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

Lorenzo Posani[1*], **Shuqi Wang**[2*], **Samuel Muscinelli**[1], **Liam Paninski**[1†], **and Stefano Fusi**[1†]

[1]Center for Theoretical Neuroscience, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027; [2]School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland

The brain is highly structured both at anatomical and functional levels. However, within individual brain areas, neurons often exhibit very diverse and seemingly disorganized responses. A more careful analysis shows that these neurons can sometimes be grouped together into specialized subpopulations (categorical representations). Organization can also be found at the level of the representational geometry in the activity space, typically in the form of low-dimensional structures. It is still unclear how the geometry in the activity space and the structure of the response profiles of individual neurons are related. Here, we systematically analyzed the geometric and selectivity structure of the neural population from 40+ cortical regions in mice performing a decision-making task (IBL public Brainwide Map data set). We used a reduced-rank regression approach to quantify the selectivity profiles of single neurons and multiple measures of dimensionality to characterize the representational geometry of task variables. We then related these measures within single areas to the position of each area in the sensory-cognitive cortical hierarchy. Our findings reveal that only a few regions (in primary sensory areas) are categorical. When multiple brain areas are considered, we observe clustering that reflects the brain's large-scale organization. The representational geometry of task variables also changed along the cortical hierarchy, with higher dimensionality in cognitive regions. These trends were explained by analytical computations linking the maximum dimensionality of representational geometry to the clustering of selectivity at the single neuron level. Finally, we computed the shattering dimensionality (SD), a measure of the linear separability of neural activity vectors; remarkably, the SD remained near maximal across all regions, suggesting that the observed variability in the selectivity profiles allows neural populations to maintain high computational flexibility. These results provide a new mathematical and empirical perspective on selectivity and representation geometry in the cortical neural code.

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

[*] Lorenzo Posani and Shuqi Wang contributed equally to this work
[†]Co-senior authors: Stefano Fusi and Liam Paninski

Correspondence: lorenzo.posani@gmail.com; sf2237@columbia.edu

The brain is highly structured both at the anatomical and functional levels. This large-scale anatomical organization contrasts with the more *local* observation that, within individual brain areas, neural responses are often complex, depend on multiple variables (mixed selectivity (1)), and, in cognitive areas, they are very diverse and seemingly disorganized (see e.g. (2–4)). However, when we look more closely at neural activity, we often see interesting structures both at the level of the responses of individual neurons and at the level of their collective behavior. Despite their complexity, some neuronal responses are similar enough to each other that suggest the existence of modular structures within brain areas ("categorical representations" (3, 5)). At the population level, we often observe low dimensional structures in the representational geometry, both in terms of neural trajectories (6, 7) and in terms of activity space organization when multiple conditions are considered (8, 9)

The organization at the level of population activity (representational geometry, Fig. 1a, blue space) and at the level of single neuron response profiles (response clustering, Fig. 1a, red space) are related to each other in a complex way, which is not fully understood. For example, highly specialized neurons with pure selectivity and linear mixed selectivity to a few independent variables correspond to low dimensional geometries, whereas it has been argued that high dimensionality requires non-linear mixing and diversity of responses (2, 3).

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 (10), Fig. 1d). 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 (11) (*PCA* and *shattering* dimensionality (1, 12)) to characterize the geometry of neural representations.

Previous work has evidenced a relation between the functional properties of neurons within cortical areas and their anatomical organization, such as their position on the visual-prefrontal hierarchy (13, 14). These results suggest that sensory and cognitive areas might employ different coding strategies. Here, we tested whether and how the selectivity and geometrical properties of neural populations systematically change along the mouse cortical hierarchy recently reported by (15) (Fig. 1e, f), which ranks regions in a sensory-cognitive axis starting from primary sensory areas and following the patterns of anatomical connectivity (Fig. 1e).

We found that the selectivity profiles within individual areas are clustered (i.e., "categorical" (3, 5)) only in primary sensory areas. 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 (16). We then developed a mathematical theory relating clustering and the geometry of neural representations, predicting that the maximal PCA dimensionality of representations should be inversely correlated to clustering. This relationship was empirically verified as the PCA dimensionality of the observed representations increased along the cortical hierarchy and was close to its maximum. Moreover, the shattering dimensionality, which measures how many different classifications in the activity space can be solved by a linear readout (1), was close to the maximum (>90%) across all areas.

