Rarely categorical, always high-dimensional: how the neural code changes along the cortical hierarchy

Lorenzo Posani[1,2*†], Shuqi Wang[2,3†], Samuel P. Muscinelli[1], Liam Paninski[1,3,4], Stefano Fusi[1,4*]

[1*]Zuckerman Institute, Columbia University, New York, NY, USA.
[2]School of Computer and Communication Sciences, EPFL, Street, Lausanne, Switzerland.
[3*]Department of Statistics, Columbia University, New York, NY, USA.
[4]Co-senior authors.

*Corresponding author(s). E-mail(s): lp2878@columbia.edu; sf2237@columbia.edu;
†These authors contributed equally to this work.

Abstract

A long-standing debate in neuroscience concerns whether individual neurons are organized into functionally distinct populations that encode information differently ("categorical" representations [1–3]) and the implications for neural computation. Here, we systematically analyzed how cortical neurons encode cognitive, sensory, and movement variables across 43 cortical regions during a complex task (14,000+ units from the International Brain Laboratory public Brain-wide Map data set [4]) and studied how these properties change across the sensory-cognitive cortical hierarchy [5]. We found that the structure of the neural code was scale-dependent: on a whole-cortex scale, neural selectivity was categorical and organized across regions in a way that reflected their anatomical connectivity. However, within individual regions, categorical representations were rare and limited to primary sensory areas. Remarkably, the degree of categorical clustering of neural selectivity was inversely correlated to the dimensionality of neural representations, suggesting a link between single-neuron selectivity and computational properties of population codes that we explained in a mathematical model. Finally, we found that the fraction of linearly separable combinations of experimental conditions ("Shattering Dimensionality" [6]) was near maximal across all areas, indicating a robust and uniform ability for flexible information encoding throughout the cortex. In conclusion, our results provide systematic evidence for a non-categorical, high-dimensional neural code in all but the lower levels of the cortical hierarchy.

Keywords: neural code; representations; neural geometry; dimensionality; categorical selectivity; specialization; structure; hierarchy; cortex; International Brain Laboratory; decision making

Neural responses are often complex, depend on multiple variables (mixed selectivity [6]), and, in cognitive areas, they are very diverse and seemingly disorganized (see e.g., [3, 7, 8]). One of the challenges in neuroscience is finding meaningful "structure" in seemingly disorganized data. This structure has been traditionally investigated in single-neuron responses in the form of easily interpretable tuning curves (e.g., orientation selectivity in the visual cortex or place cells in the hippocampus). However, it is now possible to record simultaneously from a large number of neurons, and new interesting structures are emerging, both in the statistics of the responses of individual neurons and at the level of population responses.

Here, we focused on both types of structure, how they are related to each other, and how they can be interpreted for their computational implications. To be more precise about the type of structure we investigated, consider the neural activity estimated in one particular time interval, recorded in different experimental conditions (e.g., in response to different sensory stimuli) [3, 9]. The responses of the different recorded neurons can be organized in a matrix (Fig. 1a), in which each row is the set of responses of a particular neuron to all the experimental conditions. This matrix defines two complementary spaces: the one spanned by the rows of the matrix, indexed by condition number, called the "conditions space" (shown in red in Fig. 1a), and that spanned by its columns, indexed by neuron number, called the "neural space" (shown in blue in Fig. 1a).

In the conditions space, individual points represent the response profile of single neurons to the experimental conditions. If groups of neurons respond similarly in these conditions, they will form clusters (functional groups) in this space, defining what is called a *categorical* representation [1–3]. For these representations, neurons with similar responses can be grouped together, and they form a distinct category of neurons with specific functional properties. In non-categorical representations, neurons can have diverse responses whose distribution does not display strong clusters.

Categorical representations have been reported in the orbitofrontal cortex (OFC) of rodents [2, 10], and monkeys [11], while non-categorical representations were found in the rodent posterior parietal cortex (PPC) [1] and in reward-sensitive frontostriatal brain areas in monkeys performing a variety of economic decision-making tasks [12]. In several other articles [6, 13–16], the authors did not study explicitly whether the representations are categorical, but the observation of very diverse mixed selectivity neurons and high dimensional representations suggests non-categorical representations (see also the Discussion). Whether, where, and how neural populations are subdivided into functional clusters is still an open debate [3].

In the neural space, spanned by the columns of the matrix, each point represents the population response to one experimental condition. The set of distances between all pairs of these points defines the *geometry* of the neural representations. Analyzing the representational geometry of population activity has been shown to provide insight into the brain's encoding strategies and their computational implications for learning and flexible behavior [6, 14–17]. For example, a high-dimensional neural geometry confers flexibility to a simple linear readout [6–8], as the same representation can support a large number of different output functions. When the dimensionality is maximal, it is possible to retrain the readout to perform any task without changing the representation. High dimensional representations also maximize memory capacity [15], while low dimensionality is associated with a better ability to create abstract representations and generalize

in novel situations [7, 14, 17]. How the dimensionality of population activity in the neural space is related to the clustering of response profiles in the conditions space is still an open question.

Here, we systematically analyzed the response profiles and the representational geometry in 43 cortical regions (Fig. 1c) in mice performing a decision-making task (IBL public Brainwide Map dataset [4], Fig. 1d), and related the functional properties of cortical areas with their anatomical ones as reported by the Allen Mouse Brain Connectivity Atlas ([5], Fig. 1e, f). As the neuronal responses have a complex temporal profile, we applied a reduced-rank regression model to characterize both the temporal and the selectivity components of the response profiles of individual neurons. We also employed different measures of embedding dimensionality [18] (*PCA* dimensionality and *shattering* dimensionality [6, 9]) to characterize the geometry of neural representations.

We found that the selectivity profiles within individual areas are clustered (i.e., "categorical") only in primary sensory areas. However, when multiple brain areas are considered, we observe clustering that reflects the brain's large-scale organization, similar to what has been observed in the monkey and human brain [19]. The response profiles of neurons in different brain regions are different enough that a decoder can determine to which area a particular neuron belongs, with accuracy increasing for less-connected pairs of regions. To study the computational implications of different degrees of clustering, we developed a mathematical theory relating categorical clustering and PCA dimensionality of neural representations, which was empirically verified on the data. Finally, the shattering dimensionality, which measures how many different classifications in the activity space can be solved by a linear readout [6], was close to maximal across all areas.

Together, these results reveal a systematic relationship between clustering and representational geometry, providing a new, large-scale perspective on neural selectivity and population coding across the cortex.

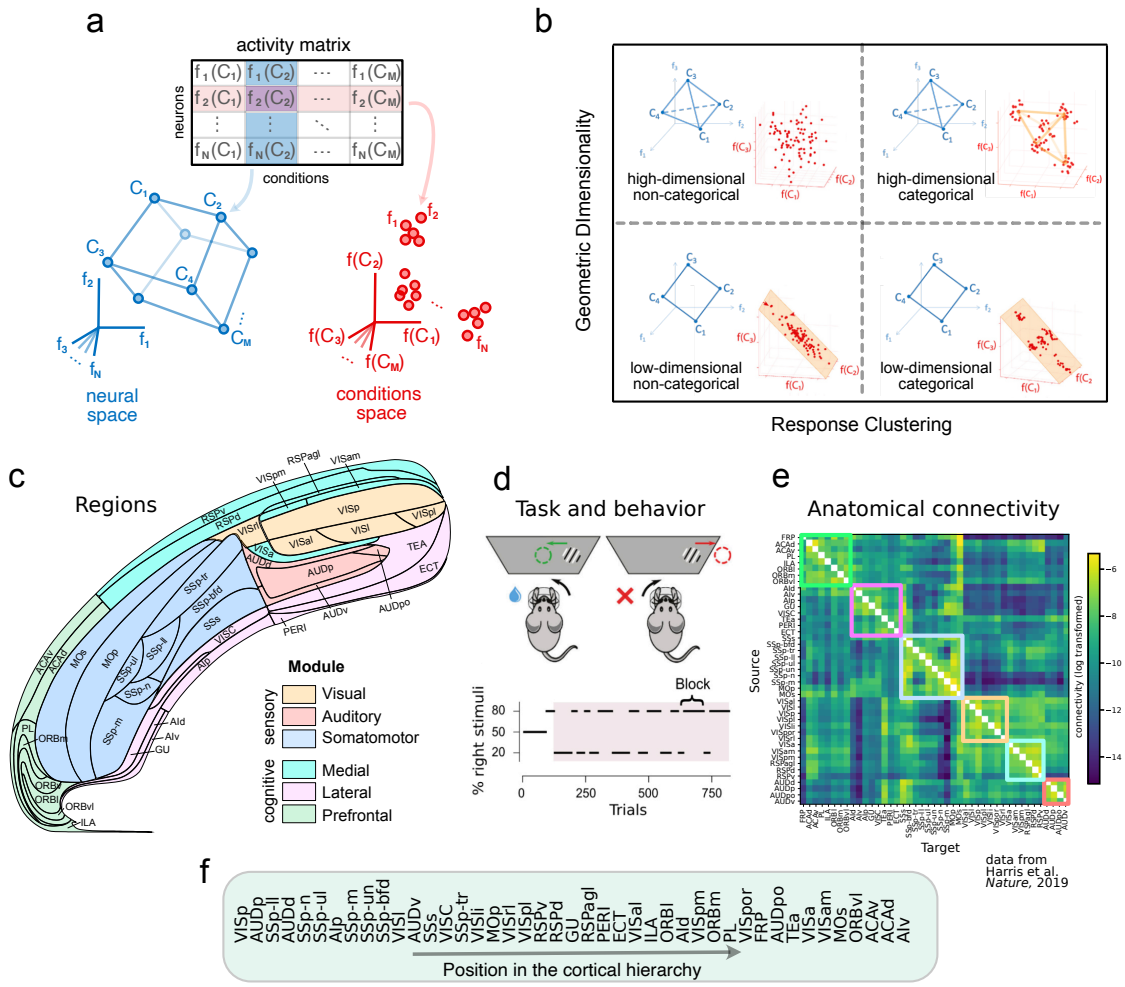


Fig. 1 Conceptual framework and data structure: (a) The matrix of neural responses of N neurons in M experimental conditions can be analyzed in either its row space (conditions space, red) or column space (neural space, blue). The relative position of conditions in the neural space defines their "representational geometry". If neurons are clustered in the conditions space, they define what is called a "categorical" representation [1–3]. **(b)** To what extent these two perspectives are related to each other is unclear, and neural populations could, in principle, occupy any portion of the clustering-dimensionality space. Image adapted from [3]. **(c)** Swanson flat map of the 43 cortical regions we analyzed. Data was recorded by the IBL consortium (IBL Brainwide Map data set) using Neuropixel probes. After selecting neurons in the cortex, we were left with $\sim 14,000$ neurons from ~ 150 recording sessions. **(d)** In the IBL task, mice need to rotate a wheel to move a visual stimulus toward the center of the screen. The stimulus appears left or right with an 80-20% biased probability in blocks of trials (bottom panel), adding "prior" contextual information to the task. **(e)** Region-to-region anatomical connectivity in the cortex, data from [5]. **(f)** Cortical hierarchy derived by the anatomical connectivity matrix of panel (e). The order was derived in [5] such that "source" regions, i.e., regions whose connectivity is unbalanced outwards, are mostly placed low, and "target" regions, i.e., regions whose connectivity is unbalanced inwards, are mostly placed high in the hierarchy. Additional details on how the hierarchy is computed can be found in [5].

Results

Characterizing the response profiles of individual neurons

We first systematically studied the response profiles of individual neurons and compared their statistics across brain regions. To characterize the response profiles, we employed a linear encoding model to estimate the effect sizes of multiple behavioral covariates of interest on single-neuron responses. Specifically, we used a reduced-rank regression (RRR) model that predicts the activity of single neurons in response to changes in a set of variables that characterize each experimental condition and the behavior of the subject (Fig. 2a, see Methods and Suppl. Fig. 1A-D). The core component of the RRR encoding model is a small set of temporal bases, learned from data and shared across neurons to describe their time-varying responses (see Methods, Suppl. Fig. 1). The sharing of the temporal bases significantly reduces the number of parameters and mitigates the issue of overfitting. Moreover, the resulting selectivity profile, composed of the estimated effect sizes of all variables, offers several advantages over the activity profiles defined by the mean responses to different experimental conditions, as it has better predictive power than the per-task condition trial average (Fig. 3-b) and mitigates the issue of unbalanced or even missing conditions.

We used this model to fit the response of each neuron (14283 neurons from 43 cortical regions) to 8 variables that describe the IBL decision-making task (Fig. 2a). In this task, subjects are shown a visual stimulus on one side of a screen and need to rotate a wheel to move the stimulus toward the center of the screen. They are given a water reward if they perform a correct wheel rotation. Stimuli are either shown on either the left or right with a 50-50 balanced probability ("50-50 block") or with an 80-20 imbalance ("left" or "right" block). In this analysis, we only used trials from left or right unbalanced blocks to avoid possible time artifacts, as the balanced block only appeared at the very beginning of each session.

The variables we used for fitting the model range from cognitive (left vs. right block), sensory (stimulus side, contrast), motor (wheel velocity, whisking power, lick), to decision-related (choice, outcome). For each trial, we considered a time window -0.2 to 0.8 seconds relative to the stimulus onset. After fitting the RRR model to predict trial-by-trial activity, each neuron is associated with a selectivity profile containing $8 \times 5 = 40$ coefficients (8 variables, 5 temporal bases). For each variable, we can then construct a time-varying coefficient by taking the weighted sum of the 5 temporal bases. These coefficients describe how sensitive the analyzed neuron activity is to each variable in trial time (Fig. 2a). The time-varying coefficients are typically stereotyped across neurons and distinct across input variables. We then summed these coefficients over time to estimate the total effect on neural responses (Methods) and obtained a time-averaged selectivity profile, which is an 8-dimensional selectivity vector (Fig. 2e, see also Suppl. Fig. 2 for examples of responses from strongly selective neurons).

We first evaluated the predictive performance of our RRR model against the average firing rate per task condition (PSTH) computed in the same time interval used to fit the model (-0.2 to 0.8 seconds relative to the stimulus onset), finding a significant improvement across the whole population of neurons (RRR: mean $R^2 = 0.16$, PSTH: mean R^2=0.09, $p \simeq 0$, Fig. 2e, see also Suppl. Fig. 3A-B for additional benchmark results of model performance and Suppl. Fig. 3C-F for

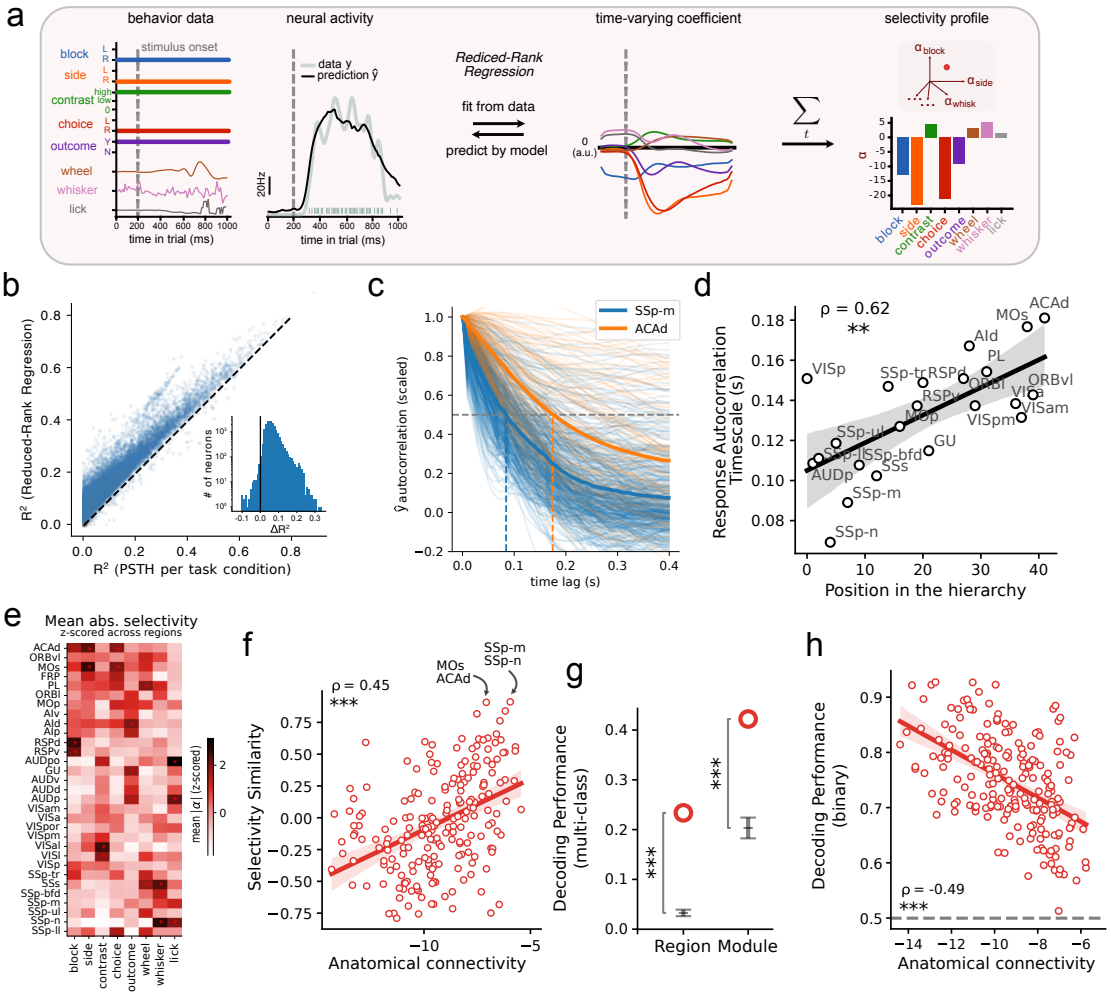


Fig. 2 The reduced-rank regression encoding model as an effective and meaningful tool to estimate single neurons' selectivity profiles. a Analysis pipeline for estimating single neurons' selectivity profiles. A linear encoding model is first fitted to describe a single neuron's temporal responses with respect to behavior data of interest (Methods). Selectivity (α) is computed as the sum of coefficients over time. Behavior data and neural activity of an example trial are shown for an example neuron, along with the estimated coefficients and selectivity. **b** Goodness-of-fit (three-fold cross-validated R^2, see Methods) achieved by our encoding model versus computing the PSTH per task condition. The outperformance (ΔR^2) is shown in the inset. **c** The autocorrelation functions of single neurons' responses to task variables estimated by the RRR (Methods). Neurons from two example brain regions (SSp-m and ACAd) are shown, along with their mean profiles (solid orange and blue lines) and estimated autocorrelation timescales (orange and blue dashed lines). **d** Correlation between the position on the hierarchy of single cortical areas with their estimated autocorrelation timescales (Methods). **e** Average (absolute) selectivity profiles for neurons in the analyzed cortical areas. Areas are first grouped by their modules (see Fig. 1) and then sorted by their hierarchy positions. The selectivity values are normalized per input variable (columns) for better visualization. The top three selective areas are marked with red stars for each input variable. **f** Correlation between region-to-region functional similarity, measured as the cosine similarity between the average selectivity profiles in panel (e) and their anatomical connectivity (data in Fig. 1e). Each dot represents a pair of cortical areas. **g** Multi-class decoding performance of the region or module labels from the time-varying response profiles of individual neurons (panel a, third panel). Decoding performance is measured as the one-vs-one multi-class accuracy of a linear SVM, while the error bars indicate the distribution of decoding performance after random label shuffling. Module labels are defined as shown in Fig. 1, following the work of [5]. **h** The binary region-to-region decoding performance using the time-varying coefficients of single neurons inversely correlates with the anatomical connectivity (data in Fig. 1e) between the two regions involved. The dashed gray line indicates the chance level for binary decoding (50%). $^{***}p < 0.001$, $^{**}p < 0.01$

visualization of example neurons). To prevent our results from being influenced by neurons that lack selectivity to any variable [4, 20], we included only neurons whose R^2 exceeded a minimum ΔR^2 threshold relative to a model that does not incorporate variables as inputs (see Methods; 4612 neurons passed the threshold). As discussed in Suppl. Fig. 7, our results are robust to different values of this threshold.

