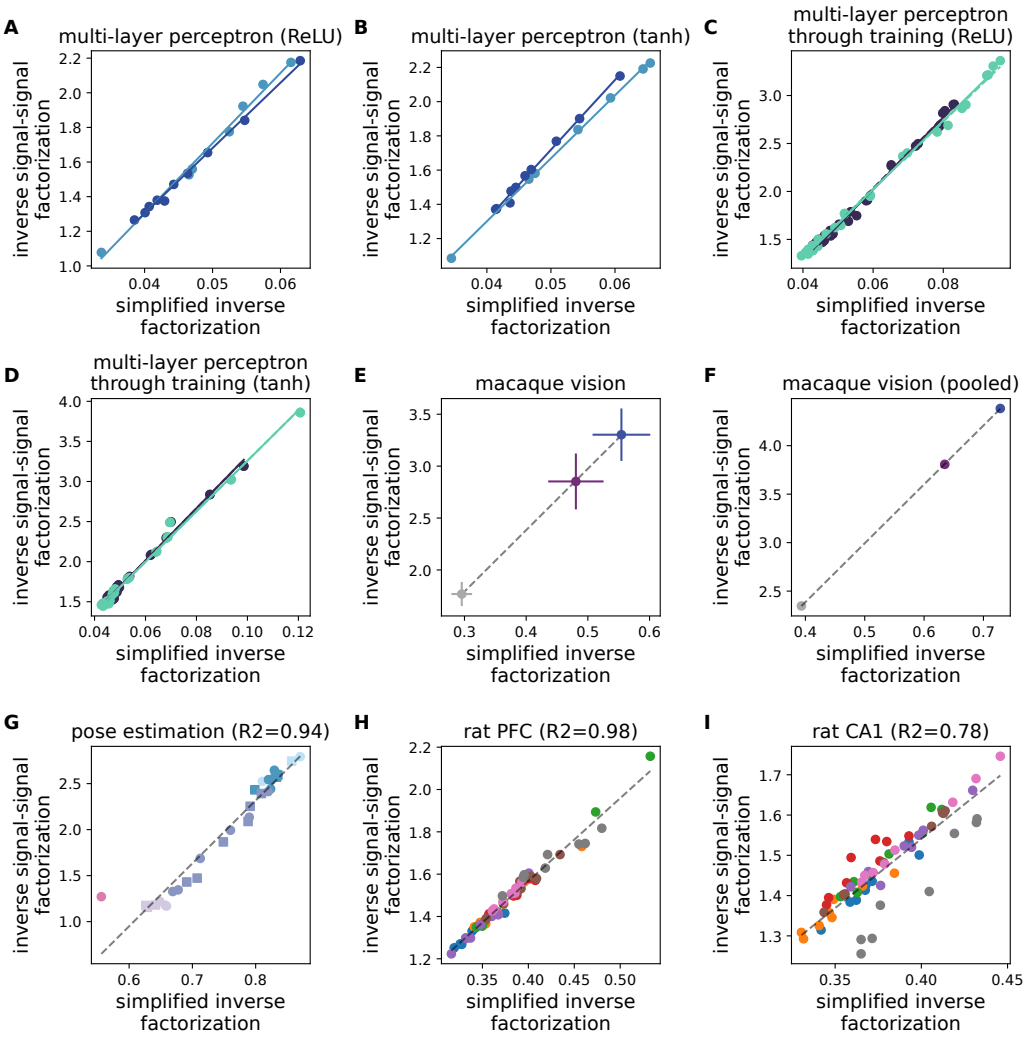

Figure 2: Almost all variation in our factorization term is driven by changing angles between coding directions and not by changes in the latent covariance Ω. We compare the reciprocal of signal-signal factorization to a simplified formula that can be written as a weighted average of squared cosines between pairs of coding directions and does not depend on the latent covariance Ω (see Eq. (73)). The strong linear relationship between $1/f$ and the simplified formula shows that we can effectively consider f to not depend on the latent covariance Ω across all the experiments presented in this work. The legend for each panel is not reprinted for brevity but matches the experimental figures. (a) Random and trained ReLU networks (Fig. 3). (b) Random and trained tanh networks (SM Fig. 5). (c) ReLU networks across training (SM Fig. 4). (d) Tanh networks across training (SM Fig. 6). (e) Macaque ventral stream (Fig. 5, error bars denote SEM and points the mean across categories). (f) Macaque ventral stream (pooled categories, SM Fig. 11). (g) Deep pose estimation network (Fig. 4) (h-i) Rat CA1 and PFC (Fig. 7)

$$\mathbb{E}_r|\rho| = \frac{\sqrt{2}}{\sqrt{\pi\Omega_{ii}\mathrm{Tr}(\Omega)}}\mathbb{E}_r|(\Omega r)_i| = \frac{2\sqrt{(\Omega^\top\Omega)_{ii}}}{\pi\sqrt{\Omega_{ii}\mathrm{Tr}(\Omega)}}. \tag{74}$$

If Ω is diagonal (or, equivalently, if we consider the correlation of the label with a principle component of Ω), then this is merely proportional to $\Omega_{ii}^{1/2}/\sqrt{\mathrm{Tr}(\Omega)}$–i.e. the standard deviation of the latent, normalized by the square root of the total variance across all latent variables, $\mathrm{Tr}(\Omega)$. More generally, this equation shows that latents with large variance are more informative of task labels.