In the next section, we will present the conceptual framework and define the two neural activity spaces that will be analyzed in the rest of the article: the space of the response profiles of individual neurons (the conditions space) and the representation space where we will study the geometry of the representations and estimate their dimensionality.

Conceptual framework: conditions space and representation space

Let us consider the activity matrix of N neurons recorded during the presentation of M different experimental conditions, such as, for example, their response to different sensory stimuli (Fig. 1a) (3, 12). This matrix defines two complementary spaces: the one spanned by the rows of the matrix, called the "conditions space" (shown in red in Fig. 1a), and that spanned by its columns, called the "representation space" (shown in blue in Fig. 1a).

In the M-dimensional conditions space, individual points represent the response profile of the N single neurons to the M recorded conditions. If neurons are specialized into subgroups that respond similarly to these conditions, they will form clusters (functional groups) in this space, defining what is called a categorical representation (3, 5, 17). Categorical representations have been reported in the orbitofrontal cortex (OFC) of rodents (5, 18), while non-categorical representations were found in the rodent posterior parietal cortex (PPC) (17). In several other articles (1, 8, 19, 20), the authors did not study explicitly whether the representations are categorical, but the observation of very diverse mixed selectivity neurons and high dimensional representations suggest 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 N-dimensional representation space, the relative arrangement of the M points representing individual conditions 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 (1, 8, 19–21). For example, a high-dimensional neural geometry has been shown to support high flexibility (1, 2, 4) and memory capacity (19), while low dimensionality is associated with a better ability to create abstractions and generalize in novel situations (2, 8, 21). Thus, dimensionality is one of the most prominent and implicative features of neural representations.

While these two approaches are, at times, discussed as in conflict with each other, they are, in fact, complementary views of the same data: a population of neurons can, in principle, encode conditions with a categorical or non-categorical representation and, independently, with a low- or high-dimensional geometry (3, 12). Here, we define the positioning within these two axes as the "coding architecture" of a neural population (Fig. 1b). As discussed above, to date, it is unknown whether there are mathematical or empirical constraints on which parts of the architecture space neural populations can occupy. In the next sections, we will introduce and apply novel data analysis methods and analytical derivations to investigate this question on the cortical recordings from the IBL dataset.