A hierarchical gradient of response timescales in the cortex

Previous and recent work have shown that neurons in regions higher up the cortical hierarchy exhibit progressively longer autocorrelation timescales, whether measured from spontaneous or task-related firing activity [21–24]. To determine if this trend extends beyond specific brain areas to the entire cortex, we analyzed the relationship between the autocorrelation timescales of neural responses to task variables – estimated using our RRR model (see Methods) – and the mouse cortical hierarchy defined by [5] from the Allen Institute Mouse Brain Connectivity Atlas. Our analysis complements prior work on intrinsic timescales of spontaneous activity using the IBL dataset [22]. We observed a significant positive correlation between a region's hierarchical position and its average response timescale (Spearman correlation=0.62, p=0.002; Fig. 2f). This finding aligns with earlier studies, reinforcing both the hierarchical organization of the cortex and the sensibility of our RRR model in capturing task-related activity.

Not "everything everywhere": response profiles reflect the anatomical organization of the cortex

Many task-relevant variables are encoded in multiple brain regions. However, this does not mean that "everything is everywhere" in the sense that there are no differences in the response properties of different brain regions. Indeed, we observed that the statistics of the response profiles of individual neurons in different regions are distinct, to the point that we can guess the region a neuron belongs to with greater-than-chance accuracy by examining its response profile.

The time-averaged response profiles averaged across the neurons in each specific brain area are reported in Fig. 2e. We then computed the cosine similarity of these selectivity profiles and compared it to the region-region anatomical connectivity from which the cortical hierarchy was derived (reported in [5] and shown in Fig. 1f). We found a significant positive correlation between selectivity similarity and anatomical connectivity (Spearman correlation $= 0.45$, $p = 6.6 \cdot 10^{-11}$, Fig. 2f), indicating that anatomically distant regions have more distinct average response profiles.

Encouraged by this result, we wondered whether selectivity profiles of individual neurons could be used to determine their region of origin. We thus labeled each neuron's selectivity profile with its region of origin and used a cross-validated multi-class decoding approach to assess to what extent this information could be decoded. We found that the region of origin could be decoded from individual neurons' selectivity profiles with above-chance accuracy using both time-varying selectivity profiles (accuracy: 0.233; null: 0.033 ± 0.003; $p \simeq 0$; Fig. 2g) and time-averaged selectivity profiles (accuracy: 0.123; null: 0.018 ± 0.002; $p \simeq 0$). Similarly, we successfully decoded the anatomical position of individual neurons on the larger scale of brain *modules* (accuracy: 0.422; null: $0.203 \pm$

0.010; $p \simeq 0$; Fig. 2g), as defined by [5], which proposed these region groupings based on anatomical connectivity and previously known functional properties. We then used the selectivity profiles of each neuron to compute the binary region-to-region decoding performance, finding that the decoding performance of a pair of regions inversely correlated with their anatomical connectivity (Spearman corr: -0.46; $p = 3.7 \cdot 10^{-14}$; Fig. 2h). Note that in all these cases, we decoded the brain region from the response profiles of individual neurons. The resulting decoding accuracy is perhaps surprisingly high, given the diversity of neuronal responses within each area, as discussed below.

Within regions, neural representations are rarely categorical

We then focused on the selectivity properties of single neurons within each area and investigated whether neurons cluster into specialized sub-populations ("categorical" selectivity) or whether information about task variables is heterogeneously distributed across the whole population ("non-categorical" selectivity).

As noted above, our RRR model associates a time-averaged selectivity profile (8-dimensional selectivity vector) to each neuron. Thus, a population of neurons within a region can be visualized as a cloud of points in an 8-dimensional selectivity space (Fig. 3a). As a measure of clustering quality, we used the silhouette score (SS) [25] of clusters identified using k-means in this space. For a given neuron, the SS compares its mean distance to neurons within the same cluster with the distance to the nearest out-of-cluster points (Fig. 3a).

Fig. 3a describes our pipeline to compute the clustering quality of a neural population: first, we used k-means to find the clustering labels. The k parameter is chosen to maximize the SS. We then computed the SS for these clusters, which we call SS_{data}. Importantly, across all the clustering analyses, we considered only clusters that are reproducible, i.e., that are not dominated by neurons from a single experimental session (see Methods). We then compared the SS observed in the data with a null model that is sampled from a single Gaussian distribution (which is unimodal, thus non-clustered by design) while preserving the mean, covariance structure, and the number of neurons of the original data. This null model was inspired by the one used for the ePAIRS test [1, 2]. By repeating many null model iterations, we could compare the SS_{data} value with a null model distribution of SSs ($\{SS_{\text{null}}\}$). As a measure of clustering quality, we took the deviation of the data point from the null distribution, measured with the z-score.

We analyzed the clustering structure of selectivity profiles in all the recorded cortical regions. To avoid small-size artifacts, we included only regions that have at least 50 neurons after our R^2 threshold was applied. In Fig. 3b, c, we show two examples of clustering results, one in a region that was found to be clustered (VISp) and one in a region that was found to be non-clustered (ACAd). To visualize the putative clusters found by k-means, we grouped neurons according to their cluster labels and sorted them according to their individual SS scores (highest at the top). As shown in Fig. 3b, the k-means algorithm clustered VISp neurons based on the block prior (c1, c4), whisking power (c2, c5), and a combination of side, contrast, and choice (c3). See also Suppl. Fig. 4 for the clustering results of all cortical regions.

The visual aspect of these representations can be misleading, as any visible structure might simply reflect the different selectivity to task variables, which could be observed even in a completely

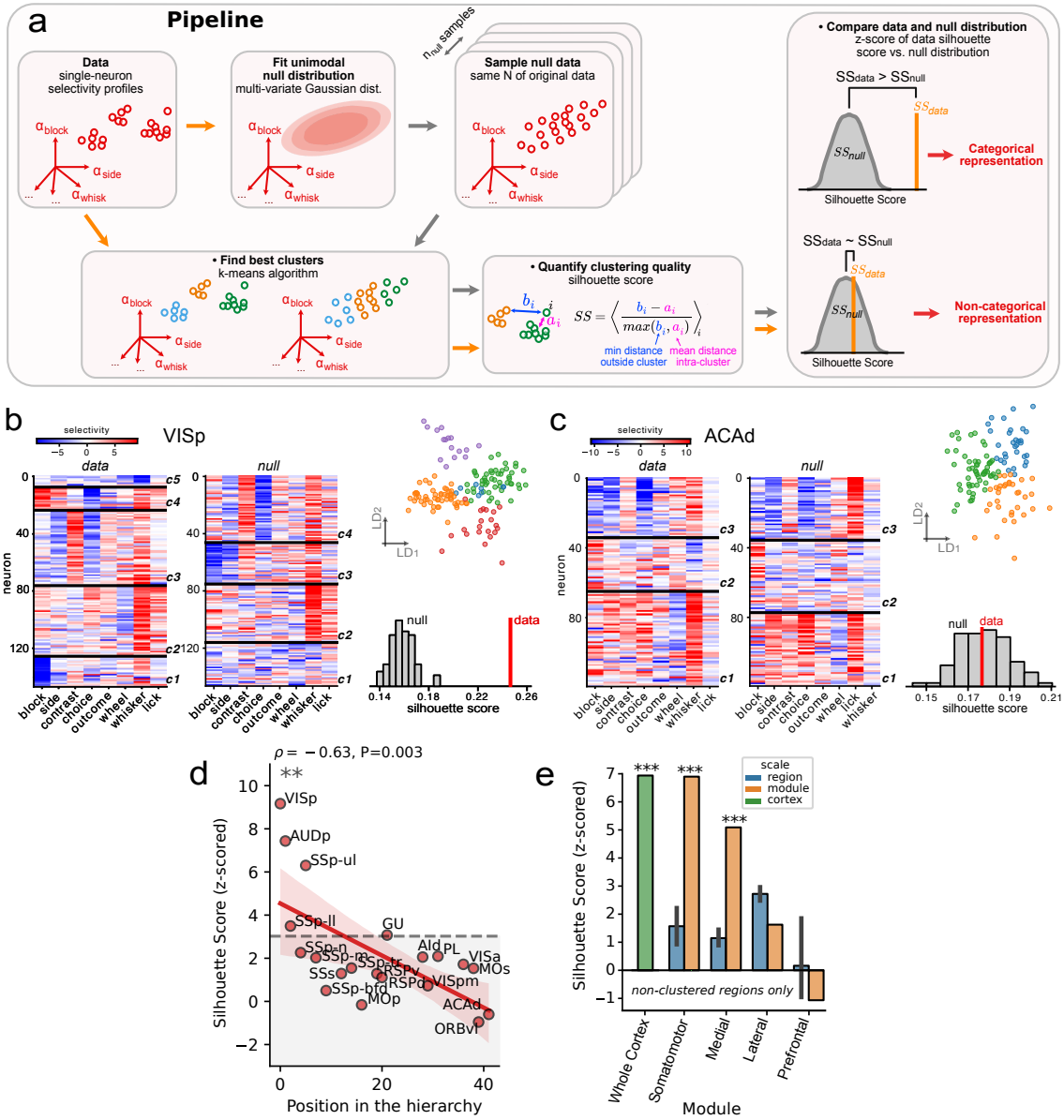


Fig. 3 The clustering quality of areas varies across the cortical hierarchy and across anatomical scales. **a** Schematic plot of our pipeline to characterize categorical versus non-categorical representations. If representations are categorical, neurons will be clustered into multiple (> 1) groups in terms of their response profiles, resulting in a silhouette score that is significantly larger than a non-categorical null model. The null model is a Gaussian distribution with a mean and covariance matrix matched to the data (see Methods and [1, 2]). For the non-categorical population, the silhouette score is comparable to the null distribution. **b,c** Examples of categorical versus non-categorical brain areas. On the left, the selectivity matrix of individual neurons is shown, with neurons sorted by the clustering labels and separated by black lines. The color indicates the selectivity value, with red being positive, blue being negative, and white being zero. In the middle, the selectivity matrix of an example null dataset is shown in the same way. The data silhouette score and its corresponding null distribution are shown at the bottom right, together with a reduced-dimensionality visualization (top right) obtained using linear discriminant analysis (individual points are neurons, colors refer to cluster labels). **d** Results of our categorical quantification pipeline in the selectivity profile space. For most cortical areas, the neuronal response profiles are very diverse, and the clustering structure is only present in primary sensory areas. **e** Clustering quality of modules (see Fig. 1e, defined in [5]) compared to their individual regions. For modules, the clustering quality was measured as in pipeline (a), using neural populations created by pooling together all the non-clustered regions (i.e., regions whose z-score did not pass a Bonferroni-corrected test for significance with a threshold of p=0.05) within a given module. These values are shown as orange bars. Blue bars show the mean and standard deviation of the clustering quality of individual regions (grouped data from panel d). The green bar shows the result for all the neurons from all the non-clustered cortical regions taken together.

unstructured null-model distribution (e.g. an elongated shape of the cloud of points in the selectivity space). In fact, after sorting, some of these clusters are still visible in a null-model sample (see c2, c3, c4 in Fig. 3b, right panel). This example showcases the necessity of comparing the silhouette score against a null model distribution to obtain a meaningful measure of clustering quality. In the case of VISp, the primary visual area, the data clusters were found to have a higher SS than the null model (Fig. 3b), indicating that neural selectivity is more categorical than a randomly distributed null model. This was not the case for ACAd, a prefrontal area, (Fig. 3c), despite some apparent clusters in the selectivity profiles mainly driven by a combination of stimulus side and choice (c3) and whisking (c1).

When applying this pipeline to all the cortical areas, we found that only four areas passed the statistical threshold of significance ($p < 0.05$ with a Bonferroni correction for multiple comparisons): the primary visual (VISp) and auditory (AUDp) areas, and two regions in the primary somatosensory area (SSp-ul, SSp-ll) (Fig. 3d). To test which variables contributed to the quality of clusters, we removed each variable individually and measured the drop in silhouette score caused by this removal. As shown in Suppl. Fig. 5, the most important variables for determining the clustering quality varied across areas and were typically multi-modal, spanning cognitive, movement, and sensory variables (VISp: block, whisking, contrast, and choice; AUDp: reward and whisking; SSp-ul: wheel and choice).

We also observed that clustering quality, measured by the z-score of the silhouette score against the null model, was significantly larger than that of the unstructured null model only in early sensory areas (Spearman correlation with position on the hierarchy: -0.63, p=0.003; Fig. 3d), suggesting that the neural code becomes more randomly mixed along the sensory-cognitive hierarchy. This negative trend was consistently found across several analysis choices such as varying the R^2 threshold for neuron inclusion (Suppl. Fig. 7b), changing the clustering algorithm (Leiden algorithm [26], Suppl. Fig. 7c), and considering the specific temporal profiles of the time-varying selectivity coefficients on top of the sum across trial time (Suppl. Fig. 7d).

More concretely, one potential weakness of the RRR model-based analysis approach is that it makes explicit assumptions about what variables are represented in the neural activity [2]. Therefore, we repeated the clustering analysis using the activity profiles in the conditions space, i.e., the space of trial-average firing rate of neurons in response to experimental conditions. We first defined a discrete set of conditions. We used the combinations of 4 variables, which were chosen so that they span different categories (sensory, motor, and cognitive) and so that each condition was well represented in the behavior: whisking motion, block prior, stimulus contrast, and stimulus side (see Methods and Suppl. Fig. 7). Continuous variables (e.g., whisking motion) were binarized to make them consistent with the other binary task variables (e.g., block prior).

Next we quantified the degree of clustering in the conditions space, with a pipeline inspired by [2]: we pre-processed the mean firing rates (z-score and dimensionality reduction), quantified the structural properties of neurons in the resulting preprocessed conditions space, and compared them to an uni-modal, non-clustered null model fitted from the data (see Methods). However, our pipeline differs from [2] in that the null distribution is fitted on the original data instead of the preprocessed firing rates (Suppl. Fig. 6a). This choice avoids the strong assumption of normality of

the preprocessed data of the original method, showing a better fit to the data (Suppl. Fig. 6b-d). Once again, we observed a negative correlation between the hierarchy and the clustering quality in this 16-dimensional conditions space (Spearman correlation: -0.52, $p = 0.039$; Suppl. Fig. 7f), with only a few regions (VISp, AUDp, SSp-n, MOp) showing more clustering than the null model.

Finally, we used ePAIRS [2] to check whether there is any additional structure in the selectivity profiles that our pipeline might not capture: ePAIRS tests for deviations from a randomly mixed selectivity distribution without explicitly conducting clustering analysis. ePAIRS also yielded few brain regions passing test for being categorical (VISp and MOs, Suppl. Fig. 7g), but showed no correlation with the hierarchy (Spearman correlation: -0.19, p¿0.05).

Together, these results provide systematic evidence that few cortical areas represent the experimental variables with a categorical representation, and become less clustered the higher they are positioned along the sensory-cognitive axis.

Mesoscale clustering in medial and somatomotor modules

The analysis of response profiles shows that, at a larger scale, the brain is functionally organized, as expected from previous studies. We should then expect more categorical representations when multiple brain regions are considered together, as each brain region would contain one or more categories of neurons with distinct response properties. We performed the same clustering analysis on modules. We ran our clustering pipeline on neural populations created by pooling together neurons from the areas of the individual cortical modules as defined in [5]: visual, auditory, somatomotor, medial, lateral, and prefrontal (Fig. 1c, e). To test whether non-categorical areas become clustered on a mesoscale, we restricted our analysis to areas that did not pass the significance threshold in the clustering analysis of Fig. 3d. Since the only areas we analyzed in the visual and auditory modules were clustered (VISp and AUDp), our analysis was restricted to four modules: somatomotor, medial, lateral, and prefrontal. After pooling different regions, we found that the representations were categorical in the medial and somatomotor modules (somatomotor: $z \simeq 7$; medial: $z \simeq 5$; Fig. 3e). Interestingly, the prefrontal and lateral modules were not found to be categorical (Fig. 3e).

Finally, when analyzing a single population spanning the entire cortex, representations were found to be categorical, reflecting the higher-order organization at the cortical level (Fig. 3e). This analysis not only confirms the existence of a functional specialization that reflects the anatomical organization of the cortex but also shows that our approach allows us to detect clustering structures and is sensitive to differences in the statistics of the neuronal response profiles.