With these considerations in place, our goal is to calculate

$$\arg\min_{\Psi,\Phi} \frac{1}{\pi}\tan^{-1}\left(\frac{\sqrt{\mathrm{Tr}(\Omega)\left[\frac{\pi}{p}\mathrm{Tr}(\Psi^2)\mathrm{Tr}(\Omega) + 2\left(\mathrm{Tr}(\Phi^\top\Psi\Phi) - \frac{(\mathrm{Tr}[\Phi\Phi^\top])^2}{\mathrm{Tr}(\Omega)}\right)\right]}}{\sqrt{2}\mathrm{Tr}(\Phi\Phi^\top)}\right), \tag{75}$$

subject to

$$\begin{pmatrix}\Psi & \Phi\\ \Phi^\top & \Omega\end{pmatrix} \succ 0. \tag{76}$$

For a positive matrix Ω, this condition is equivalent to

$$\Psi \succ 0, \quad \Psi - \Phi\Omega^{-1}\Phi^\top \succ 0. \tag{77}$$

Rewriting the argument of the $\tan^{-1}$ function, we can see that the objective may be written as

$$\arg\min_{\Psi,\Phi} \frac{\frac{\pi}{p}\,\mathrm{Tr}(\Psi^2)\,\mathrm{Tr}(\Omega) + 2\,\mathrm{Tr}(\Phi^\top\Psi\Phi)}{\mathrm{Tr}(\Phi\Phi^\top)^2}. \tag{78}$$

Our approach is to first minimize Eq. (78) holding $\mathrm{Tr}(\Phi\Phi^\top) = \gamma$ fixed and show that the objective is ultimately invariant to the choice of γ. The optimization can be done by introducing the Lagrangian

$$L(\Psi,\Phi,\rho,\zeta) = c\mathrm{Tr}(\Psi^2) + 2\mathrm{Tr}(\Phi^\top\Psi\Phi) + \zeta(\gamma - \mathrm{Tr}(\Phi\Phi^\top)) \tag{79}$$

$$- \int d^n x\rho(x)\mathrm{Tr}(xx^\top(\Psi - \Phi\Omega^{-1}\Phi^\top)), \tag{80}$$

where the $\rho(x)$ are KKT multipliers enforcing the positive definite constraint, ζ is a Lagrange multiplier, and $c := \frac{\pi}{p}\mathrm{Tr}(\Omega)$. Note that we do not explicitly enforce the positivity constraint on Ψ, as we find that this is unnecessary. The KKT equations are

$$\partial_\Psi L = 2c\Psi + 2\Phi\Phi^\top - \langle xx^\top \rangle_\rho = 0, \tag{81}$$

$$\partial_\Phi L = 4\Psi\Phi + 2\langle xx^\top \rangle \Phi\Omega^{-1} - 2\zeta\Phi = 0, \tag{82}$$

$$\delta_\rho L = \mathrm{Tr}(xx^\top(\Psi - \Phi\Omega^{-1}\Phi^\top)) \geq 0, \tag{83}$$

$$\rho(x)\mathrm{Tr}(xx^\top(\Psi - \Phi\Omega^{-1}\Phi^\top)) = 0, \tag{84}$$

$$\mathrm{Tr}(\Phi\Phi^\top) = \gamma. \tag{85}$$

We try for a solution with $\Psi = \Phi\Omega^{-1}\Phi^\top$. Since the variance of x conditional on z is given by $\Psi - \Phi\Omega^{-1}\Phi^\top$, this is equivalent to looking for solutions in which the neural code has no signal-unrelated noise. Then Eq. (81) gives $\langle xx^\top \rangle_\rho = 2\Phi(c\Omega^{-1} + I)\Phi^\top$. Using these two formulae, Eq. (82) gives the condition

$$\Phi(4\Omega^{-1}\Phi^\top\Phi + 4(c\Omega^{-1} + I)\Phi^\top\Phi\Omega^{-1} - 2\zeta I) = 0. \tag{86}$$

This suggests looking for a Φ whose right singular vectors are aligned with those of Ω. The singular values of Φ are then given by:

$$\phi_i^2 = \frac{\gamma}{\sum_i \frac{\omega_i^2}{2\omega_i + \frac{\pi}{P}\mathrm{Tr}(\Omega)}} \frac{\omega_i^2}{2\omega_i + \frac{\pi}{p}\mathrm{Tr}(\Omega)}. \tag{87}$$

From the assumption that $\Psi = \Phi^\top\Omega^{-1}\Phi$, we can see that the eigenvectors of Ψ and the left singular vectors of Φ can be chosen however we want, so long as they are the same. The eigenvalues of Ψ are then simply

$$\psi_i = \frac{\gamma}{\sum_i \frac{\omega_i^2}{2\omega_i + \frac{\pi}{p}\mathrm{Tr}(\Omega)}} \frac{\omega_i}{2\omega_i + \frac{\pi}{p}\mathrm{Tr}(\Omega)}. \tag{88}$$

Plugging this solution into the objective, we indeed find that γ drops out of the picture entirely. Since we are free to choose γ, we obtain that the optimal representation satisfies the following properties: (1) The eigenvectors of Ψ are the left singular vectors of Φ and (2) The eigenvectors of Ω are the right singular vectors of Φ.

We now show that the optimal representation described above is disentangled. More precisely, we show that the principal components of z directly map onto the principal components of x. If we let o be the eigenvector of Ψ corresponding to the neural direction and u the eigenvector of Ω corresponding to the latent direction, we can see that the value of

$$\mathbb{E}[\langle o, x\rangle\langle u, z\rangle] = u^\top\Phi o. \tag{89}$$

is either 0 or equal to a singular value of Φ. It is then easy to check that $\mathrm{corr}(\langle o, x\rangle, \langle u, z\rangle)$ is similarly either 0 or 1. Thus the principal components of the neurons and latents stand in a one to one relationship to one another. When the individual latent variables are uncorrelated

with one another (i.e. when Ω is diagonal), this implies that distinct latent variables are represented along mutually orthogonal directions.