Diverse response profiles are important for high-dimensional representations

The analysis of the response profiles revealed a highly distributed code, with responses that are diverse and progressively less clustered as we move along the cortical hierarchy. What are the implications of these different degrees of clustering for cognitive functions? We determined how the degree of clustering of neural response profiles is related to the dimensionality of the representational geometry in the neural space. Several studies have shown that representations characterized by a high-dimensional geometry in the neural space, specifically geometries with a high *embedding*

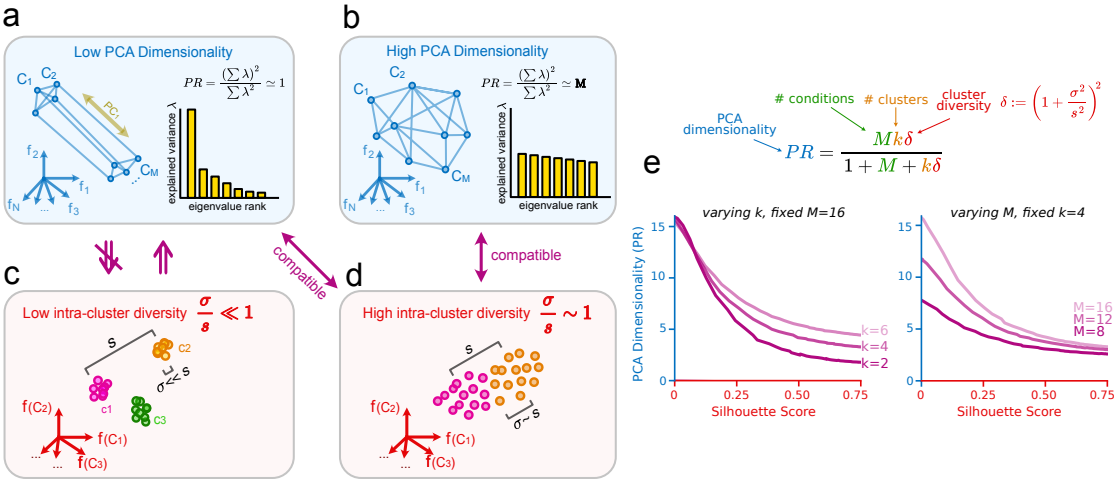


Fig. 4 Relation between clustering of neural response profiles and embedding dimensionality of neural representations (a) Schematic examples of a low-dimensional representational geometry in the neural space. As one direction dominates the covariance structure of the neural data (one data point per condition is shown), the participation ratio (PR) is low. **(b)** Schematic example of a high-dimensional geometry in the neural space. Here, the covariance structure is evenly distributed along different directions, yielding a high participation ratio. **(c)** Schematic example of a categorical representation in the space of single neuron responses to the same conditions of **a**,**b**. The intra-clustering diversity is quantified as the ratio of the dispersion of points within each cluster (denoted as σ) to the distance between different clusters (denoted as s). While low intra-clustering diversity implies a low-dimensional geometry, the opposite is not true. **(d)** Schematic example of a non-categorical representation (high intra-clustering diversity). This representation is compatible with both high- and low-dimensional geometries. **(e)** Top: analytical relation between the features of a collection of clusters in the selectivity space (number of clusters k, number of conditions M, intra-cluster diversity δ) and their PCA dimensionality (PR) in the neural space. Bottom: visualization of how the PCA dimensionality (PR) is limited by clustering quality and how this limit varies with the number of clusters k (left panel) and the number of conditions M (right panel). For these visualizations, intra-clustering diversity was converted into a silhouette score to aid comparison with the data (see Methods).

dimensionality [9], are important for cognitive flexibility [6, 7, 27] or high memory capacity [15]. Here, we link the statistics of the response profiles of individual neurons to an important aspect of the representational geometry - its dimensionality - showing that clustering, which limits diversity, reduces the maximal dimensionality that is achievable.

One measure of dimensionality is the *PCA dimensionality*, which is the number of independent dimensions needed to linearly *embed* the geometry in the neural space. The PCA dimensionality of a set of patterns of neural activity can be expressed as the *participation ratio* (PR) (see e.g. [28, 29]), which is a measure of the sparseness of the eigenvalues of the covariance matrix of the patterns. The participation ratio will be low if a few eigenvalues dominate (suggesting that only a few dimensions explain most of the variance, see Fig. 4a) and high if the eigenvalues are more evenly distributed (Fig. 4b).

Diversity in neural responses is required to achieve a high embedding dimensionality in the neural space [6, 7, 27]. Consider a scenario in which the response profiles of neurons are organized into $k < M$ small, well-defined clusters whose centers are randomly positioned in the M-dimensional conditions space (Fig. 4c). Neurons in the same cluster will have correlated response profiles. Thus, if we were to perform a PCA analysis on these response profiles, we would find that a high portion

of the variance can be explained by the first k principal components, one for each cluster, yielding a low participation ratio $PR \sim k < M$. As explained in Fig. 1a, response profiles can be thought of as rows of a matrix X whose columns define the geometry of conditions in the neural space. Since the eigenvalue spectrum of a matrix and its transpose are the same, their dimensionality must also be the same, i.e., $PR(X) = PR(X^T)$. Thus, in this strongly clustered example, the embedding dimensionality of the geometry in the neural space will also be constrained to $PR \sim k$. Note that the converse is not true, i.e., a low embedding dimensionality does *not* imply clustering of the response profiles. Neurons could be organized in a low-dimensional manifold in the conditions space without necessarily forming well-defined clusters (see the lower-left corner of Fig. 1b and [3]). For the same reason, non-clustered response profiles are compatible with both high- and low-dimensional neural geometries (Fig. 4d). Finally, a well-clustered yet high-dimensional case is also possible, but it requires a number of clusters of the same order as the number of conditions $k \sim M$.

This argument provides a qualitative scheme of the relation between categorical clustering and dimensionality in the neural space, summarized in Fig. 4(a-d). To be quantitative, we developed a mathematical model that allowed us to analytically derive the PCA dimensionality in the neural space as a function of the *intra-cluster diversity*, defined as the standard deviation of points around each cluster center divided by the distance between cluster centers (Fig. 4c,d). In the assumption of clusters with randomly placed centers and Gaussian dispersion (see Methods), we can analytically derive the participation ratio of the resulting activity matrix X as

$$\mathrm{PR} = \frac{M\ k\delta}{1 + M + k\delta} \quad , \tag{1}$$

where M is the number of conditions, k is the number of clusters, and δ is a notation-convenient *intra-clustering diversity* defined as

$$\delta := \left(1 + \frac{\sigma^2}{s^2}\right)^2 \quad , \tag{2}$$

where s is the average distance between cluster centers and σ is the standard deviation of the Gaussian dispersion around each cluster center (in the model, this is the same for all clusters). Note that this model assumes that the positions of these clusters in the conditions space are random (i.e., there is no additional structure besides clustering). Hence, we can imagine the participation ratio we derived as an upper limit to the PCA dimensionality (see Methods and Discussion).

This model makes precise predictions on how the dimensionality is limited by clustering. As shown in Fig. 4e, the allowed PCA dimensionality, as expected by the intuition above, decreases with clustering quality while increasing with the number of clusters k and the number of conditions M (note that we converted the intra-cluster diversity in its corresponding silhouette score to allow comparison with the data). In summary, while a low number of small, well-defined clusters constrains the PCA dimensionality of the geometry in the neural space to be low, highly diverse clusters allow (but do not imply) higher dimensionalities.

Testing the theory: the importance of estimating independent conditions

Our theory can be tested by computing the PCA dimensionality of experimental conditions in the neural space and comparing it with the PR predicted by Eq. 1 when plugging in the clustering properties of the conditions space, i.e., M, k, and δ. While the k-means algorithm gives the number of clusters k, and the intra-cluster diversity δ can easily be measured in the conditions space (see Methods), it is less trivial to measure M, the number of conditions. Indeed, we should use the number of *distinguishable*, or *independent* conditions that are encoded in the neural activity of a particular area, which we call M_{IC}. For example, if an area only represents a single variable (say, stimulus side), the effective number of independent conditions will be $M_{IC} = 2$ (left and right). As the number of conditions sets an upper bound on the dimensionality (assuming $N > M_{IC}$, see also [28]), M_{IC} will act as an upper limit to the PR. This point is explained in Fig. 5. In Fig. 5b, we have $M_{IC} = 4$ independent conditions, which are represented as four small clouds in the neural space representing the noisy responses to four conditions. The conditions are distinguishable from each other as the clouds are well separated, so the max dimensionality in general is $M - 1 = 3$. However, because of the specific geometric arrangement of the clouds, the measured PCA dimensionality is lower (PR $\simeq 1$). The case depicted in Fig. 5c is characterized by a similar PCA dimensionality, PR $\simeq 1$; however, here, the neural code only represents two *independent* conditions (green+blue clouds vs. red+purple clouds). So, in both cases, the dimensionality is $PR \simeq 1$, but in the second case, it is the largest possible dimensionality, whereas in the first case, it could legitimately be called a *low-dimensional* representation because the maximal one is 3.

To estimate M_{IC} in the data, we started by selecting the 16 conditions defined by four binary variables that span cognitive, sensory, and movement information while being well represented in the experiment: (1) stimulus side, (2) whisking power (binarized), (3) block, and (4) stimulus contrast (binarized) (Fig. 5a, see Methods). By experimental design, variables like context and stimulus side are highly correlated in the biased blocks of trials that we analyzed. Hence, adding additional variables (such as choice or reward) was impractical due to the very limited number of trials for certain variable combinations (e.g., errors in context-side matched trials).

We then estimated M_{IC} using an iterative algorithm based on the cross-validated decoding performance of a linear classifier trained to report the condition label of individual trials from their corresponding time-averaged population activity vectors (Fig. 5b, c, see Methods). The number of independent conditions varied widely across different cortical regions, ranging from a minimum of $M_{IC} = 5$ (SSp-n) to a maximum of $M_{IC} = 16$ (MOs). M_{IC} increased significantly along the cortical hierarchy (Spearman corr: 0.77, p=0.0005; Fig. 5d). Thus, cortical regions encode an increasing number of combinations of task variables as they move higher up the cortical hierarchy.

The PCA dimensionality of neural representations increases along the cortical hierarchy

We next analyzed the PCA dimensionality of the representational geometry of the M_{IC} independent conditions in the different cortical regions. Our theoretical analysis suggests that PR should increase with M_{IC}, which was found to increase along the hierarchy (Fig. 5d) and decrease with

clustering quality, which was found to decrease along the hierarchy, (Fig. 3g and Suppl. Fig. 7). Thus, we expect the PCA dimensionality to be larger in cognitive areas than in primary sensory ones. Our analysis revealed that this is indeed the case, as we found a positive correlation across cortical areas between their PR and their position along the hierarchy (Spearman corr: 0.79, p=0.0003; Fig. 5e). The PCA dimensionality was observed to be inversely correlated to clustering quality across cortical areas (Spearman corr: -0.77, p=0.0004; Fig. 5f). This result, compatible with the predictions from the theory, suggests that areas with more categorical representations, such as primary sensory areas, exhibit a lower PCA dimensionality, while highly diverse cognitive areas encode task conditions with a higher-dimensional geometry.

We are now equipped to test the predictive power of our analytical result. Using the form in Eq. 1, we compared the value of PR found in the data with that predicted by the theory as a function of M_{IC}, using k, and δ from the clusters found by k-means in the conditions space. Strikingly, the theoretical prediction of PR is not only correlated (Spearman corr: 0.80, p=0.0002; Fig. 5g), but it actually matches individual data points (linear regression slope=1.02 ± 0.02 SE). We also tested whether the cluster diversity δ could be predicted by our analytical theory using the values M_{IC} and PR found across cortical regions. Again, our theory proved to be able to quantitatively predict the intra-cluster diversity δ (Spearman corr: 0.80, p=0.0002; Linear regression slope: 0.97 ± 0.03; Fig. 5h). Together, these results validate our new theoretical framework, providing a quantitative relation between the clustering of neural responses and the dimensionality of neural representations.

The shattering dimensionality of neural representations is close to maximal across all regions

Dimensionality is important because it can restrict the type of outputs that can be generated by a downstream readout when the complexity of the readout mechanism is limited, as in the case of a downstream neuron, which we can model as a linear readout. In particular, high dimensional representations allow such a readout to separate the experimental conditions into two arbitrary classes and respond differently to them, providing a computational substrate to a diverse array of classification and decision-making tasks.

There are geometries that, according to PCA dimensionality, are low dimensional but still enable a linear readout to separate all the conditions arbitrarily: in Fig. 6a, we show two geometries with similar PR (low PCA dimensionality) but different properties in terms of separability. In the first one (top panel), conditions are all distinguishable, but they are represented as small clouds placed along a line in the activity space (low PCA dimensionality). This arrangement is disruptive for flexibility: out of the three possible balanced dichotomies (ways to divide the conditions into two equally sized groups), only one is separable by a linear readout (Fig. 6a). In the second case (bottom panel), the geometry is still stretched along a particular coding direction, but the centroids of the four conditions define a stretched tetrahedron. As the PR considers the sparseness of the eigenvalues of the covariance matrix, this representation still has a low PCA dimensionality because one eigenvalue is much larger than the others. However, this geometry differs from the one above in

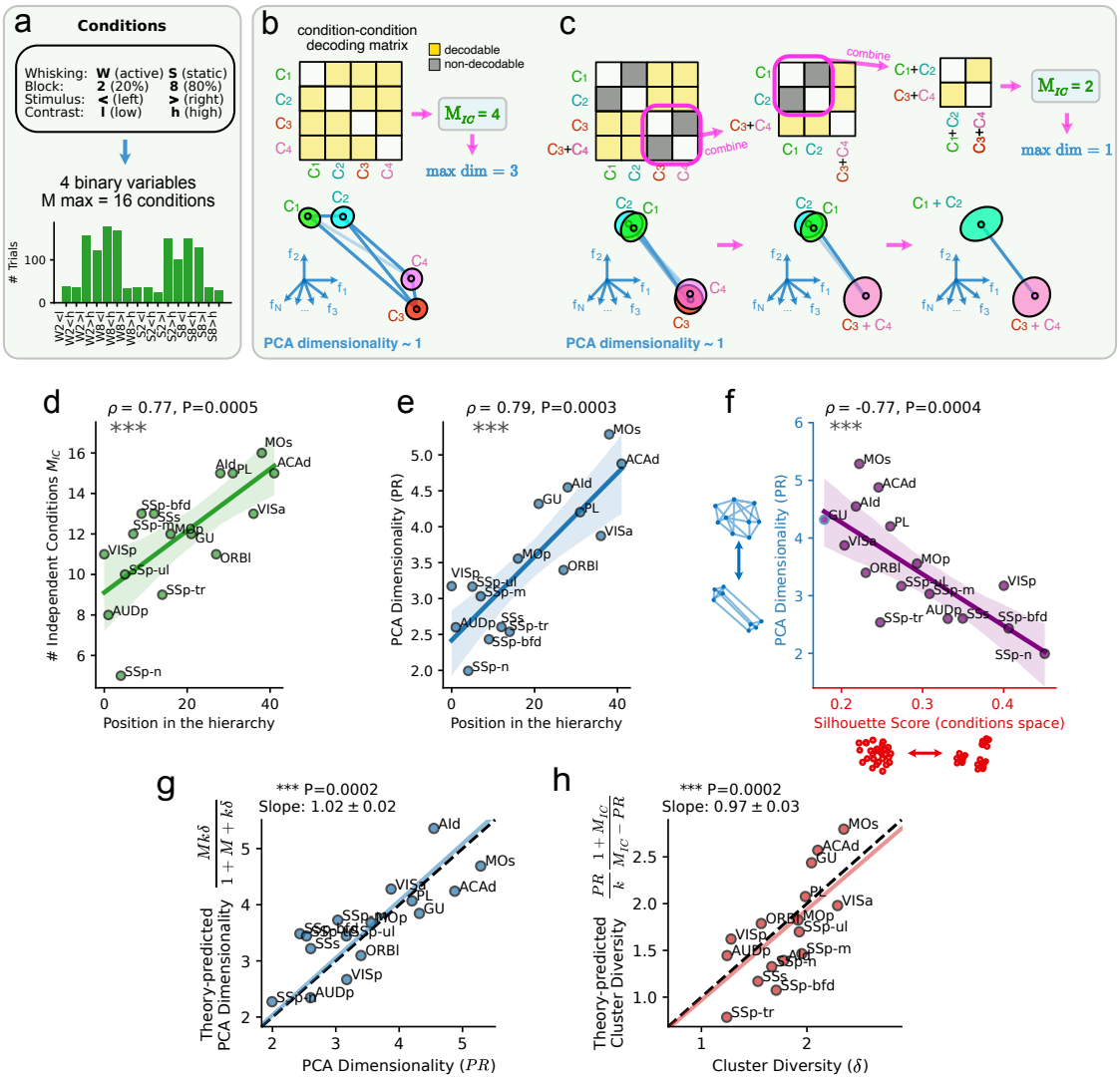


Fig. 5 Analysis of the representational geometry of independent conditions in the neural space. (a) To select conditions, we chose the largest set of combinations of motor, sensory, and cognitive variables that are well represented in the behavior. The bottom panel shows an example of the number of trials for each condition in a session. **(b,c)** Schematic of the algorithm to find the number of independent conditions in the data, with two examples of geometries with a similarly low PCA dimensionality but with a different number of independent conditions ($M_{IC} = 4$ in **(b)** and $M_{IC} = 2$ in **(c)**). See Methods and Suppl. Fig. 8. **(d)** The number of independent conditions was found to vary across areas and increase along the hierarchy. **(e)** The PCA dimensionality, estimated by the PR of the M_{IC} independent conditions, increased significantly along the cortical hierarchy. **(f)** The PCA dimensionality inversely correlated with the clustering quality measured in the space of the mean firing rates of the 16 conditions (data from Suppl. Fig. 7g-h). **(g)** Correlation between the value for PR predicted by the relation in Eq. 1 using k, δ and M_{IC} (y-axis) and that computed from the centroids of the M_{IC} conditions in the neural space. The blue line is the best fit zero-intercept slope, while the dashed black line is the identity function $y = x$. **(h)** Similar to **g**, the intra-cluster diversity δ in the conditions space could also be quantitatively predicted from M_{IC}, PR, and k using an inversion of Eq. 1.

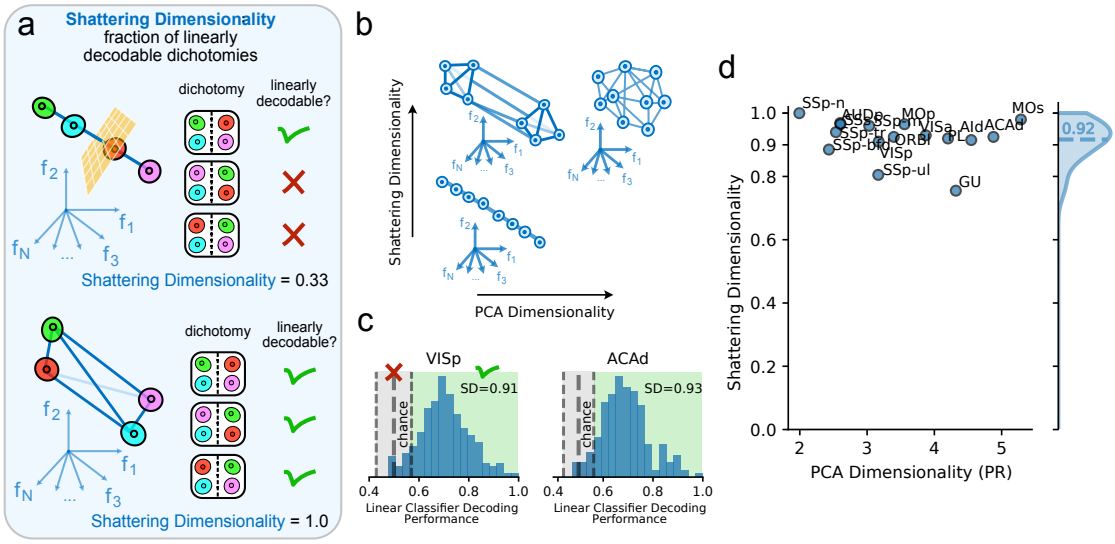


Fig. 6 The shattering dimensionality of neural representations is close to maximal across all regions **(a)** Definition of Shattering Dimensionality (SD) and schematic representations of two geometries with similar Participation Ratio (PR) and different SD. In this schematic case, both geometries represent $M = 4$ independent conditions. However, in the top panel, only one dichotomy of these conditions is linearly separable (green-blue vs. red-purple), resulting in an SD value of 1/3. In the bottom case, all dichotomies are linearly separable, resulting in a maximal SD= 1. **(b)** The combination of SD and PR can be used to more precisely characterize the representational geometry of the encoded conditions. **(c)** Two examples of the distribution of linear decoding accuracies for $n = 200$ random dichotomies defined over the M_{IC} independent conditions encoded by two regions at the opposite sides of the hierarchy (VISp and ACAd). Here, a dichotomy is defined as *decodable* if its associated cross-validated decoding performance is larger than the 95th percentile of a null model population of decoding performances obtained by shuffling the labels of individual conditions (see Methods). The SD is defined as the fraction of decodable dichotomies. **(d)** Using this definition of SD, all regions were found to represent their independent conditions with close to maximal dimensionality (mean SD $= 0.92 \pm 0.06$ STD). Dashed lines in panel **c** show the 5th percentile, the mean, and the 95th percentile of the null distribution.

that, here, all possible dichotomies of conditions are linearly separable (Fig. 6b), assuming that the trial variability (the clouds' "noise" around the centroids) is smaller than the shortest distances.

One measure of dimensionality that directly quantifies the linear separability of the conditions is the "shattering dimensionality" (SD) [6, 14], which is defined as the fraction of dichotomies that are decodable using a cross-validated (across trials) linear classifier. Specifically, the linear classifiers are trained to report the dichotomy label (0 or 1, depending on each condition) from the population activity vectors of individual trials (see Methods). In the example of Fig. 6a, the first geometry has SD= 0.33, while the second one has maximum dimensionality (SD= 1.0). The combination of SD and PR allows for a more comprehensive quantification of the representational geometry (Fig. 6b).

We measured the shattering dimensionality of the M_{IC} independent conditions represented in the population activity of the different regions in the cortex. We sampled $n = 200$ random balanced dichotomies (binary labels on conditions) and used a linear classifier to decode the corresponding labels from single trial population vectors. We computed the shattering dimensionality (SD) as the fraction of *decodable* dichotomies, i.e., those dichotomies whose decoding performance was

larger than the 95th percentile of a null distribution obtained by shuffling the condition labels of individual trials (gray area in Fig. 6c, see Methods). To our surprise, we found no difference in shattering dimensionality between sensory and cognitive areas (see two examples at the ends of the hierarchy (VISp and ACAd) in Fig. 6c). For almost all regions, the shattering dimensionality was found to be close to maximal (average $SD = 0.92$, Fig. 6d, see also Suppl. Fig. 10).

This result shows that all regions, despite encoding a different number of conditions and with varying clustering quality, represent these conditions in the maximum possible dimensionality, maximizing their readout flexibility.

Discussion

The brain has a clear anatomical organization, and, not too surprisingly, we observe that it is reflected in the organization of the response profiles of individual neurons in different brain areas (functional organization). However, when one looks at the neurons within each brain area, only the responses of some primary sensory areas seem to be organized into functional clusters (categorical representations). We showed in a simple model how these two different aspects of the representations are related to each other: categorical clustering limits the dimensionality of the representational geometry in the neural space, making cluster diversity necessary for high-dimensional representations.

Finally, all our analyses revealed that several aspects of the representational geometry and the statistics of the neuronal response profiles vary in a systematic way along the cortical hierarchy: in particular, clustering decreases with the position in the hierarchy, and the PCA dimensionality increases, together with the number of independent conditions. All these results show that the structure in the response profile space is related to the geometrical structure of the activity space. This relation can help us understand the computational implications of the neuronal response properties. All the analyses we performed can be applied to non-cortical regions, which is one of our future directions.

Why is the shattering dimensionality maximal in all brain areas?

Dimensionality is directly linked to several computational properties of neural networks, which include flexibility, memory capacity, and the ability to solve the binding problem [7, 8, 30]. In feed-forward multi-layer artificial neural networks, it is important to have high dimensional representations in the last layer, because low dimensionality would severely restrict the number of input-output functions that the output units can implement. However, this is not a strict requirement for intermediate layers.

So why is maximal dimensionality so ubiquitous in the mouse cortex? One possible computational reason is related to the strong recurrent connectivity of the cortex. In artificial recurrent neural networks (RNNs), many of the computational properties depend on the dimensionality of the representations. One famous example is the memory capacity of the Hopfield network [31], which is rather limited for low-dimensional memories [15, 32]. Moreover, any RNN that needs to transition to a new state that depends on both the previous state and the external input requires

neurons to non-linearly mix these two sources of information, which otherwise would constitute a concatenated input/recurrent state representation that is typically low dimensional [8, 27].

Fortunately, increasing the dimensionality is relatively easy as non-linear random projections already do a surprisingly good job, both in RNNs [27, 33–35] and feed-forward networks [29, 36, 37]. More generally, if a network is initialized with random connectivity and the parameters are properly tuned, it is likely that the representations are already as high dimensional as they can be. Learning can certainly improve generalization and robustness to noise, but starting from high-dimensional representations is not that difficult. Moreover, high shattering dimensionality does not imply a complete absence of structure. Representations can have the generalization properties of low dimensional disentangled representations and still possess the maximal shattering dimensionality [14]. So, a learning process can lead to other interesting computational properties typically associated with low-dimensional representations without compromising the flexibility conferred by high-shattering dimensionality.

Notice that in our article and this Discussion, we always refer to the embedding dimensionality of the set of points representing different experimental conditions. This dimensionality is close to maximal. Instead, when the points along a trajectory for a single condition are analyzed, the representations are typically low dimensional [38]. We did not perform this analysis, but we would not be surprised to see low dimensional trajectories in the IBL dataset.

What are the implications of maximal shattering dimensionality for single neuron response properties?

Single neuron response properties and the geometrical structure of the activity space are related to each other. In particular, as we have shown mathematically, highly clustered representations can limit the dimensionality of representations, even when the positions of the clusters are random. Thus, to achieve the maximal dimensionality, the neuronal responses need to be diverse enough, which often means that they exhibit mixed selectivity [7, 8]. Our analysis of the response profiles revealed that they are indeed very diverse, as previously observed in other studies (see e.g. [1, 6, 12, 13]): even when there is some significant clustering, there is always the possibility that there is some additional structure within each cluster, given the cluster size. This diversity of responses within each cluster can greatly contribute to increasing the shattering dimensionality, and hence, this diversity is potentially computationally important. As we discussed above, a maximal shattering dimensionality does not mean a complete absence of structure in the representational geometry. Analogously, it does not necessarily mean that the neuronal responses are completely random, but only that the responses are sufficiently diverse. Clustering is one form of structure in the response space which can coexist with high dimensionality.

Modularity and cell types.

In the brain, there are several different neuronal types [39], highlighting a biological complexity that is especially marked for inhibitory neurons [40]. It is not unreasonable to expect that these different types of neurons would cluster in the response profile space. However, our analysis shows that only primary sensory areas are clustered, and even in these areas, the clusters are not well

separated. One possible explanation is that the discrete nature of neuronal type is not directly reflected in their functional response to task variables. Another possibility is that the response profile reflects neuronal types that are not discrete or well separated. A recent article from the Harris-Carandini lab [41] analyzed the transcriptomic profile of neurons in the visual cortex and found a relation between the position of each neuron along a "transcriptomic axis" (found as the principal component in the high-dimensional space of 72 genes) and their activity modulation in different behavioral states of the subject. Crucially, while clustering in the transcriptomic space was found to correlate with putative cell types, this axis was defined as a "continuum," hence, without clear-cut clusters. The transcriptomic space looks only partially clustered, with some different neuronal types arranged in continuous filaments. Further studies will be needed to assess whether other structures emerge in experiments with more complex tasks (see also the last section of the Discussion). Future work will also explore the relationship between functional tuning (e.g., the RRR coefficients) and cell type embeddings estimated from spike waveform, autocorrelation [42], and eventually also transcriptomic/anatomical identity.

Clustering of response profiles in other species.

The recent article by the Kanwisher group [19] performed an analysis that also systematically analyzes structures in the space of neuronal response profiles. Interestingly, they analyzed the visual and auditory systems of humans (fMRI) and monkeys (electrophysiology). Consistent with our results, when they looked at mesoscale brain structures, they found privileged neuronal axes that are preserved across individuals and reflect the large-scale organization of the brain. Their result is compatible with our observation in Fig. 3f, which shows clear clustering in the case of mesoscale sensory brain structures. Instead, when they repeated the analysis on a more local scale, within category-selective regions of the high-level visual cortex, they did not find the same structure, and they reported a distribution of the response profiles that would be comparable to the ones of our null models (these results are preliminary and mentioned in their Discussion). As the authors say, it is possible that this is a limitation of the resolution of fMRI, and that interesting structures emerge in electrophysiological recordings.

The computational advantages of clustered response profiles.

The brain is highly organized into functional and anatomical structures, which can be considered "modules" of the brain. This large-scale organization is well known; it has computational implications, and its emergence can be explained using general computational principles, at least in the visual system [43–45]. However, when one looks inside a particular brain area, the picture is less clear, though in several cases, it is possible to identify some form of modularity, which would lead to clustering in the response profile space. Two main forms of modularity have been studied in theory and experiments. The first modularity is observed for specific forms of disentangled representations that are aligned to the neuronal axes. Disentangled abstract representations are widely observed in the brain [14, 15, 17, 46] and are important for generalization. However, this geometry is compatible with both highly diverse neuronal responses (e.g., in the case of linear mixed selectivity) and with more organized categorical representations with modular structures in the

neuronal response profile space. In the second case, each relevant variable would be represented by a segregated population of neurons (module), which would be a cluster in the selectivity space. This specific type of modular representation has been shown to be computationally equivalent to non-modular ones but more efficient in terms of energy consumption and number of required connections, at least under some assumptions [19, 47].

A second type of modularity has been characterized in artificial neural networks that are trained to perform multiple tasks [48, 49] or operate in different contexts [9, 50, 51]. In this case, it is possible to observe modularity, even when no additional constraints are imposed on metabolic costs and a number of connections [51]. In particular, two subtypes of modularity are considered: 1) *explicit* modularity, for which there are segregated populations of neurons that are activated in different contexts. In the activity space, this would imply that the conditions of different contexts are represented in orthogonal subspaces. Interestingly, the same geometry would be observed in the response profile space, and these representations exhibit clustering (i.e., they are categorical). 2) *implicit* modularity, for which the geometry in the activity space is the same as in the case of explicit modularity (it is a rotated version of the explicit modularity case), but in the response space, it is difficult to say whether clustering would be observed or not.

All these different types of modular representations share similar computational properties and allow us to generalize more easily, learn new structures more rapidly [51], and even enable some form of compositionality, in which the dynamics of sub-circuits can be reused in different tasks [49].

Limitations of our study.

Given all these computational advantages of modular structures, and the fact that categorical representations have been observed in some experiments [2, 10, 11], it is surprising that we observe significant clustering only in a few sensory areas. One possible explanation is that the assumptions made in the theoretical studies are incorrect and that the biological brain operates in different regimes. For example, the metabolic advantage of modular representation could be too modest compared to the enormous baseline consumption [52].

However, there are also other possible explanations. We looked for clustering using three ways of characterizing the response profile of each neuron (conditions space, regression coefficients of RRR, and time-averaged regression coefficients), two different algorithms for finding clusters (k-means and Leiden algorithm), and two statistical tests (one based on silhouette score, and one on ePAIRS). It is possible that other approaches will reveal some form of clustering or entirely different interesting structures in the statistics of the response profiles [53].

Finally, perhaps the most notable limitation of our study is that the IBL task is relatively simple, and the recorded animals are trained (in this case, often over-trained) to perform a single task. It is possible that repeating our analysis on a dataset that involves multiple tasks would actually reveal more clustering and specific modular structures. Indeed, Dubreil, Valente, et al. [50] showed that, in RNNs, simple tasks do not require clusters while complex tasks do. Similar considerations apply to other theoretical studies [49, 51]. Future studies on multiple complex tasks will reveal whether the organizational principles that we identified are more general and valid in experiments that are closer to real-world situations.

Competing interests

Authors have no competing interests to report.

Funding

This work received support from NIH RF1AG080818 and U19NS123716, the Simons Foundation, the Kavli Foundation, the Gatsby Foundation GAT3708, and the Swartz Foundation. LP was supported by the NIH 1K99MH135166-01 grant.

Acknowledgments

We are grateful to Fabio Stefanini, Mattia Rigotti, Marcus Benna, Ramon Nogueira, and Genevera Allen for numerous discussions about categorical representations and their analysis, to Tatiana Engel for the deep exchange of perspectives on dimensionality and clustering, and to our colleagues in the IBL for many fruitful discussions. We thank Yizi Zhang and Ji Xia for a code review of the RRR encoding model. LP is thankful to Jeff Johnston and James Whittington for their help in co-organizing the workshop "Are Neurons Interpretable?" (COSYNE, 2023), which played a big role in inspiring this work.

References

[1] Raposo, D., Kaufman, M. T. & Churchland, A. K. A category-free neural population supports evolving demands during decision-making. *Nature neuroscience* **17**, 1784–1792 (2014).

[2] Hirokawa, J., Vaughan, A., Masset, P., Ott, T. & Kepecs, A. Frontal cortex neuron types categorically encode single decision variables. *Nature* **576**, 446–451 (2019).

[3] Kaufman, M. T. *et al.* The implications of categorical and category-free mixed selectivity on representational geometries. *Current opinion in neurobiology* **77**, 102644 (2022).

[4] Laboratory, I. B. *et al.* A brain-wide map of neural activity during complex behaviour. *Biorxiv* 2023–07 (2023).

[5] Harris, J. A. *et al.* Hierarchical organization of cortical and thalamic connectivity. *Nature* **575**, 195–202 (2019).

[6] Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).

[7] Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology* **37**, 66–74 (2016).

[8] Tye, K. M. *et al.* Mixed selectivity: Cellular computations for complexity. *Neuron* (2024).

[9] Ostojic, S. & Fusi, S. Computational role of structure in neural activity and connectivity. *Trends in Cognitive Sciences* (2024).

[10] Hocker, D. L., Brody, C. D., Savin, C. & Constantinople, C. M. Subpopulations of neurons in lofc encode previous and current rewards at time of choice. *Elife* **10**, e70129 (2021).

[11] Onken, A., Xie, J., Panzeri, S. & Padoa-Schioppa, C. Categorical encoding of decision variables in orbitofrontal cortex. *PLoS computational biology* **15**, e1006667 (2019).

[12] Blanchard, T. C., Piantadosi, S. T. & Hayden, B. Y. Robust mixture modeling reveals category-free selectivity in reward region neuronal ensembles. *Journal of neurophysiology* **119**, 1305–1318 (2018).

[13] Hardcastle, K., Maheswaranathan, N., Ganguli, S. & Giocomo, L. M. A multiplexed, heterogeneous, and adaptive code for navigation in medial entorhinal cortex. *Neuron* **94**, 375–387 (2017).

[14] Bernardi, S. *et al.* The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967 (2020).

[15] Boyle, L. M., Posani, L., Irfan, S., Siegelbaum, S. A. & Fusi, S. Tuned geometries of hippocampal representations meet the computational demands of social memory. *Neuron* **112**, 1358–1371 (2024).

[16] O'Neill, P.-K. *et al.* The representational geometry of emotional states in basolateral amygdala. *bioRxiv* 2023–09 (2023).

[17] Courellis, H. S. *et al.* Abstract representations emerge in human hippocampal neurons during inference. *Nature* 1–9 (2024).

[18] Jazayeri, M. & Ostojic, S. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current opinion in neurobiology* **70**, 113–120 (2021).

[19] Khosla, M., Williams, A. H., McDermott, J. & Kanwisher, N. Privileged representational axes in biological and artificial neural networks. *bioRxiv* 2024–06 (2024).

[20] Steinmetz, N. A., Zatka-Haas, P., Carandini, M. & Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576**, 266–273 (2019).

[21] Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nature neuroscience* **17**, 1661–1663 (2014).

[22] Zeraati, R., Shi, Y., Levina, A., Engel, T. *et al.* *A census of neural timescales across the mouse brain* (2024).

[23] Song, M. *et al.* Hierarchical gradients of multiple timescales in the mammalian forebrain. *bioRxiv* 2023–05 (2023).

[24] Rudelt, L. *et al.* Signatures of hierarchical temporal processing in the mouse visual system. *PLOS Computational Biology* **20**, e1012355 (2024).

[25] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53–65 (1987).

[26] Traag, V. A., Waltman, L. & Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 1–12 (2019).

[27] Rigotti, M., Rubin, D. B. D., Wang, X.-J. & Fusi, S. Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Frontiers in computational neuroscience* **4**, 24 (2010).

[28] Gao, P. & Ganguli, S. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology* **32**, 148–155 (2015).

[29] Litwin-Kumar, A., Harris, K. D., Axel, R., Sompolinsky, H. & Abbott, L. Optimal degrees of synaptic connectivity. *Neuron* **93**, 1153–1164 (2017).

[30] Johnston, W. J., Fine, J. M., Yoo, S. B. M., Ebitz, R. B. & Hayden, B. Y. Semi-orthogonal subspaces for value mediate a binding and generalization trade-off. *Nature Neuroscience* 1–13 (2024).

[31] Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences* **79**, 2554–2558 (1982).

[32] Amit, D. J. *Modeling brain function: The world of attractor neural networks* (Cambridge university press, 1989).

[33] Jaeger, H. & Haas, H. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *science* **304**, 78–80 (2004).

[34] Maass, W., Natschläger, T. & Markram, H. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation* **14**, 2531–2560 (2002).

[35] Buonomano, D. V. & Maass, W. State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience* **10**, 113–125 (2009).

[36] Barak, O., Rigotti, M. & Fusi, S. The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. *Journal of Neuroscience* **33**, 3844–3856 (2013).

[37] Babadi, B. & Sompolinsky, H. Sparseness and expansion in sensory representations. *Neuron* **83**, 1213–1226 (2014).

[38] Cueva, C. J. *et al.* Low-dimensional dynamics for working memory and time encoding. *Proceedings of the National Academy of Sciences* **117**, 23021–23032 (2020).