Finally, the eigen/singular-values of the matrices are given by

$$\phi_i^2 \propto \frac{\omega_i^2}{2p\omega_i + \pi \mathrm{Tr}(\Omega)}, \tag{90}$$

$$\psi_i \propto \frac{\omega_i}{2p\omega_i + \pi \mathrm{Tr}(\Omega)}. \tag{91}$$

This gives the optimal representation's spectrum. As stated in the main text, we can see that as p grows relative to d, the spectrum becomes increasingly flat, indicating that more variance in the neural state space is being allocated to the less informative directions in the latent space.

5 Comparison of all analyses to SVC

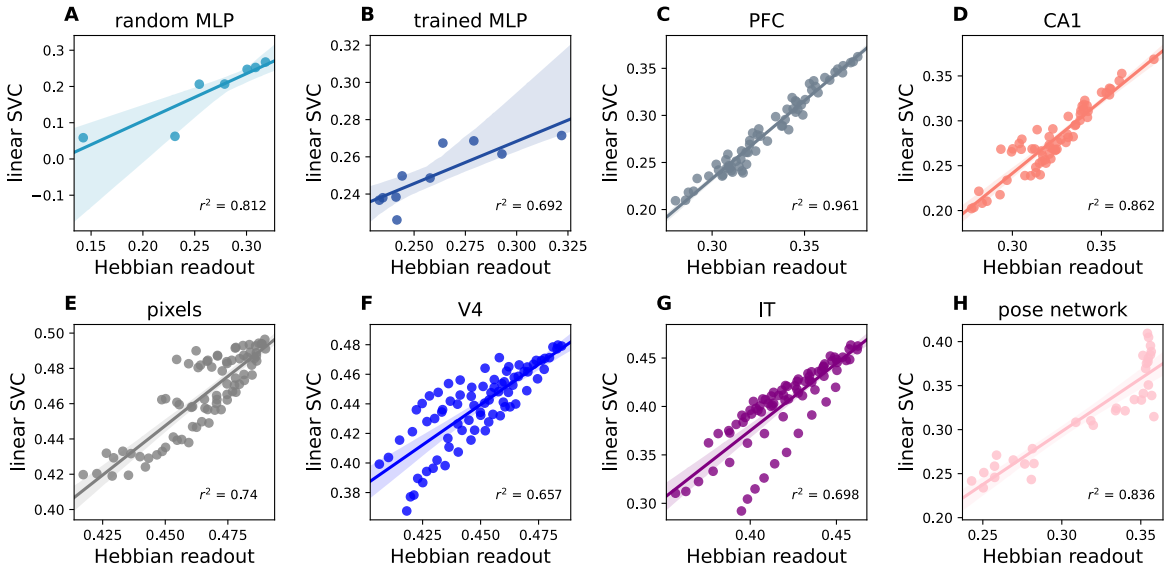

Figure 3: Comparison of the theoretically calculated Hebbian readout and a linear support vector classifier (SVC) on the multi-task learning problems analyzed in the main text. We fit a linear SVC using the default parameters in scikitlearn and average over train/test splits, as well as distinct shatterings of the latent space. Across all analyses reported in the main text, we find a strong linear association between the performance of the two readouts, suggesting that the geometric statistics we derived using Hebbian readouts are also informative of the performance of non-Hebbian readouts. (a-b) Random and trained MLP analyses across all layers for fixed p. (c-d) PFC and CA1 analyses across all rats and sessions. (e-g) Pixels, V4, and IT analyses for all values of p reported in the main text. (h) DeepLabCut results using the same values of p reported in the main text.

6 Additional MLP analyses

Here, we present results regarding the evolution of the geometric terms through training. Additionally, we redo the MLP analyses using a tanh non-linearity in the random and trained networks, and we consider a trained network with a similar fanout structure to the random network.