[39] Zhang, M. *et al.* Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* **624**, 343–354 (2023).

[40] Markram, H. *et al.* Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience* **5**, 793–807 (2004).

[41] Bugeon, S. *et al.* A transcriptomic axis predicts state modulation of cortical interneurons. *Nature* **607**, 330–338 (2022).

[42] Yu, H. *et al.* In vivo cell-type and brain region classification via multimodal contrastive learning. *bioRxiv* 2024–11 (2024).

[43] Long, B., Yu, C.-P. & Konkle, T. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences* **115**, E9015–E9024 (2018).

[44] Doshi, F. R. & Konkle, T. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances* **9**, eade8187 (2023).

[45] Prince, J. S., Alvarez, G. A. & Konkle, T. Contrastive learning explains the emergence and function of visual category-selective regions. *Science Advances* **10**, eadl1776 (2024).

[46] Chang, L. & Tsao, D. Y. The code for facial identity in the primate brain. *Cell* **169**, 1013–1028 (2017).

[47] Whittington, J. C., Dorrell, W., Ganguli, S. & Behrens, T. E. Disentanglement with biological constraints: A theory of functional cell types. *arXiv preprint arXiv:2210.01768* (2022).

[48] 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. *Nature neuroscience* **22**, 297–306 (2019).

[49] Driscoll, L. N., Shenoy, K. & Sussillo, D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nature Neuroscience* **27**, 1349–1363 (2024).

[50] Dubreuil, A., Valente, A., Beiran, M., Mastrogiuseppe, F. & Ostojic, S. The role of population structure in computations through neural dynamics. *Nature neuroscience* **25**, 783–794 (2022).

[51] Johnston, W. J. & Fusi, S. Modular representations emerge in neural networks trained to perform context-dependent tasks. *bioRxiv* 2024–09 (2024).

[52] Raichle, M. E. & Gusnard, D. A. Appraising the brain's energy budget. *Proceedings of the National Academy of Sciences* **99**, 10237–10239 (2002).

[53] Chen, Y. T. & Witten, D. M. Selective inference for k-means clustering. *Journal of Machine Learning Research* **24**, 1–41 (2023).

[54] Williams, A. H. *et al.* Discovering precise temporal patterns in large-scale neural recordings through robust and interpretable time warping. *Neuron* **105**, 246–259 (2020).

[55] Pillow, J. W. *et al.* Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995–999 (2008).

[56] Izenman, A. J. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis* **5**, 248–264 (1975).

[57] Posani, L. Decodanda: a python package for decoding and geometrical analysis of neural activity. *In preparation. Available on Github: https://www.github.com/lposani/decodanda* (2024).

[58] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

[59] Bron, C. & Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM* **16**, 575–577 (1973).

[60] Langdon, C. & Engel, T. A. Embedding dimension of neural manifolds and the structure of mixed selectivity. *In preparation.* (2024).

[61] Laboratory, I. B. *et al.* Reproducibility of in-vivo electrophysiological measurements in mice. *bioRxiv* 2022–05 (2022).

Methods

Data structure

We used the International Brain Laboratory (IBL) public data release [4]. For each experimental session (178 in total), we collected time-series data on task, behavior, and electrophysiological recordings. These were segmented into trials based on key task events. The task recordings collected for each trial included information on the block prior as well as the stimulus contrast and location. The behavior recordings for each trial comprised the choice made, the outcome/ reward received, and the time-varying movement such as wheel movement velocity, whisker motion energy, and licks. Other behavior movements, such as paw movement, body motion energy, and pupil diameter traces, could be potentially included. However, we did not include them due to missing values in many sessions. The electrophysiological recordings for each trial contained time-varying spike trains of recorded neurons. All these recordings can be accessed directly via the IBL's open API. The following section details the steps for preprocessing these raw data into data matrices for the encoding model.

Criteria for session inclusion

We iterated over all cortical regions and downloaded the related sessions. Sessions were included as long as all the behavior recordings (wheel velocity, whisker motion energy and licks) and electrophysiological data were in place. Some analyses required additional inclusion criteria, such as a minimum number of trials per condition. These analysis-specific criteria are discussed in the relevant sections below.

Criteria for trial inclusion

All trials from the left or right unbalanced blocks were included except when the animals did not respond to the stimulus in time (first movement time was longer than 0.8 s). Trials from the "50-50" balanced block were excluded from the analysis to avoid possible time artifacts arising from the fact that all these trials were exclusively recorded in the first 90 trials of the session.

Criteria for neuron inclusion

All neurons were included in the downloaded data as long as their mean firing rate was larger than 0.5 Hz and smaller than 50 Hz. For the selectivity and geometry analyses, we included only neurons that were predicted above a minimal threshold of min ΔR^2 using the RRR model described below. Unless specified differently, we used min $\Delta R^2 = 0.015$. This threshold was necessary to avoid the confounding effects of neurons that do not encode any relevant variable, including those neurons that were recorded with a low signal-to-noise ratio. But the exact value of the threshold is not crucial for obtaining our results.

Reduced-rank regression encoding model

In this section, we describe the reduced-rank regression (RRR) model used to analyze the selectivity profiles of single neurons. We start by describing the input variables and the target variables of

the model, followed by the description of the model itself and its fitting procedure. Finally, we introduce a few quantities resulting from the fitted model that are key to the follow-up analysis. The notation that will be used are summarized in Table 1. The code for implementing and fitting the encoding model is available at https://github.com/realwsq/brainwide-RRR-encoding-model.

Input and target variables

Target variables

For each trial, we used the spike trains of the time window -0.2 to 0.8 seconds relative to stimulus onset, as the first movement time is typically less than 0.8 seconds [4]. The activity of each neuron was first binned at 0.01 s, divided by the size of the time bin, and then smoothed with a Gaussian filter with a standard deviation of 0.02 s. We tried to apply the linear time warping technique [54] so that the stimulus onset time and first movement or response time aligned across trials, but the results did not differ substantially. The resulting activity of neuron n, denoted as fr_n, was organized into a matrix of shape $K_n \times T$, where K_n is the number of trials and $T = 100$ is the number of time steps per trial. K_n depends on n as neurons may have different numbers of trials if they were recorded in different sessions. Finally, for each neuron and each time step, we Z-scored the activity fr across trials to obtain the target variable y as follows (Suppl. Fig. 1 B-D):

$$y_n(k,t) = \frac{fr_n(k,t) - \mu_n(t)}{\sigma_n(t)}, \text{ where } \mu_n(t) = \frac{\sum_k fr_n(k,t)}{K_n} \text{ and } \sigma_n(t) = \sqrt{\frac{\sum_k (fr_n(k,t) - \mu_n(t))^2}{K_n}}. \tag{3}$$

Since the squared error between the preprocessed data and model predictions was used in the loss function for optimizing the model (Sec 4), the applied normalization prevented biases due to inherent differences in activity scales and ensured that the predictions for all the neurons and time steps were optimized equally. Notably, the Z-score transformation is easily invertible, allowing the model's predictions to be mapped back to the original units of firing rate, thus preserving interpretability.

Input variables

The input variables (x, Suppl. Fig. 1A) we considered can be divided into two types: discrete task-based variables and continuous movement variables. Discrete task-based variables include task-related features, such as the block prior, stimulus contrast, stimulus side, choice, and outcome. These are listed below:

- **Block**: The prior probability for the stimulus to appear on the left side is either $p(\text{left}) = 0.2$ (right block), or $p(\text{left}) = 0.8$ (left block). We used one input variable to encode the block prior: $-1.$ representing $p(\text{left}) = 0.2$, and $+1.$ representing $p(\text{left}) = 0.8$. As noted above, we excluded trials from $p(\text{left}) = 0.5$ unbiased block.
- **Contrast**: The stimulus contrast is 0%, 6.25%, 12.5%, 25%, or 100%. One variable was used to encode the stimulus contrast: 0. representing 0% contrast, 1. representing the low contrast ($\leq 12.5\%$), and 4. representing the high contrast ($> 12.5\%$).
- **Stimulus**: The stimulus location is either on the left side ($+1.$) or the right side ($-1.$).

- **Choice**: The choice is indicated by the turning of the wheel: clockwise (+1.) or counterclockwise (−1.).
- **Outcome**: The outcome is either a water reward (+1.), or negative feedback (−1.).

Since these values are static, all the time points share the same values (e.g., see Suppl. Fig. 1A).

Continuous movement variables included both instructed (e.g., licking and wheel velocity) and uninstructed (e.g., whisker motion energy) movement. These are listed below.

- **Wheel**: The velocity of the wheel movement (radian per second) per time bin.
- **Whisking**: The whisker motion energy per time bin is calculated as the motion energy for a square of the left/ right camera roughly covering the whisker pad. The maximum value between the left and right whisker motion energy was used.
- **Lick**: The number of licks per time bin.

A few preprocessing steps were applied separately to each movement variable: for each trial, we first read out the continuous behavior of the time window −0.2 to 0.8 seconds relative to stimulus onset and interpolated into 0.01 s time bins. Then, to account for the activity that was shifted in time, for each session, we computed the mean time-lagged correlation between the neuronal activity and the movement traces averaged across neurons and trials and shifted the movement traces so that the zero-lagged correlation was maximized. Lastly, significant differences were observed in the variance of input values across trials, both for different input variables and different time steps. To ensure optimal performance and clarity in interpretation, we Z-scored the values for each input variable and time step across trials in the same way as Eq 3. Examples of the resulting input variables are shown in Suppl. Fig. 1A.

The reduced-rank regression model

Linear encoding model

For each neuron n, we describe its temporal responses as a linear, time-dependent combination of input variables (Suppl. Fig. 1-AB):

$$y_n(k,t) \approx \hat{y}_n(k,t) = \sum_v \beta_n^v(t)\, x^v(k,t), \quad \forall k,t,n, \tag{4}$$

where

- $y_n(k,t)$ is the preprocessed neuronal activity of the trial $k \in \{1,\cdots,K_n\}$ and time step $t \in \{1,\cdots,T\}$.
- $\hat{y}_n(k,t)$ is the corresponding model prediction, given by the value of the equation on the right-hand side.
- v represents the relevant input variables included in the model. $x^v(k,t)$ is the preprocessed value of the input variable v for the trial k and time step t.
- $\beta_n^v(t)$ is the effect size of the input variable v at time step t. It is further referred to as the regression *coefficient*.

Low-rank coefficient matrix

The time-varying coefficients $\boldsymbol{\beta}_n^v \in \mathbb{R}^T$ are the weighted sum of a set of temporal basis vectors shared across all the neurons and input variables, that is

$$\boldsymbol{\beta}_n^v = \boldsymbol{U}_n^v \boldsymbol{V}, \quad \forall n, v. \tag{5}$$

Here, $\boldsymbol{U}_n^v \in \mathbb{R}^d$ is the neuron n and input variable v-dependent loading of *temporal basis vectors* $\boldsymbol{V} \in \mathbb{R}^{d \times T}$. Specifically, we considered sharing a single set of temporal basis vectors across all the neurons from all the brain regions and across all the input variables. We verified that this restriction did not compromise the goodness-of-fit. The rank d is generally a value much smaller than the number of time steps T. See Suppl. Fig. 1E for an example decomposition. Sharing the temporal bases across neurons and input variables significantly reduces the number of parameters (Suppl. Fig. 1F). Let N be the number of neurons, T be the number of time steps, and $|v|$ be the number of input variables. An unconstrained full-rank coefficient matrix employs $N \times |v| \times T$ parameters, while a reduced-rank coefficient matrix of the same shape only needs $N \times |v| \times d + d \times T$ parameters. Since N, T are typically much larger than d, the reduction in parameters is on the order of T, i.e., around 100-fold.

Comparison to previous regression models

Well-established linear encoding models include generalized linear model (GLM) [4, 55] and kernel regression model [20]. Below, we first describe these two models and then discuss how they compare to our RRR model. GLM is expressed as

$$y_n(k,t) \approx \hat{y}_n(k,t) = \sum_v \sum_{t'} \beta_n^v(t')\, x^v(k, t-t'), \quad \forall k, t, n, \tag{6}$$

where the input filter vector, $\boldsymbol{\beta}_n^v$, composed of $\beta_n^v(t')$ over a neighboring time window, is factorized as

$$\boldsymbol{\beta}_n^v = \boldsymbol{U}_n^v \boldsymbol{\kappa}^v, \quad \forall n, v. \tag{7}$$

Here, κ^v represents pre-specified temporal basis vectors, which are not trained. Adapted from [4] [1], we used the same set of raised cosine "bump" functions in log space for Suppl. Fig. 3B. The kernel regression model employs the same linear formulation as the GLM (Eq 6). However, instead of relying on pre-defined temporal basis vectors, it allows these vectors to be trainable:

$$\boldsymbol{\beta}_n^v = \boldsymbol{U}_n^v \boldsymbol{V}^v, \quad \forall n, v. \tag{8}$$

Both the kernel regression model and our RRR model fall into the category of reduced-rank regression models. The primary distinction lies in the set of input features used by each approach.

In our RRR model, the influence of input variables on neural responses, $\beta_n^v(t)$ can vary over time. In contrast, both the GLM and the kernel regression model assume that the influence is time-independent. Therefore, for example, whether the subject response time is early or late makes no

[1] Here we considered neural responses over a different time window from [4] and enriched set of behavior movements.

difference and will modulate the neural responses the same way. Instead, these models provide a more descriptive account of input effects by allowing neural responses to depend on inputs from neighboring time steps. While this feature is absent in our current RRR model (Eq. 4), it could be incorporated in future extensions.

Estimation of the parameters

The parameters of the RRR model include a shared temporal bases matrix $\boldsymbol{V}$ of size $d \times T$ and loading vectors $\boldsymbol{U}_n^v$ of length d for each input variable v and neuron n. The approach we adapted to fit the parameters was to minimize the ridge-penalized mean square loss:

$$\mathcal{L}\left(\boldsymbol{V}, \{\boldsymbol{U}_n^v\}_{n,v}\right) = \sum_n \left(\sum_k \sum_t \left(y_n\left(k,t\right) - \hat{y}_n\left(k,t\right)\right)^2 + \lambda \sum_v \sum_t \beta_n^v(t)^2 \right). \tag{9}$$

Minimizing this particular loss function is straightforward as a closed-form solution exists [56]. In practice, we chose to use the L-BFGS optimization algorithm to compute the optimum.

Moreover, to optimize the model hyperparameters, namely the rank d and the regularization penalty λ, we implemented a 3-fold cross-validation technique across trials. First, the dataset was stratified based on a composite target label that included the block prior and stimulus contrast to ensure each fold was representative of the entire dataset. Then, for each combination of d and λ, the dataset was partitioned into three subsets by trials, utilizing each subset in turn for testing the model while the remaining data served as the training set. Finally, the combination of d and λ that yielded the lowest average test error across all folds was selected. $d = 5$ turned out to be the optimal number of temporal bases (Suppl. Fig. 1E).

Estimating the goodness of fit

We used the 3-fold cross-validated R^2 to measure the goodness-of-fit of single-trial predictions. For each session's data, we randomly sampled one-third of the trials as the test set held out during training. Once the model is trained - using the remaining two-thirds of trials - we computed the R^2 between the model predictions $\hat{fr}_n(k,t)$ [2] and the actual neuronal activity $fr_n(k,t)$ on the test set. We repeated the whole split-train-test process three times and computed the mean of the three cross-validated R^2 as the measure of goodness-of-fit.

Selectively modulated neurons

Conceptually, we distinguish two types of task modulation: the *selective modulation* and the *non-selective modulation*. Selective modulation, captured by $\hat{y}_n(k,t) = \sum_v \beta_n^v\left(t\right) x^v(k,t)$ is induced by the input variables and varied trial-by-trial. Non-selective modulation, captured by the mean time-varying response $\mu_n(t) = \frac{\sum_k fr_n(k,t)}{K_n}$ is locked to the key events of the trial (stimulus onset in this case) and does not vary trial-by-trial. Both types play a significant role in modulating neuronal responses. See Suppl. Fig. 3 for example selectively modulated and non-selectively modulated neurons. In this work, we focus mostly on the neuronal responses selectively modulated by the

[2] $\hat{fr}_n(k,t)$ is calculated by inversing the z-score transformation applied in the preprocessing step $\hat{fr}_n(k,t) = \sigma_n(t)\hat{y}_n(k,t) + \mu_n\left(t\right)$ (Suppl. Fig. 1-BCD).

task. To distinguish the variation explained by the selective modulation from the non-selective modulation, we use the trial-average estimate as the null model that does not consider any effect of the input variables:

$$\hat{y}_n^{\text{null}}(k,t) = 0,\ \forall k,t,n. \tag{10}$$

The outperformance, ΔR^2, defined as:

$$\Delta R^2(\text{model}) = R^2(\text{model}) - R^2(\text{null}), \tag{11}$$

captures the overall selective modulation of all the input variables combined.

Only selectively modulated neurons, identified as $\Delta R^2(\text{RRR}) \geq 0.015$ (Suppl. Fig. 3-A), are included in the further analysis.

Computing the selectivity profiles of single neurons to the input variable

To compute the selectivity profiles of single neurons to the individual variables, we used the estimated coefficient $\beta_n^v(t)$. For the clustering analysis of Fig. 3, Suppl. Fig. 7), we took the sum of the coefficients across time as a measure of the total selectivity of neuron n to input variable v.

$$\alpha_n^v = \sum_t \beta_n^v(t) \tag{12}$$

Note that by normalizing the neuronal responses and input variables in the preprocessing steps, we ensured that the unit-free coefficients $\beta_n^v(t)$ are not affected by the neuron's mean firing rate or the inherently different scales in different input variables and can be compared directly across neurons, input variables, and time steps. Thus, $\beta_n^v(t)$ can be interpreted as the expected change in normalized neuronal activity y_n per one standard deviation change in the input variable x^v at time t, and α_n^v can hence be thought as the expected total change across the whole trial. Also, note that the sign of $\beta_n^v(t)$ typically does not change over time t. See Suppl. Fig. 1-A for examples.

The selectivity α_n^v captures whether individual neurons are selectively modulated by the given variable v or not (see Suppl. Fig. 2 for examples of responses from strongly selective neurons). In the selectivity analysis (Fig. 2ef), when the goal is to estimate the absolute modulation of an input variable, α_n^v is calculated as $\alpha_n^v = \sum_t |\beta_n^v(t)|$.