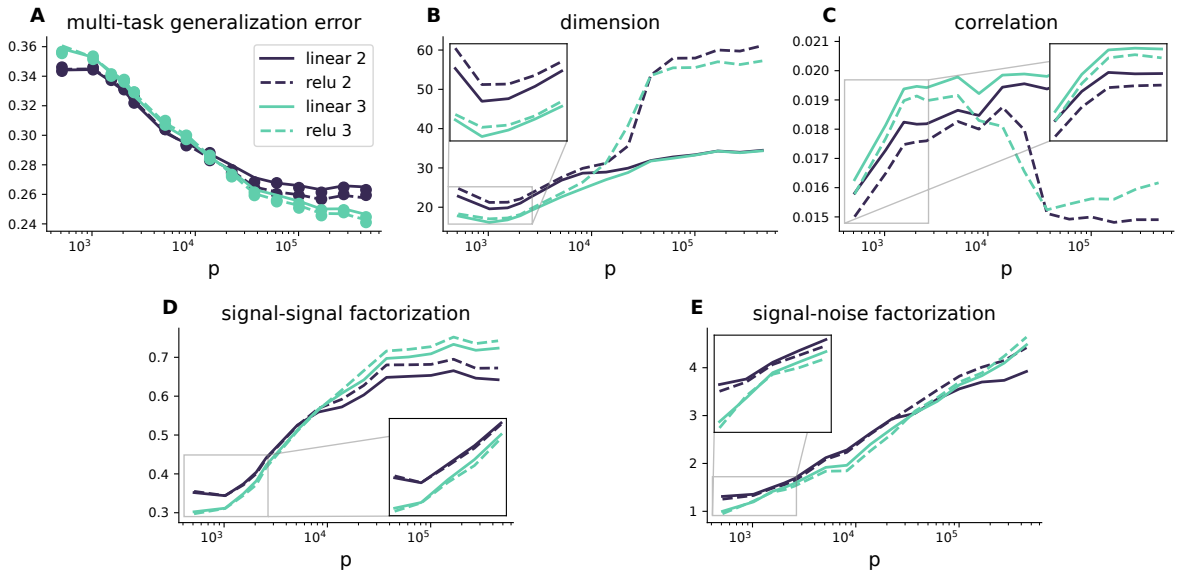

Figure 4: Evolution of generalization error and representational geometry through training. Note that p here denotes the number of stochastic gradient descent (SGD) training steps. (a) Generalization error of the Hebb rule applied to four different layers on a previously unseen set of tasks. We examine layers from the early stages of the MLP (black) as well as late stages (green) for both linear (solid line) and relu (dashed line) layers. We can see that the error decreases across all layers with the number of training samples, p. (b-e) Evolution of the geometry over the course of learning. Insets show the geometric terms during the first five steps of SGD. Just as in the rat analyses of Sec. 3.6, we find that nearly all geometric terms improve uniformly at the beginning of learning. Moreover, we find a similar fall-then-rise trend in the total dimension in these early stages. Late in training, we find that the dimension increases, and we see a non-monotonic trend in the correlation, whereby it begins to decrease late in training in the relu layers. Both factorization terms uniformly increase through training across the network, though the SSF begins to plateau at the end of training. The geometric trends we observe late in training mirror the effect of increasing the number of training samples on the optimal neural code.

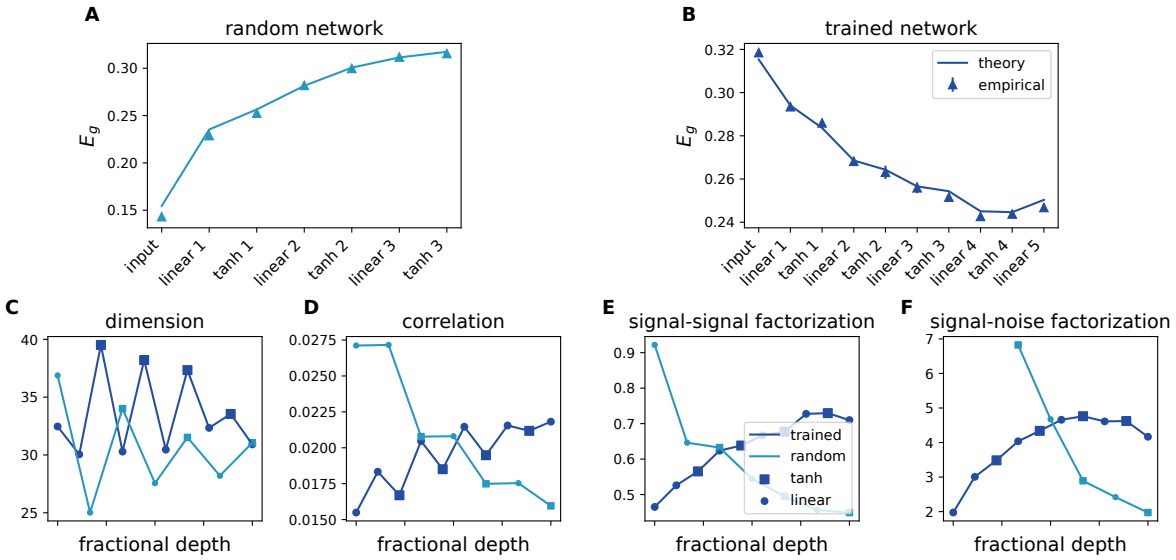

Figure 5: Generalization error and geometry of the tanh MLPs. (a-b) Generalization error for the random and trained network. (c-f) Layer-wise geometry of the networks. We find similar trends as with the relu non-linearity.

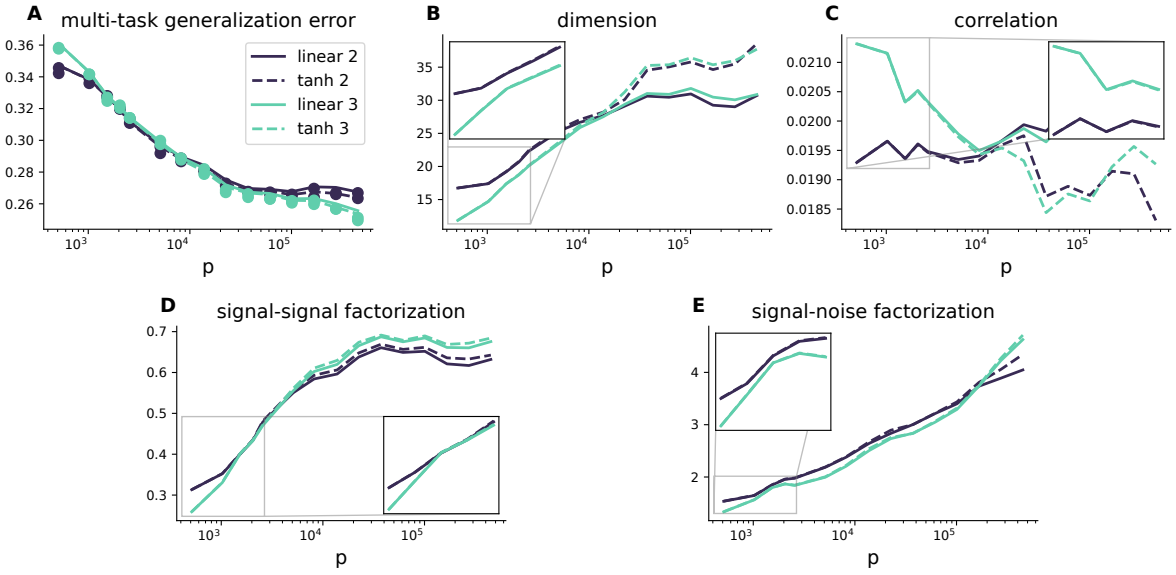

Figure 6: Dynamics of generalization error (a) and geometry (b-e) through training using networks with a tanh non-linearity. Again, we show 2 early and 2 late layers. At each of the early/late stages, we show a linear and tanh layer. Inset denotes geometric quantities over the first five steps of SGD on a linear scale. We see that with the exception of the correlation, the geometric terms uniformly improve early in learning, with non-monotonic effects beginning in the later stages of learning.

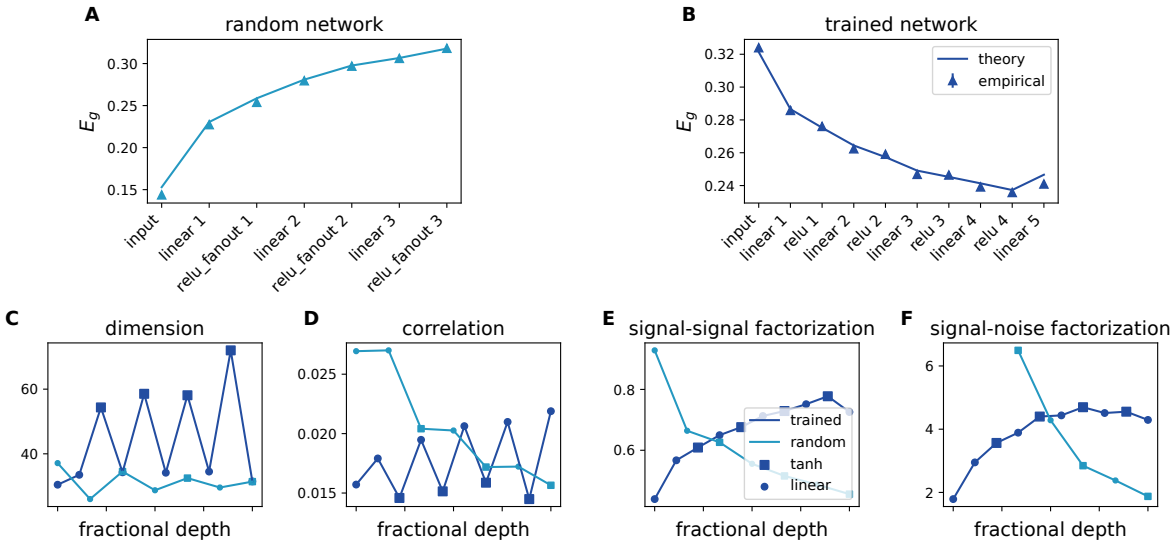

Figure 7: Generalization error and geometry of the relu MLPs in which the trained network has a similar expansionary structure as the random network. Specifically, the trained network has a layer structure in which each layer is 1.5 times the size of the previous one. We find similar trends as those reported in the main text. (a-b) Generalization error in both networks. (c-f) Geometry across layers.