Computing the autocorrelation timescale of neural responses to task variables

Given a matrix of single-neuron responses to task variables $\hat{\boldsymbol{y}}_n \in \mathbb{R}^{K\times T}$ with K being the number of trials and T the number of time steps per trial, we can compute the corresponding autocorrelation timescale. To compute the timescale, we first calculate the time-lagged autocorrelation sequence

$$c_n(i) = \frac{\sum_k \sum_t \hat{y}_n(k,t)\hat{y}_n(k,t+i)}{K},\ i \geq 0.$$

We then linearly interpolate the autocorrelation sequence so that $c_n(i)$ is spaced at 1 ms resolution (original 10 ms). The timescale τ_n is approximated by the time the autocorrelation function first

reaches half its peak value. The timescale of brain area a (Fig. 2d) is further determined by averaging the values over all the selectively modulated neurons within this area.

Table 1 Notation

p0.1X
Indices:
n, N index and number of neurons ($n \in \{1, \cdots, N\}$)
k, K index and number of trials ($k \in \{1, \cdots, K\}$)
t, T index and number of time steps ($t \in \{1, \cdots, T\}$)
v, $\|v\|$ index and number of input variables ($v \in$ {block prior, stimulus contrast, stimulus side, choice, outcome, wheel velocity, whisker motion energy, lick})
i, d index and rank of the RRR model ($i \in \{1, \cdots, d\}$)
RRR model:
$\beta_n^v(t)$ regression coefficient (a.u.) of neuron n, input variable v at time step t
$\boldsymbol{U}_n^v \in \mathbb{R}^d$ neuron n and input variable v-dependent loading of temporal basis vectors (a.u.)
$\boldsymbol{V} \in \mathbb{R}^{d \times T}$ temporal basis vectors (a.u.)
Random variables:
$x^v(k,t)$ preprocessed value (a.u.) of input variable v, trial k at time step t
$y_n(k,t)$ preprocessed neuronal response (a.u.) of neuron n, trial k at time step t
$\hat{y}_n(k,t)$ model prediction of preprocessed neuronal response (a.u.) of neuron n, trial k at time step t
$fr_n(k,t)$ smoothed, binned firing rate (Hz) of neuron n, trial k at time step t
$\hat{fr}_n(k,t)$ model prediction of smoothed, binned firing rate (Hz) of neuron n, trial k at time step t
$\mu_n(t)$ mean of smoothed, binned firing rate (Hz) of neuron n at time step t
$\sigma_n(t)$ standard deviation of smoothed, binned firing rate (Hz) of neuron n at time step t

Clustering Analysis

To test for the presence of functional clusters, we followed the steps explained below. The required inputs include each relevant neuron's response profile and original session ID. Two types of response profiles can be considered: the estimated selectivity to individual input variables (Eq 12, Fig. 3, Suppl. Fig. 7, referred to as clustering analysis in the variable selectivity space) or the average response in each experimental condition (Suppl. Fig. 7, referred to as clustering analysis in the conditions space). Performing clustering analysis in the selectivity space arguably has a few advantages:

- It reduces the dimensionality in an interpretable and informed way. If we have $|v|$ variables, then there are at least $2^{|v|}$ conditions, assuming all the variables are discrete and have more than one different value.

- It mitigates the issue of unbalanced or even missing conditions.
- It reduces the noise in the estimation of the response profile. As shown in Suppl. Figure 3, neural responses are very noisy, and simple averaging may be non-satisfactory (Figure 2b). The selectivity estimated from the encoding model, in comparison, provides a more reliable account of the task-driven variance in the neural responses.

The code for the clustering analysis is available at https://github.com/realwsq/clustering-analysis.

Clustering analysis in the variable selectivity space.

We summarize our clustering pipeline below (see also Fig. 3a for schematic illustration).

1. Check whether there are more than 50 neurons and only continue if so.
2. Given the selectivity profile of each neuron, run k-means clustering algorithm (with 100 random initializations) with the number of clusters k varied from 3 to 20.
3. Then, select the optimal clustering result by maximizing the silhouette score. The Silhouette score is defined as $\left\langle \frac{b_i - a_i}{\max(b_i, a_i)} \right\rangle_i$ where i is the index of the neuron, $a_i = \frac{1}{|C_I|-1} \sum_{j \in C_I, i \neq j} d(i,j)$ is the mean Euclidean distance intra-cluster and $b_i = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$ is the min Euclidean distance outside cluster (Fig. 3a).
4. Iterate over the resulting clusters and check whether there is a cluster whose total silhouette score summed over all neurons is mainly contributed by neurons from one single session ($> 90\%$). If so, remove neurons from that cluster and session, and repeat steps 1-4.
5. Sample the same number of data points from the Gaussian distribution with the mean and covariance matrix matched to the data values and compute the sampled data's null silhouette score following steps 2-3.
6. Repeat step 5 100 times and pool the null silhouette scores to form the null distribution. Finally, compute the z-score of the data silhouette score with respect to the null distribution.

Clustering analysis in the conditions space.

When clustering in the space of mean firing rate, two additional preprocessing steps are required. First, we normalized each neuron's mean firing rates separately across conditions to prevent clustering driven solely by overall firing rate differences between neurons. Second, since the number of conditions and the dimensionality of the activity profile is high, we reduced the dimensionality using principal component analysis. Additionally, we modified the null model by replacing the multivariate Gaussian distribution with a multivariate log-normal distribution to better capture the lower-bounded and heavy-tailed nature of mean firing rates. Neurons with zero activity were excluded from this analysis.

The clustering analysis in the conditions space follows these steps (see also Suppl. Fig. 6 for schematic illustration):

1. Check whether there are more than 40 neurons and only continue if so.
2. Z-score the mean firing rates for each neuron.
3. Reduce dimensionality using PCA, retaining components that capture 90% of the total variance.

4. Run k-means clustering algorithm (with 100 random initialization), varying the number of clusters k from 3 to 20.
5. Select the optimal clustering result by maximizing the silhouette score.
6. Iterate over the resulting clusters and check whether there is a cluster whose total silhouette score summed over all neurons is mainly contributed by neurons from one single session ($> 90\%$). If so, remove neurons from that cluster and session, and repeat steps 1-5.
7. Sample the same number of data points from the multivariate log-normal distribution with the mean and covariance matrix matched to the log data values, then compute the sampled data's null silhouette score following steps 2-5.
8. Repeat step 7 100 times and pool the null silhouette scores to form the null distribution. Finally, compute the z-score of the data silhouette score with respect to the null distribution.

Modified ePAIRS test

Given the average responses of single neurons in each experimental condition, we performed the ePAIRS test as follows (see also Suppl. Fig. 6 for schematic illustration):

1. Z-score the mean firing rates for each neuron.
2. Reduce dimensionality using PCA, retaining components that capture 90% of the total variance.
3. Calculate the cosine distance to its nearest neighbor for each neuron.
4. Calculate the empirical median of these nearest-neighbor distances as the aggregated nearest-neighbor angle.
5. Sample the same number of data points from the multivariate log-normal distribution with the mean and covariance matrix matched to the log data values, then compute the null nearest-neighbor angle of the sampled data following steps 1-4.
6. Repeat step 5 5000 times and pool the null nearest-neighbor angles to form the null distribution. Finally, the z-score of the data aggregated nearest-neighbor angle with respect to the null distribution is computed.

Analysis of population representations

Data Preparation

For the analysis of population neural representations, we used the same sessions as described above. For each trial within a session, we thus have a collection of N-dimensional population activity vectors $\mathbf{f}^{t,k}$, where $k \in \{1, P\}$ is the trial index within the session, and $t \in \{1, T\}$ is the time-bin index within each trial. For the analysis below, we used data from 0 to 1000ms after the stimulus onset, so to capture a variety of sensory and behavioral variables. We then labeled each time bin according to the value of four binarized cognitive, sensory, and movement variables:

- **Block**: left (20-80) vs. right (80-20) prior block.
- **Contrast**: we binarized the contrast into low (0-0.125) vs. high (0.25-1.0) values.
- **Stimulus**: left vs. right side of the screen.

- **Whisking**: we binarized the whisking power using the distribution of whisking power values within each session. Time bins where the mouse was whisking with a power larger than the 50 percentile across the distribution were annotated as *high*, while those below the 50 percentile were annotated as *low*.

These variables were chosen so that they span movement, cognitive, and sensory variables while ensuring that all the $M = 16$ conditions (combinations of the four variables) were well represented in the data. For example, we could not add Choice as a variable since mice are overtrained in the task and, as a consequence, make very few mistakes when Block and Stimulus are aligned (e.g., choose "left" when the block and the stimulus are both "right".)

For each condition $c \in \{1, M\}$ (for example, Whisking = high, Contrast = low, Stimulus = left, Block = right), we first identified those trials where that specific combination of variables was present. We then defined a collection of "conditioned trial" population activity vectors $\{\mathbf{f}^{k,c}\}$ as the mean firing rate of the population of neurons conditioned to the specific condition in each trial. Given a trial k and a neuron index i, the mean firing rate was computed as

$$f_i^{k,c} = \frac{\sum_t \delta^{t,k}(c)\ f_i^{t,k}}{\sum_t \delta^{t,k}(c)} \quad , \tag{13}$$

where i indicates the neuron index and $\delta^{t,k}(c) = 1$ if the time bin t in trial k corresponds to the condition c, and $= 0$ otherwise. These conditioned trial population vectors are the data samples that will be used for the dimensionality and decoding analyses below. Across all analyses, we considered only those recording sessions where each condition was present in at least $M_{min} = 5$ trials.

Participation Ratio

The Participation Ratio (PR) quantifies the effective dimensionality of a set of data points by measuring how evenly the variance is distributed across the eigenvalues of its PCA decomposition [29]. For each neural population, we computed the PR of the *centroids* $\mathbf{f}_c$ of the M conditions, defined as the average activity pattern across all trials of the same condition:

$$\mathbf{f}^c = \left\langle \mathbf{f}^{k,c} \right\rangle_k \quad , \tag{14}$$

To compute the PR for the set of centroids, we first calculated their covariance matrix, normalizing each neuron's mean activity vector by subtracting from each f_i^c the mean across conditions c and dividing them for their standard deviation. We then performed Principal Component Analysis (PCA) on this covariance matrix to obtain its eigenvalues, λ_j. The PR is defined as:

$$\mathrm{PR} = \frac{\left(\sum_j \lambda_j\right)^2}{\sum_j \lambda_j^2} \quad , \tag{15}$$

where λ_j is the j-th eigenvalue of the covariance matrix. A higher PR indicates that the variance is more evenly spread across multiple dimensions, suggesting a higher effective dimensionality of the neural representations. Conversely, a lower PR implies that the variance is concentrated in

fewer dimensions, indicating a lower effective dimensionality. This metric provides insight into the complexity and dimensionality of the neural population's response patterns.

Cross-validated Decoding

We used Decodanda [57] (www.github.com/lposani/decodanda) to perform a cross-validated, class-balanced decoding analysis of different combination of condition labels from the neural activity within individual trials (condition trial vectors $\mathbf{f}_c^k$). See individual sections below for additional detail on the data input structure of our decoding analyses. As a decoder, we used a scikit-learn SVM classifier with linear kernel [58]. To ensure that results were comparable across regions, which might have a different number of recorded neurons, we created a pseudo-population by over-sampling (or down-sampling) all the recorded neurons within each region to a fixed number $N = 640$. Similarly, we re-sampled the same number of pseudo-population for each analysis ($T = 100$ patterns per condition). Note that simultaneously recorded neurons were always kept together during resampling, so as to keep the noise correlations intact within the pseudo population [57]. All cross-validated decoding analyses were performed using the following Decodanda parameters: `training_fraction` $= 0.8$, `cross_validations` $= 100$, `ndata` $= 100$. These analyses were used for Fig. 5, Fig. 6, Suppl. Figures 9, 8, 10.

Finding the Independent Conditions

To find the number of independent conditions encoded in the activity of a population of neurons, we developed an iterative algorithm based on linear decoding. The algorithm followed the steps below, and is shown in action on one example region in Suppl. Fig. 8.

1. First, we decoded individual condition trial population vectors $\mathbf{f}^{k,c}$ based on their condition label c using a linear classifier (as explained in the section above). For each pair of conditions (c_i, c_j), we estimated a cross-validated decoding performance $\varphi(c_i, c_j)$, resulting in an initial $M \times M$ condition-condition decoding matrix (C_0, see Suppl. Fig. 8) defined as $\mathrm{C}_0(ij) = \varphi(c_i, c_j)$.
2. We then chose a decoding threshold $\varphi_{\min} = 0.666$; the pairs of conditions whose 1-vs-1 decoding performance was smaller than $\varphi_{\min}$ were defined as "dependent". Using this threshold, we defined a binary *dependency* matrix D defined as $\mathrm{D}_0(i, j) = 1$ if $\varphi(c_i, c_j) < \varphi_{\min}$, and $\mathrm{C}_0(i, j) = 0$ otherwise.
3. Then, we used the Bron-Kerbosch algorithm [59] to find all the cliques, i.e., subgroups of fully-connected nodes, in the undirected graph defined by the dependency matrix D_0. This process allows us to identify whether there are groups of conditions that are all non-decodable from each other (dark squares in the sorted matrix in Suppl. Fig. 8).
4. Next, we identified the largest clique and grouped together all the trials of the conditions within that group into a new, *merged* condition (see arrows and "mrg" conditions in Suppl. Fig. 8).
5. We repeated steps 1 and 2 with the new reduced set of conditions, yielding a new C_t and a new D_t matrix of a different size M_t; t denotes the iteration step.
6. We then repeated step 3 and 4, and iterated the whole process (1-4) until all the merged and remaining conditions were found to be independent, i.e, the dependency matrix $\mathrm{D}_{\tilde{t}}$ was diagonal.

The number of independent conditions was then defined as the size of the final dependency matrix: $M_{IC} := M_{\tilde{t}}$.

Shattering Dimensionality

Shattering Dimensionality (SD) [6, 14] is a functional definition of dimensionality that quantifies how many random classifications of a given set of points a linear readout can solve. To estimate the SD of a neural population, we followed the following steps:

1. First, we randomly divided the set of M_{IC} independent conditions into two equally-sized groups (dichotomy).
2. Given the dichotomy d, we then measured the cross-validated decoding performance φ_d of a linear classifier trained to report whether individual condition trial vectors $\mathbf{f}^{k,c}$ belonged to conditions within one or the other dichotomy groups. This decoding analysis was performed as described in the Decoding section above, resampling a fixed number of neurons and a fixed number of trials per condition for all regions to ensure the comparability of the performances across regions.
3. The random dichotomy assignment and decoding (steps 1, 2) was then repeated $n = 200$ times to obtain a population of decoding performances $\{\varphi_d\}$.
4. the decoding analysis was then repeated for $n = 200$ iterations with shuffled condition labels across the population vectors to obtain a distribution of null decoding performance values $\{\varphi^{\text{null}}\}$
5. The SD was then defined as the fraction of decodable dichotomies, i.e., the fraction of φ_d larger than the 95 percentile of the null population $\{\varphi^{\text{null}}\}$.

The distributions of decoding performances for all the analyzed regions are shown in Suppl. Fig. 10.

Theoretical derivation of the relationship between embedding dimensionality and clustering

We consider a data model with N features (neurons) and M observations (conditions), in which observations are sampled i.i.d. as

$$\boldsymbol{x}^{\mu} = \boldsymbol{z}^{\mu} + \boldsymbol{\eta}^{\mu} \quad , \tag{16}$$

where $\boldsymbol{z}^{\mu}$ and $\boldsymbol{\eta}^{\mu}$ are both vectors in $\mathbb{R}^N$ and represent the clustered and heterogeneous part of the data, respectively. More precisely, $\boldsymbol{z}^{\mu}$ is sampled from a normal distribution $\mathcal{N}(0, \mathbf{B})$ that has a clustered covariance matrix, i.e. $B_{ij} = 1$ if i and j belong to the same cluster and $B_{ij} = 0$ otherwise. We call k the number of clusters and assume that all clusters have the same number of neurons $N_c = N/k$. In contrast, the heterogenous part $\boldsymbol{\eta}^{\mu}$ is sampled from $\mathcal{N}\left(0, \sigma^2\mathbf{I}\right)$, where $\mathbf{I}$ is the identity matrix. Our goal is to compute the participation ratio (PR) of this representation, which we define as

$$\text{PR} = \frac{\text{Tr}(\mathbf{C})^2}{\text{Tr}\mathbf{C}^2} = \frac{N\langle C_{ii}\rangle^2}{\langle C_{ii}^2\rangle + (N-1)\langle C_{ij}^2\rangle} \quad , \tag{17}$$

where the averages are across neurons and the matrix $\mathbf{C}$ is the *sample* neuron-by-neuron covariance matrix, i.e. $\mathbf{C} = \frac{1}{M}\sum_{\mu=1}^{M} \boldsymbol{x}^{\mu}(\boldsymbol{x}^{\mu})^T$. We note that this definition assumes that the sample mean of both $\boldsymbol{z}$ and $\boldsymbol{\eta}$ are negligible or have been subtracted.