7 Additional macaque analyses

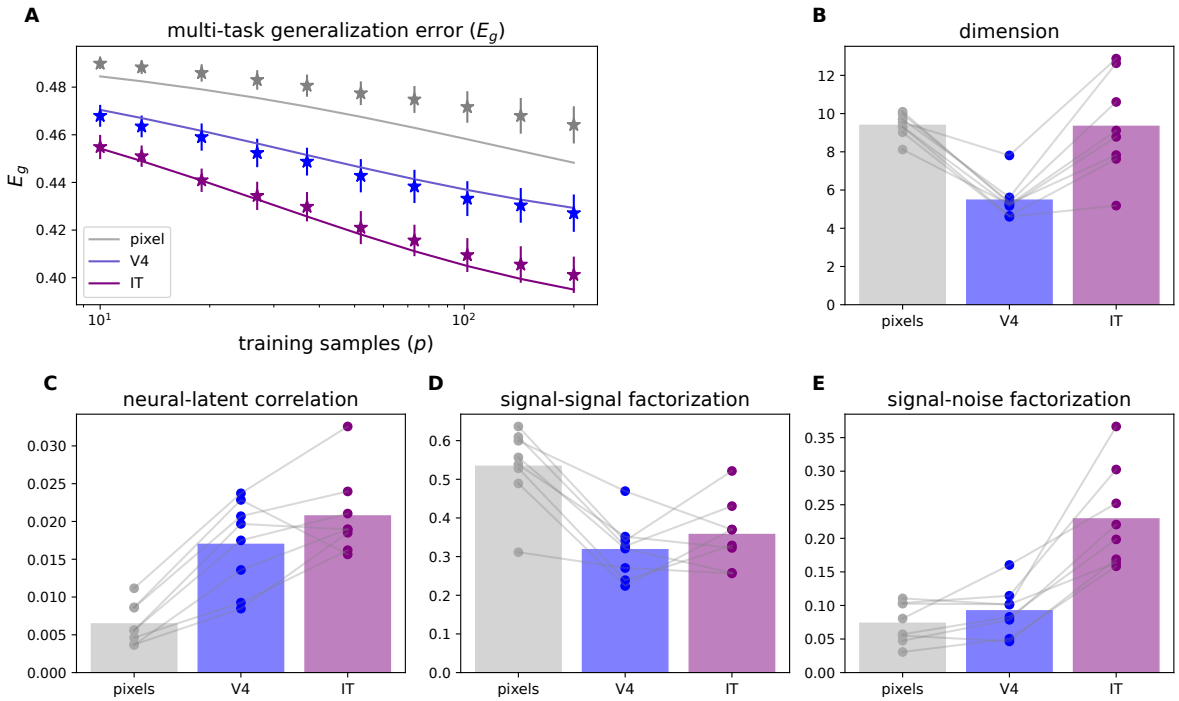

Figure 8: Geometry and generalization error using the same number of units across regions [2]. Here, we repeat the analysis presented in the main text, only we project the pixels and IT data down to 88 dimensions using Gaussian random projection. (a) Generalization error for the three representations. (b-e) Geometric terms. We find qualitatively similar trends to those presented in the main text.

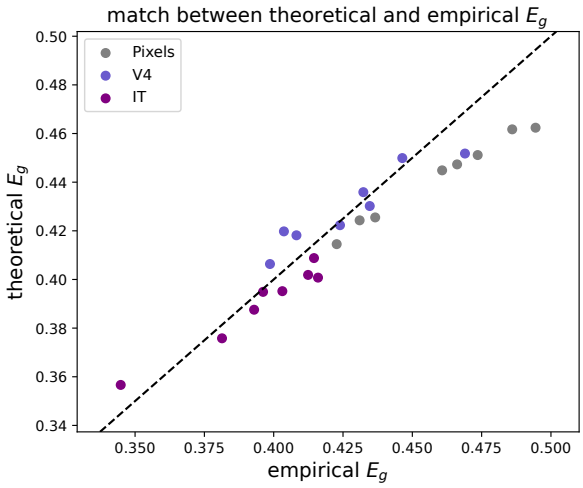

Figure 9: Match between theoretical and empirical generalization error across each of the 8 individual categories. In the main text, we presented the theoretical and empirical generalization errors averaged over individual object categories. Here we show that the theory predicts the empirical generalization error well across individual object categories.

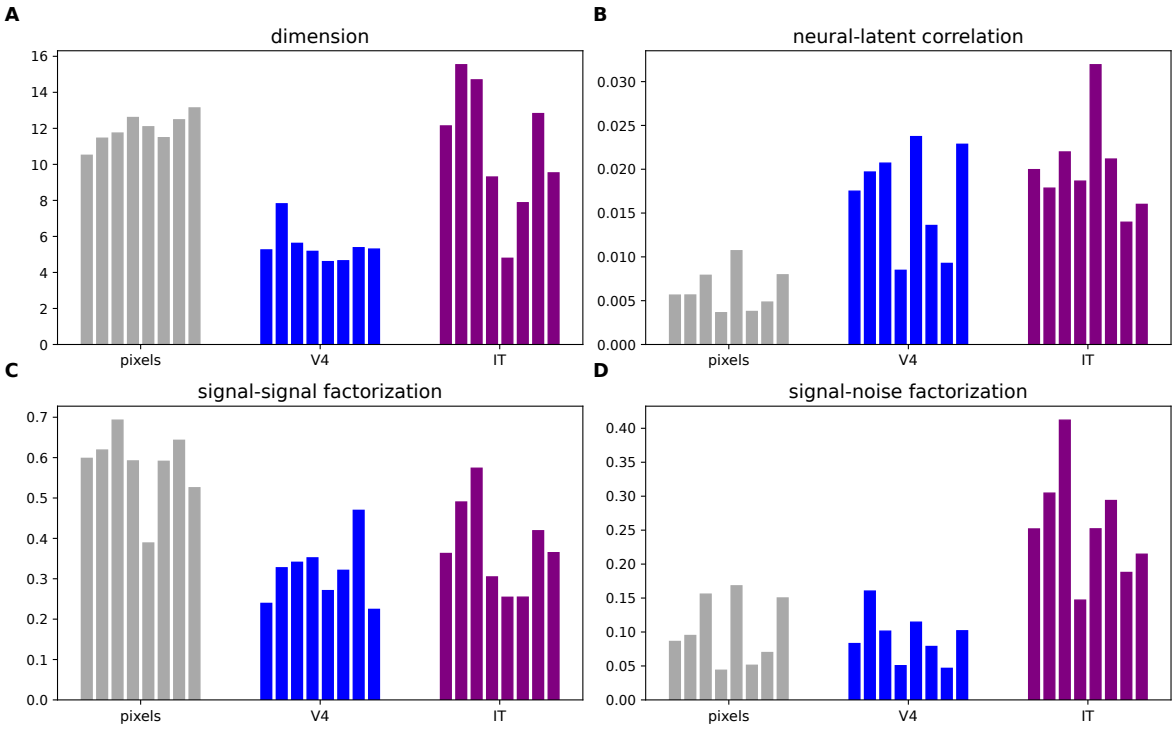

Figure 10: Geometry across individual object categories. Here, we present the distribution of geometric terms, calculated on subsets of the data corresponding to each of the 8 object categories.

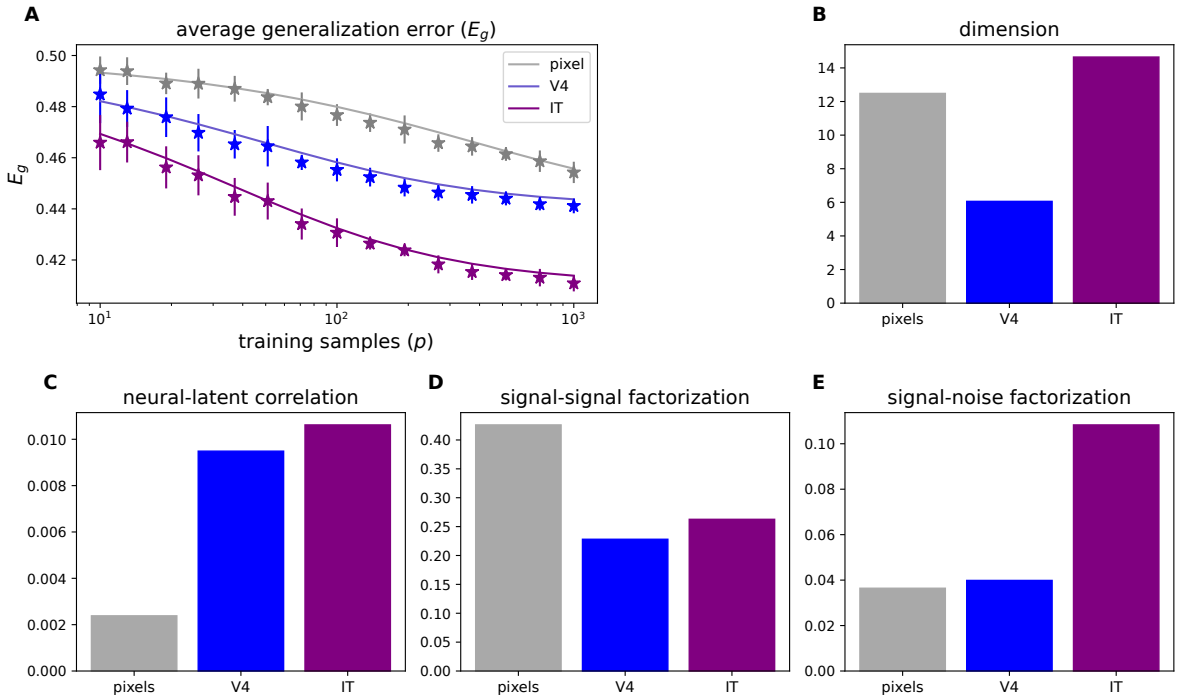

Figure 11: Generalization error and geometry of the pooled monkey data. Here, we pool data from all categories together, rather than considering data subsets corresponding to stimuli coming from the same object category as done in the main text. We can see that the trends are largely the same and that the theory predicts the empirical error. (a) Theoretical and empirical generalization error. (b-e) Geometry of the pooled data.

Measure	Group 1	Group 2	t	p	p (adj.)
dimension	pix	V4	13	3.8e-06	4.6e-05
dimension	pix	IT	0.791	0.455	0.496
dimension	V4	IT	-5.17	0.00129	0.00194
corr.	pix	V4	-7.49	1.4e-04	4.2e-04
corr.	pix	IT	-9.23	3.6e-05	1.8e-04
corr.	V4	IT	-1.6	0.154	0.205
ssf	pix	V4	8.89	4.6e-05	1.8e-04
ssf	pix	IT	7.13	1.9e-04	4.5e-04
ssf	V4	IT	-1.44	0.192	0.231
snf	pix	V4	0.682	0.517	0.517
snf	pix	IT	-5.94	5.7e-04	9.8e-04
snf	V4	IT	-6.84	2.4e-04	4.9e-04

Table 1: Paired sample two-sided t-test results presented in Fig. 5. All p-values were adjusted using a Benjamini-Hochberg method.