We are interested in the regime in which $N \to \infty$ while M is allowed to be small, as it often happens in controlled experiments. Small M might cause the sample covariance matrix to differ substantially from the true covariance matrix. Defining $\mathbf{C}^c$, $\mathbf{C}^h$, and $\mathbf{C}^{ch}$ as the sample covariance matrices of $\boldsymbol{z}$, $\boldsymbol{\eta}$, and the cross-covariance between $\boldsymbol{z}$ and $\boldsymbol{\eta}$ respectively, we have that

$$\mathrm{PR} = N \frac{\langle C^h_{ii} + C^c_{ii} + 2C^{ch}_{ii} \rangle^2}{\langle (C^h_{ii} + C^c_{ii} + 2C^{ch}_{ii})^2 \rangle + (N-1)\langle (C^h_{ij} + C^c_{ij} + 2C^{ch}_{ij})^2 \rangle} \quad . \tag{18}$$

Therefore, we need to evaluate the first and second moments of both diagonal and off-diagonal elements of all covariance and cross-covariance matrices. The diagonal elements of these matrices have the following statistics:

$$\begin{aligned}
C^c_{ii} &= \frac{1}{M}\sum_{\mu=1}^{M}(z^\mu_i)^2 \quad \Rightarrow \quad \langle C^c_{ii} \rangle = 1\,, \quad \langle (C^c_{ii})^2 \rangle = \frac{M+2}{M} \\
C^h_{ii} &= \frac{1}{M}\sum_{\mu=1}^{M}(\eta^\mu_i)^2 \quad \Rightarrow \quad \langle C^h_{ii} \rangle = \sigma^2\,, \quad \langle (C^h_{ii})^2 \rangle = \frac{M+2}{M}\sigma^4 \\
C^{ch}_{ii} &= \frac{1}{M}\sum_{\mu=1}^{M} z^\mu_i \eta^\mu_i \quad \Rightarrow \quad \langle C^{ch}_{ii} \rangle = 0\,, \quad \langle (C^{ch}_{ii})^2 \rangle = \frac{1}{M}\sigma^2 \quad ,
\end{aligned} \tag{19}$$

and

$$\langle C^c_{ii} C^h_{ii} \rangle = \sigma^2, \quad \langle C^c_{ii} C^{ch}_{ii} \rangle = 0, \quad \langle C^h_{ii} C^{ch}_{ii} \rangle = 0 \quad . \tag{20}$$

For the off-diagonal elements, we have

$$\begin{aligned}
C^c_{ij} &= \frac{1}{M}\sum_{\mu=1}^{M} z^\mu_i z^\mu_j \quad \Rightarrow \quad \langle C^c_{ij} \rangle = \frac{1}{k}\,, \quad \langle (C^c_{ij})^2 \rangle = \frac{1}{M} + \frac{1}{kM} + \frac{1}{k} \\
C^h_{ij} &= \frac{1}{M}\sum_{\mu=1}^{M} \eta^\mu_i \eta^\mu_j \quad \Rightarrow \quad \langle C^h_{ij} \rangle = 0\,, \quad \langle (C^h_{ij})^2 \rangle = \frac{1}{M}\sigma^4 \\
C^{ch}_{ij} &= \frac{1}{M}\sum_{\mu=1}^{M} z^\mu_i \eta^\mu_j \quad \Rightarrow \quad \langle C^{ch}_{ij} \rangle = 0\,, \quad \langle (C^{ch}_{ij})^2 \rangle = \frac{1}{M}\sigma^2 \quad ,
\end{aligned} \tag{21}$$

and

$$\langle C^c_{ij} C^h_{ij} \rangle = 0, \quad \langle C^c_{ij} C^{ch}_{ij} \rangle = 0, \quad \langle C^h_{ij} C^{ch}_{ij} \rangle = 0 \quad . \tag{22}$$

Most of the expressions above can be straightforwardly derived by writing down the definition of the sample covariance matrix for a zero-mean variable and then performing the average over neurons. To illustrate this procedure, let us consider one of the most involved terms:

$$\begin{aligned}
\langle (C^c_{ij})^2 \rangle &= \frac{1}{M^2}\sum_{\mu,\nu=1}^{M} \langle z^\mu_i z^\mu_j z^\nu_i z^\nu_j \rangle \\
&= \frac{1}{M^2}\sum_{\mu=1}^{M} \langle (z^\mu_i)^2 (z^\mu_j)^2 \rangle + \frac{1}{M^2}\langle z^\mu_i z^\mu_j \rangle \langle z^\nu_i z^\nu_j \rangle
\end{aligned} \tag{23}$$

The probability that z_i and z_j are part of the same cluster is given, for large N, by $\frac{1}{k}$. The expression for $\langle (C^c_{ij})^2 \rangle$ then becomes

$$\begin{aligned}\langle (C^c_{ij})^2 \rangle &= \frac{1}{kM^2}\sum_{\mu=1}^{M}\langle (z_i^\mu)^4 \rangle + \frac{1}{M^2}\left(1-\frac{1}{k}\right)\sum_{\mu=1}^{M}\langle (z_i^\mu)^2 \rangle^2 + \frac{1}{kM^2}\sum_{\mu\neq\nu}^{M}\langle (z_i^\mu)^2 \rangle^2 \\ &= \frac{3}{kM} + \frac{1}{M} - \frac{1}{kM} + \frac{1}{kM}(M-1) \\ &= \frac{1}{M} + \frac{1}{kM} + \frac{1}{k} \quad .\end{aligned} \tag{24}$$

The other terms can be computed following the same steps.

Given that k, M are finite, we can approximate the PR for large N as:

$$\mathrm{PR} \simeq \frac{\langle C^h_{ii} + C^c_{ii} + 2C^{ch}_{ii} \rangle^2}{\langle (C^h_{ij} + C^c_{ij} + 2C^{ch}_{ij})^2 \rangle} \quad . \tag{25}$$

Expanding the square and using the results above for the first and second moments of the covariance matrices, we get to our final expression:

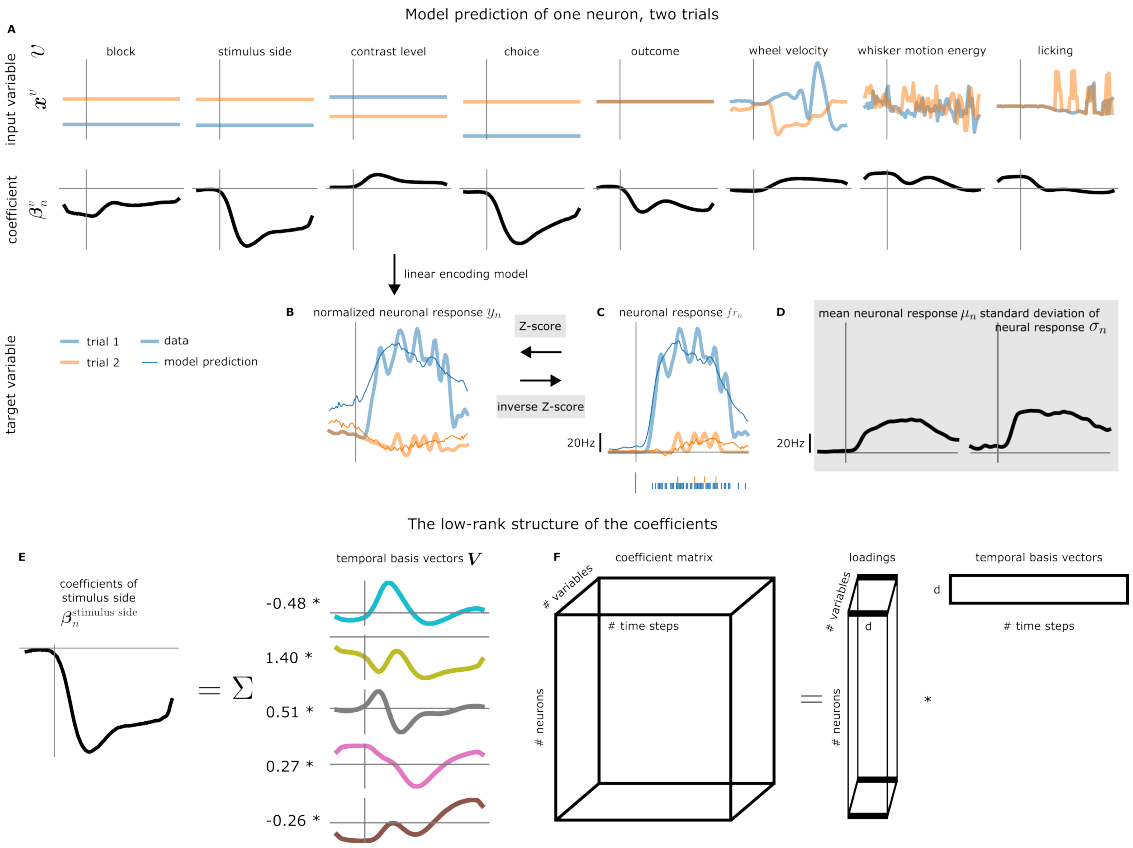
$$\mathrm{PR} \;\simeq\; \frac{\left(1+\sigma^2\right)^2}{\frac{1}{k}+\frac{1}{kM}+\frac{1}{M}\left(1+\sigma^2\right)^2} \;=\; M\frac{k\left(1+\sigma^2\right)^2}{1+M+k\left(1+\sigma^2\right)^2} \tag{26}$$

From the mathematical expression in Eq. 1, we can see that, in the limit of perfect clusters ($\sigma \to 0$), the function is either limited by the number of rows-neurons (in this case, the k perfect clusters) or columns-conditions M, coherently with the intuition above:

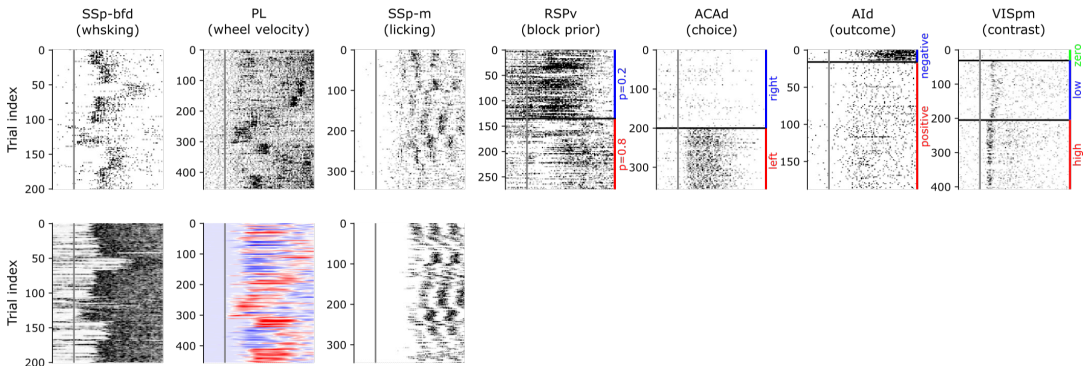
$$\mathrm{PR} \xrightarrow[k\to\infty]{} M \qquad\qquad \mathrm{PR} \xrightarrow[M\to\infty]{} k \quad , \tag{27}$$

However, things become more nuanced when k, M, and σ are finite and non-zero. First, as shown in Fig. 4e, when k and M are kept fixed, the PCA dimensionality decreases with the clustering quality (expressed as the average silhouette score of a population of Gaussian clusters with given k, M, and σ). Second, if we fix the quality of clusters and the number of conditions (Fig. 4e, left), we see that dimensionality increases with the number of clusters, with a magnitude that is larger for high silhouette scores (categorical representations). This behavior is complementary to a recent work that showed the inverse relation, i.e., that constraining the dynamics generated by recurrent neural networks to a low dimensional manifold implies a small number of functional types (categorical clusters) [60]. Finally, when fixing the number and quality of clusters, the dimensionality is determined by the number of conditions, with a magnitude that is larger for low silhouette scores (non-categorical representations, Fig. 4e, right).

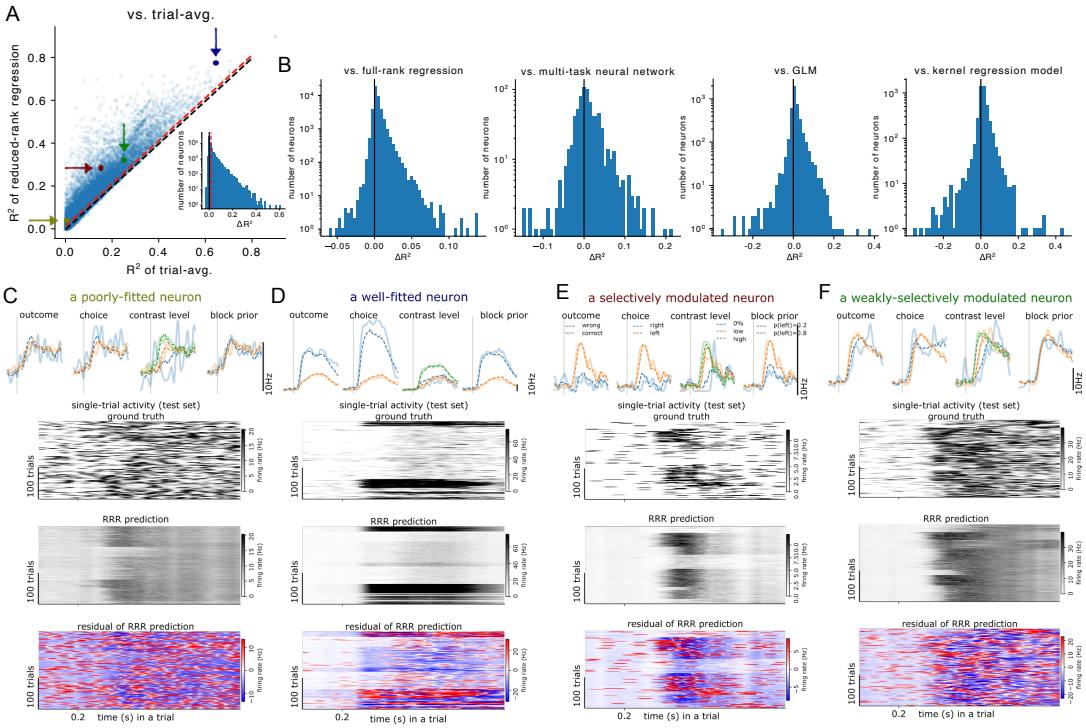
Extended Data



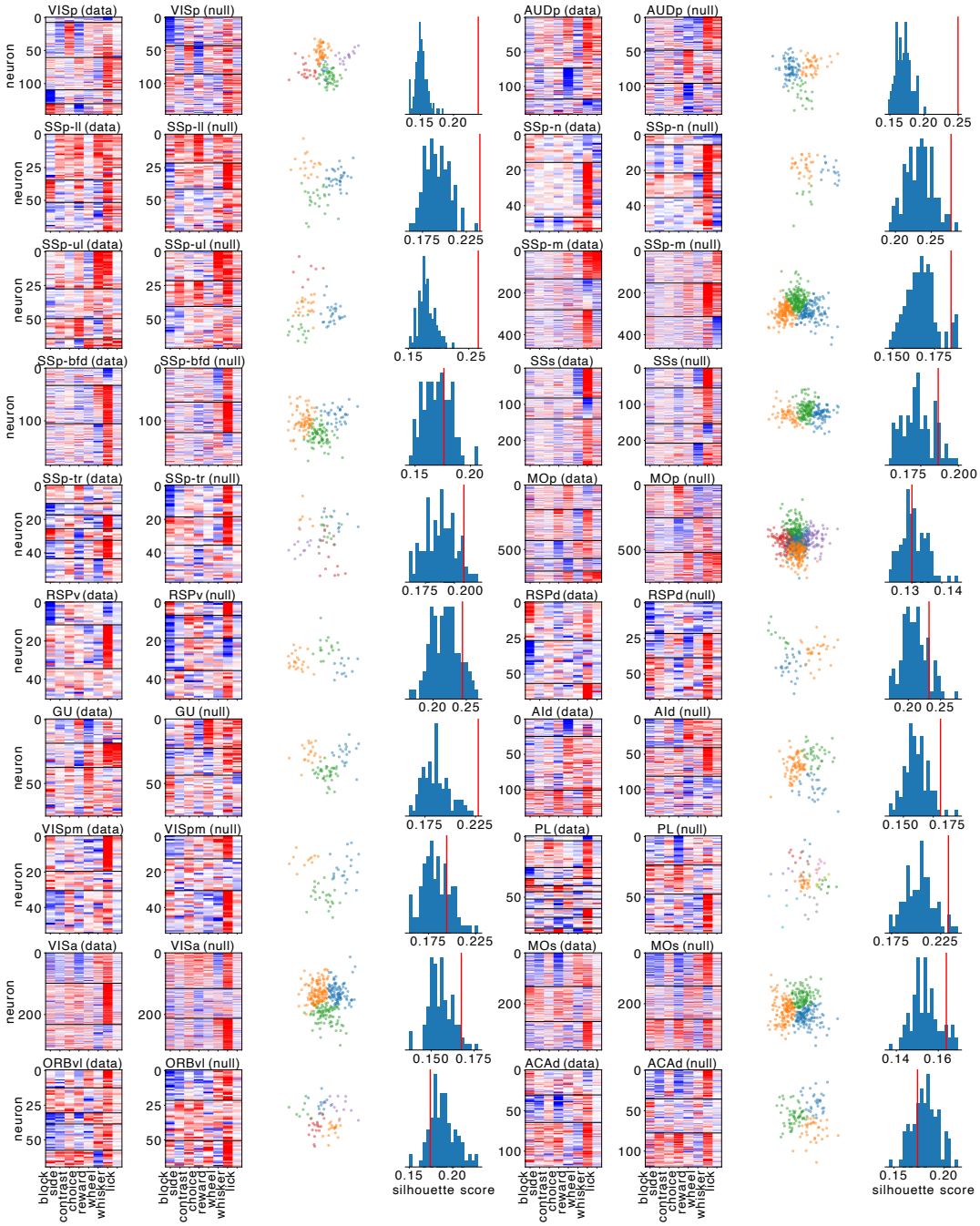
Supplementary Figure 1 Visualization of the RRR encoding model. A-D) Illustration of how the RRR encoding model predicts the neuronal responses using two example trials (blue and orange lines). Importantly, the model coefficients are time-dependent (A), and the inputs (x) and outputs (y) of the RRR encoding model are normalized across trials (B-D, Methods). The prediction (thin and opaque lines) overlays the data (thick and transparent lines). The gray vertical lines indicate the onset of the stimulus (0.2 s), and the gray horizontal lines indicate the zero value. E-F) Coefficients are enforced to be weighted sums of a small set of temporal basis vectors. E) Example decomposition of the stimulus side coefficients in A). F) Schematic plot showing the low-rank structure of the coefficient matrix.



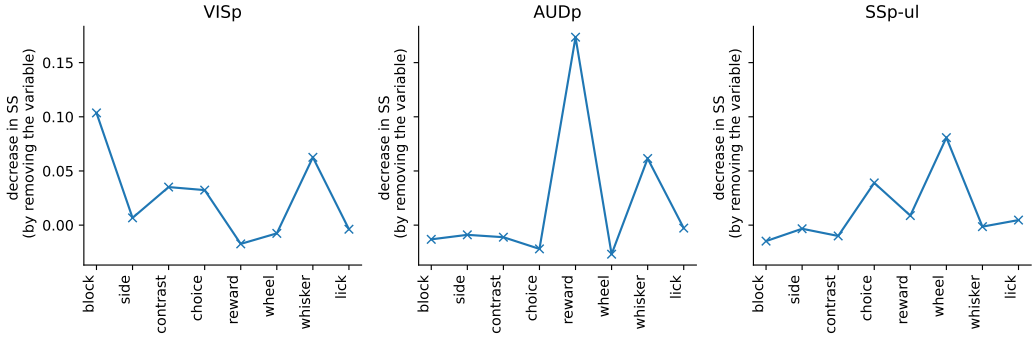
Supplementary Figure 2 Neurons with high selectivity are strongly modulated by the corresponding input variable. Examples of neurons with strong selectivity to each input variable are displayed: the first row shows neuronal activity, while the second row illustrates the associated behavioral movements.