8 Additional rat analyses

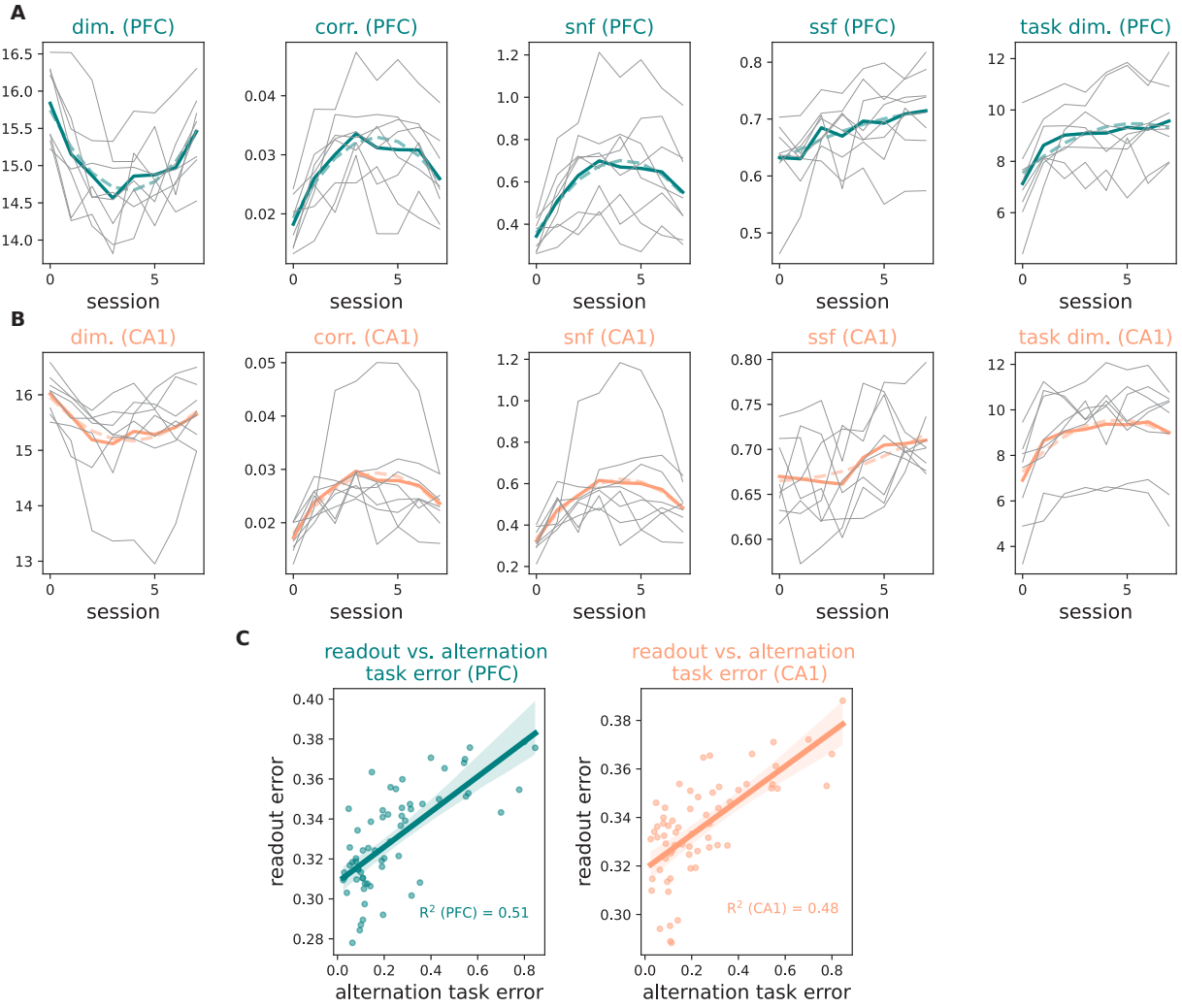

Figure 12: (A-B) Example quadratic model fits (dashed lines) for PFC (A) and CA1 (B). Geometric measures for individual rats ($N = 8$) shown in light grey and population averages in solid colored lines. (To test the effect of the outlier rat in the CA1 data, we reran all statistical tests after excluding this rat's CA1 data. Besides the linear model's CA1 SSF term, no other term ceased to be statistically significant at $\alpha = 0.01$ after performing this exclusion.) (C) Relationship between single session error rates and theoretically calculated E_g.

Area	Model	Measure	γ	p_γ	p_γ (adj.)	η	p_η	p_η (adj.)
PFC	Quad.	task dim.	0.716	7.0e-06	9.4e-06	-0.0662	0.00247	0.00297
PFC	Quad.	snf	0.164	2.8e-14	2.3e-13	-0.0197	3.1e-11	1.2e-10
PFC	Quad.	ssf	0.0204	0.0267	0.0305	-0.00117	0.352	0.375
PFC	Quad.	corr.	0.00677	1.3e-17	3.1e-16	-8.3e-04	1.8e-14	2.1e-13
PFC	Quad.	dim.	-0.574	6.9e-14	4.2e-13	0.0765	3.7e-13	1.8e-12
PFC	Linear	ssf	0.0122	1.4e-06	2.2e-06			
PFC	Linear	task dim.	0.252	7.6e-08	1.8e-07			
CA1	Quad.	task dim.	0.997	8.8e-11	3.0e-10	-0.108	2.8e-07	5.7e-07
CA1	Quad.	snf	0.139	2.4e-08	7.3e-08	-0.0169	8.4e-07	1.3e-06
CA1	Quad.	corr.	0.00557	2.9e-08	7.7e-08	-6.9e-04	5.7e-07	1.1e-06
CA1	Quad.	ssf	5.3e-04	0.947	0.947	1.0e-03	0.36	0.375
CA1	Quad.	dim.	-0.409	1.8e-07	4.0e-07	0.0532	7.9e-07	1.3e-06
CA1	Linear	ssf	0.0075	5.5e-04	7.0e-04			
CA1	Linear	task dim.	0.238	3.1e-06	4.3e-06			

Table 2: Parameter estimates and associated p-values for the quadratic regression models (Methods). Here, we fit mixed-effects models of geometric terms. A given geometric measure y for rat i in session t was modeled as $y_{it} = \beta_i + \gamma t + \eta t^2$. We additionally fit a linear model to the SSF term of the form $y_{it} = \beta_i + \gamma t$. All p-values were adjusted using a Benjamini-Hochberg method.

References

[1] Radoslaw Adamczak. "A note on the Hanson-Wright inequality for random vectors with dependencies". In: (2015).

[2] Najib J Majaj et al. "Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance". In: *Journal of Neuroscience* 35.39 (2015), pp. 13402–13418.

[3] Martin Raič. "A multivariate Berry–Esseen theorem with explicit constants". In: *Bernoulli* 25.4A (2019), pp. 2824–2853. DOI: 10.3150/18-BEJ1072. URL: https://doi.org/10.3150/18-BEJ1072.

[4] Carsen Stringer et al. "High-dimensional geometry of population responses in visual cortex". In: *Nature* 571.7765 (2019), pp. 361–365.

[5] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.