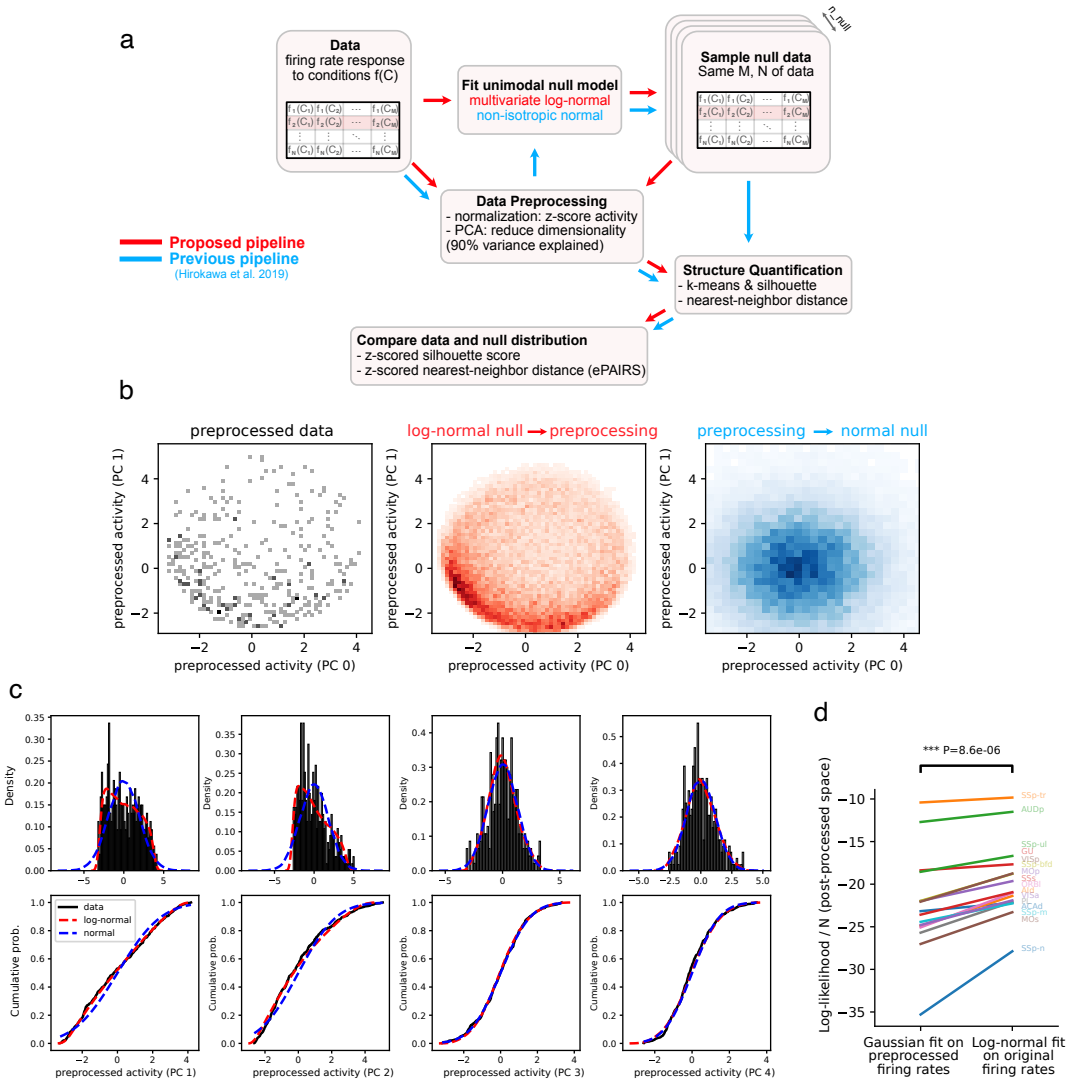
Supplementary Figure 3 The reduced-rank regression (RRR) encoding model effectively reduces trial-to-trial variability and captures task-related variance. A) Goodness-of-fit analysis comparing the RRR encoding model to the trial-averaged neural activity across all trials (i.e., the null model). Each point represents a neuron. The black line marks equal performance for both methods, while the red line highlights cases where RRR outperforms the trial-average estimate (threshold: 0.015). The four colored dots indicate the neurons shown in C-F. B) Comparison of RRR performance relative to other baseline models, measured as $\Delta R^2 = R^2(RRR) - R^2(\text{baseline method})$. Baseline models include full-rank regression (which assumes a linear encoding model without rank constraints), a multi-task neural network (a nonlinear model with feed-forward and recurrent layers as in [61]), a generalized linear model (GLM, adapted from [4]), and a kernel regression model (adapted from [20]). C-F) Example neurons illustrating different response characteristics and how well RRR captures their activity. Top: Peri-stimulus time histograms (PSTHs) for different task conditions (color-coded). Middle: Single-trial activity (test set) alongside RRR predictions. Bottom: Residuals of the RRR predictions. C) A neuron with low R^2, where a large portion of neural variability is unexplained by RRR and not clearly linked to task variables. D) A neuron with high R^2, where RRR successfully captures most of the task-related variance. E) A selectively modulated neuron, showing strong dependence on task conditions, reflected in both the PSTH and single-trial activity. RRR effectively models this neuron's responses. F) A weakly-selectively modulated neuron. Unlike E, this neuron does not show strong differences in PSTHs across task conditions. However, a consistent step-like pattern appears in all PSTHs shortly after stimulus onset, suggesting that its response is primarily driven by task onset.



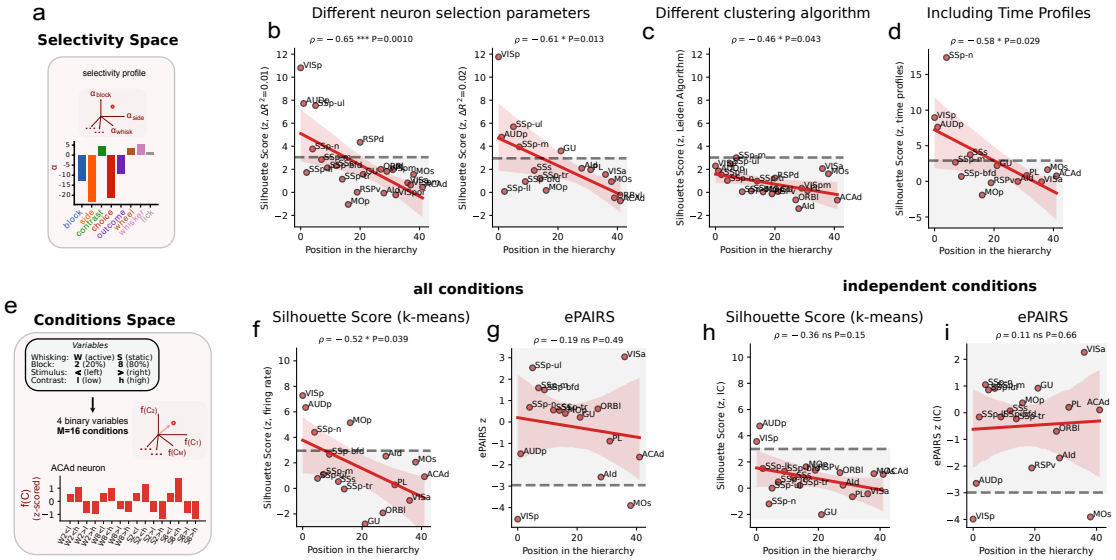
Supplementary Figure 4 Clustering results of all cortical regions in the selectivity space. Subfigures are presented in the same format as Fig. 3bc.



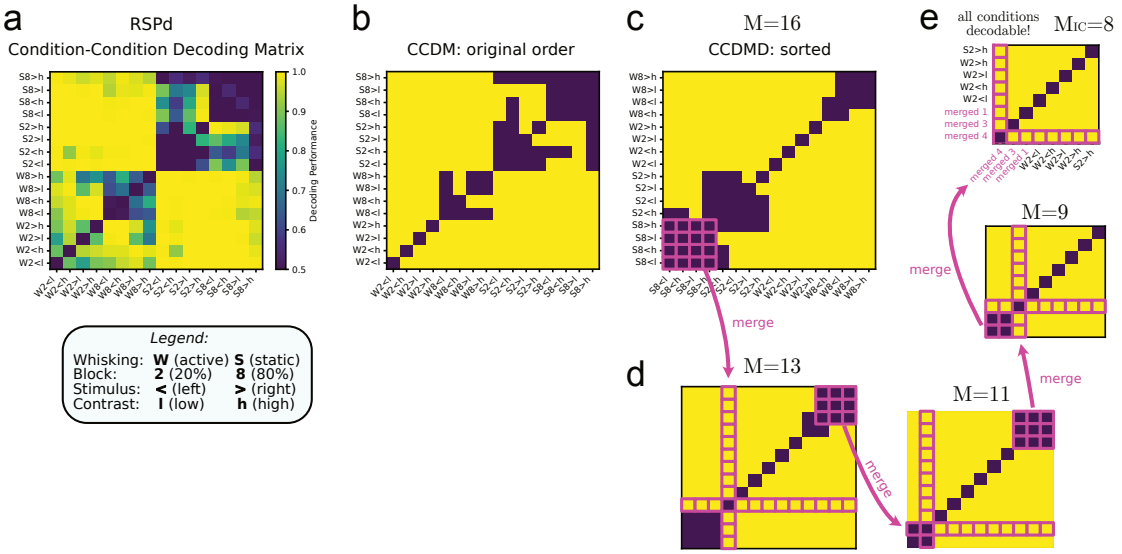
Supplementary Figure 5 Key variables influencing clustering quality varied across brain areas and were often multi-modal, encompassing cognitive, movement, and sensory domains. Variable importance was measured by the reduction in the silhouette score when removing each input variable. (VISp: block, whisker, contrast and choice; AUDp: reward and whisker; SSp-ul: wheel and choice.)



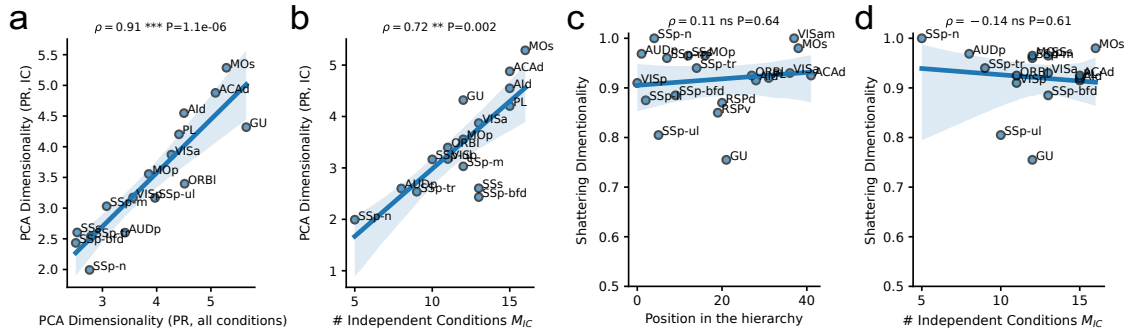
Supplementary Figure 6 Pipeline comparison for analysis of structure in the conditions space. (a) Flow charts for the analysis of structure in the conditions space defined by the mean firing rates of neurons in response to a given set of experimental conditions, comparing the proposed pipeline with the pipeline developed in [2] for the ePAIRS null model comparison (blue arrows). The main methodological difference is that, in our pipeline, the uni-modal null model is fitted on the original data instead of the z-scored and dimensionality-reduced firing rate profiles. Note that in both cases, structural measures such as clustering quality or ePAIRS are computed in the post-processed, z-scored space. **(b)** Comparison of the joint distribution of the post-processed z-scored data (black) compared with the null model inferred using our pipeline (red: log-normal fitting in the original space, then z-scoring and dimensionality reduction) or using the pipeline as described in [2] (blue: z-scoring and reducing the data dimensionality and *then* fitting a Gaussian null model), for an example region (MOs). **(c)** Marginal distributions of the data along the first 4 principal components of the reduced-dimensionality z-scored conditions space (that in which we compute structure). Top row: distribution of the data (black) compared with the null model distribution given by our pipeline (red) and the pipeline in [2] (blue). Bottom row: cumulative distributions of the data in the top row (same color scheme). Note that, in this data, the distributions of the data along the first dimensions of the reduced-dimensionality space (noted as PC0 and PC1) are typically asymmetric and non-Gaussian. The asymmetry is well-captured by a log-normal null model in the original space that undergoes the same preprocessing of the data (red curve), while it is less well-captured by a post-processed Gaussian fit (blue curve). In this example, PCs after the second are more Gaussian. **(d)** Comparison of the goodness of fit, quantified as the normalized log-likelihood of the data given the fitted null distribution, between the two pipelines. p-value computed via a paired t-test.



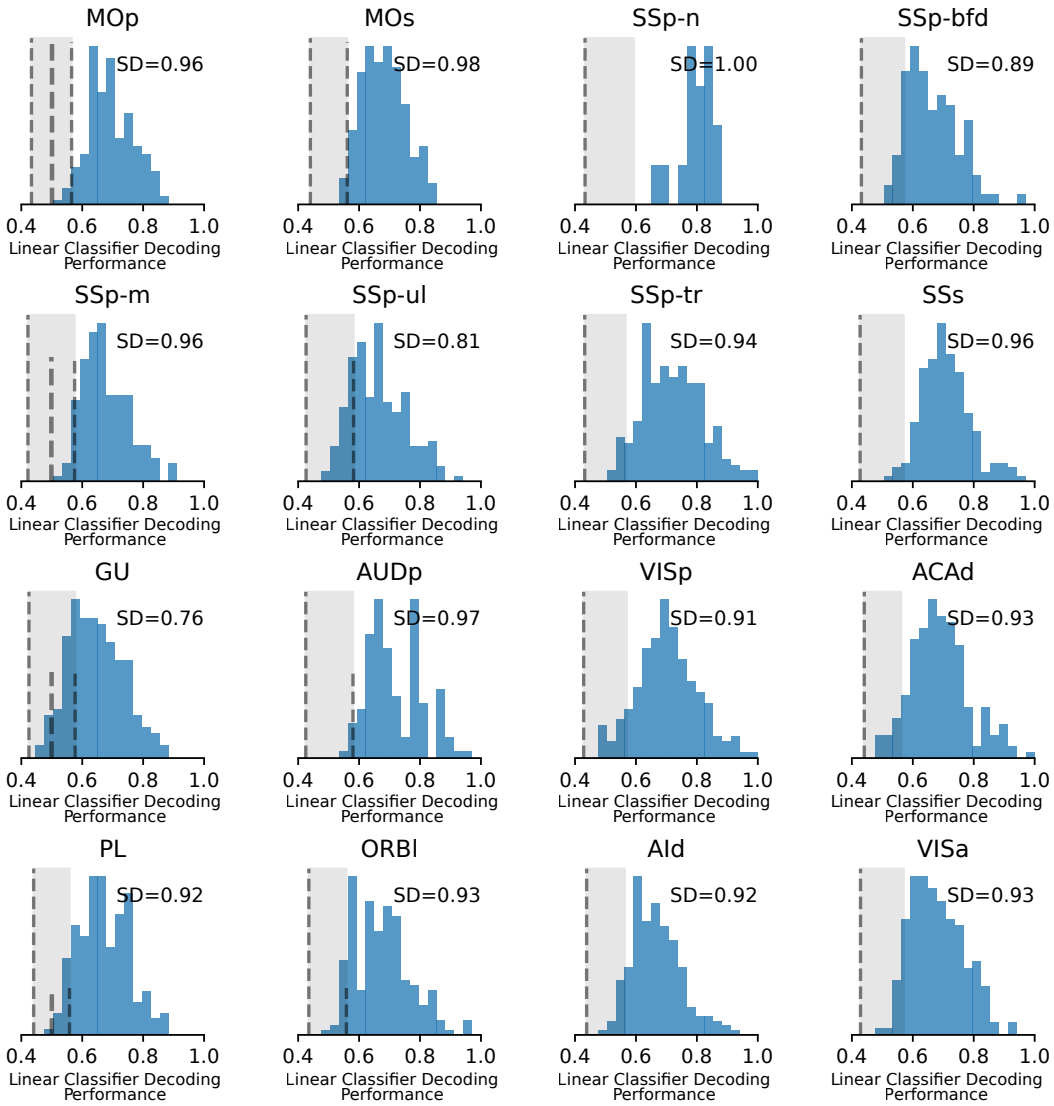
Supplementary Figure 7 Additional analysis of structure in the selectivity and condition firing rate space. Comprehensive clustering analyses show consistent patterns in both selectivity space (a-d) and condition mean firing rate space (e-i) in agreement with the results in Fig. 3. **(a-d)** Analysis of structure in the selectivity space with different parameters and algorithms. **(a)** Visualization of an example neuronal selectivity profile in the selectivity space. (b-d) Same analysis as Fig. 3d (correlation of clustering quality in the selectivity space with the position of the hierarchy) reproduced with **(b)** different neuron selection ΔR^2 thresholds, **(c)** a different algorithm than k-means for finding the clusters (Leiden clustering algorithm), and **(d)** expanding the selectivity space by including the time-varying selectivity profiles (see Methods). **(e-i)** Results of structure analysis in the conditions mean firing rate space using the analysis pipeline explained in Suppl. Fig. 6. **(e)** Illustration of the conditions space framework incorporating four binary variables resulting in 16 experimental conditions. **(f)** Analysis of clustering quality in the 16-dimensional mean firing rate space of the 16 experimental conditions defined by the combinations of the four selected variables (see panel **e** and main text). **(g)** Structure quantified by the ePAIRS analysis (mean nearest-neighbor distance compared to a null model through z-scoring, see Methods) in the same conditions mean firing rate space of panel **f**. **(h, i)** Same analyses of (f, g) in the space defined by the mean firing rate of independent conditions of each individual area.



Supplementary Figure 8 Example of all the individual steps of the independent conditions algorithm applied to the population activity of an example region (RSPd). First, a condition-condition decoding matrix (CCDM) is computed using cross-validated decoding with a linear SVC **(a)**, starting from the $M = 16$ conditions defined by the combinations of 4 binarized variables. The CCDM is then binarized into decodable and non-decodable pairs **(b)** by using a threshold (0.666). Then, the iterative algorithm finds the largest group of conditions that are non-decodable from each other, which corresponds to a dark block in the sorted matrix in **(c)**. These conditions are grouped together and given a new single-condition label **(d)**. The algorithm iterates until all the conditions are decodable from each other **(e)**. The size of the resulting CCDM is the number of independent conditions M_{IC}.



Supplementary Figure 9 **a** Relation between the participation ratio of all $M = 16$ conditions and that of the M_{IC} independent conditions. **b** Relation between the number of independent conditions M_{IC} and the participation ratio of independent conditions. **(c, d)** The Shattering Dimensionality does not correlate with the position on the cortical hierarchy nor with the number of independent conditions M_{IC}.



Supplementary Figure 10 Distribution of decoding performance of $n = 200$ random dichotomies of the independent conditions for all the regions used in Fig.s 5 and 6. Each performance was obtained by a cross-validated analysis of the decoding performance of a random balanced binary classification of the independent conditions found for each region. The cross-validation was performed on mean firing rate vectors corresponding to individual trials. Within each analysis, individual conditions were balanced within each side of the dichotomy to avoid grouped or large conditions dominating the analysis (see Methods). The Shattering Dimensionality was then computed as the fraction of dichotomy decoding performances that was larger than a null distribution found by shuffling the labels of individual trials.