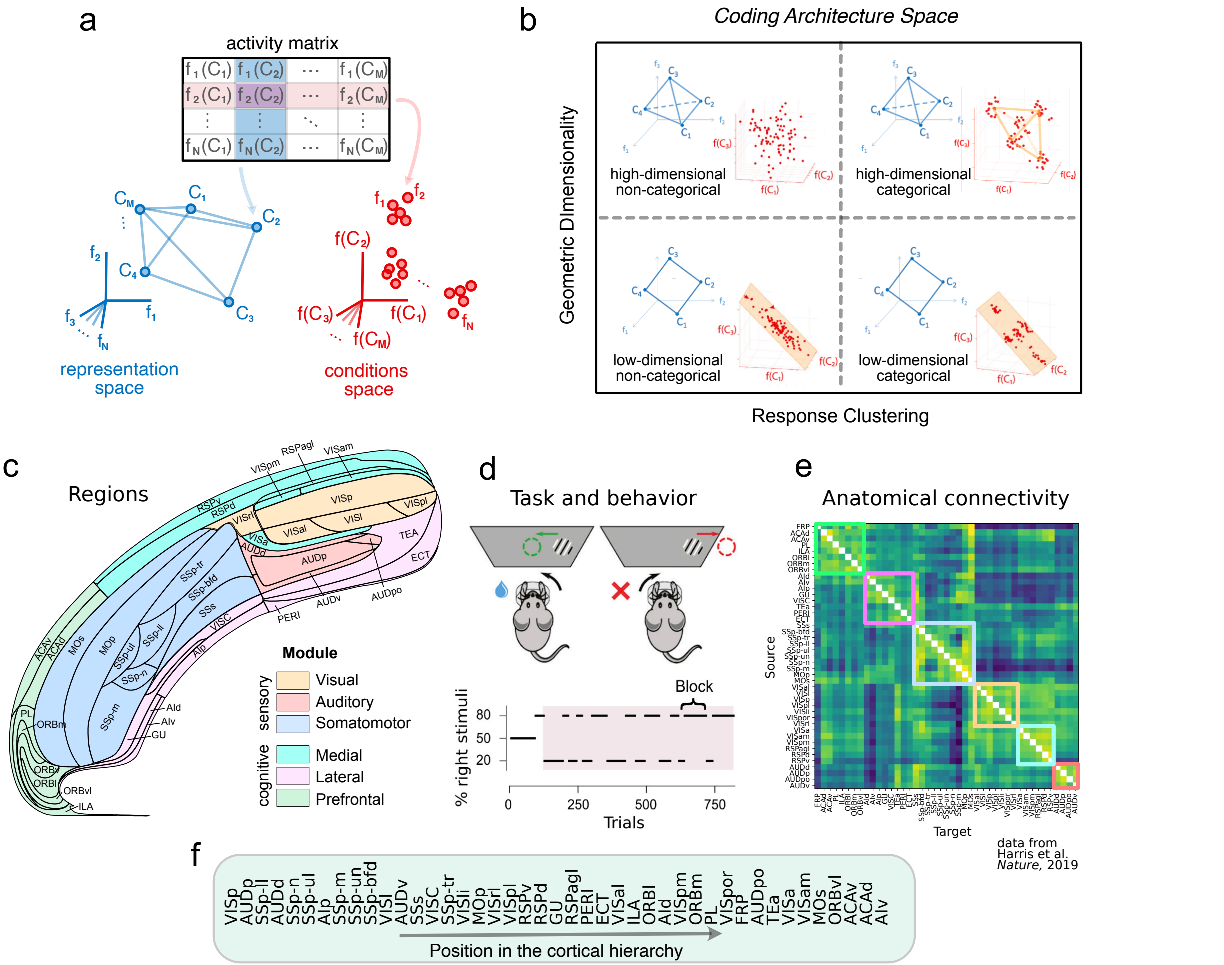

a
activity matrix
f1(C1) f1(C2) ⋯ f1(CM)
f2(C1) f2(C2) ⋯ f2(CM)
fN(C1) fN(C2) ⋯ fN(CM)
representation space
conditions space
CM
C1
C2
C3
C4
f1
f2
f3
fN
f(C1)
f(C2)
f(C3)
f(CM)
b
Coding Architecture Space
Geometric DImensionality
Response Clustering
high-dimensional non-categorical
high-dimensional categorical
low-dimensional non-categorical
low-dimensional categorical
c
Regions
Module
sensory
Visual
Auditory
Somatomotor
cognitive
Medial
Lateral
Prefrontal
d
Task and behavior
Block
% right stimuli
80
50
20
0
250
500
750
Trials
e
Anatomical connectivity
Source
Target
data from Harris et al. Nature, 2019
f
VISp AUDp SSp-ll AUDd SSp-n SSp-ul AIp SSp-m SSp-un SSp-bfd VISl AUDv SSs VISC SSp-tr VISli MOp VISrl VISpl RSPv RSPd GU RSPagl PERI ECT VISal ILA ORBl AId VISpm ORBm PL VISpor FRP AUDpo TEa VISa VISam MOs ORBvl ACAv ACAd AIv
Position in the cortical hierarchy

Fig. 1. Conceptual framework and data structure: (a) The matrix of neural responses of N neurons to M conditions can be analyzed in either its row space (conditions space, red) or column space (representation space, blue). The relative position of conditions in the representation space defines their "representational geometry". If neurons are clustered in the conditions space, they define what is called a "categorical" representation (3, 5, 17). **(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. Here, we call the combination of clustering and geometry the "architecture" of the neural code. 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 (15). **(f)** Cortical hierarchy derived by the anatomical connectivity matrix of panel (e). The order was derived in (15) 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 (15).

The statistics of the neural response profiles reflect the cortical anatomical organization

We first focused on the response properties of single neurons. Since the IBL task is structured with both continuous and discrete variables, we initially focused on neural selectivity, i.e., the profile of responses of single neurons to behavior variables. To analyze neural selectivity, we developed a reduced-rank regression (RRR) model that predicts the activity of single neurons in response to changes in a set of variables from the experimental conditions and the subject's behavior (Fig. 2a-d, see Methods and Suppl. Fig. 1A-D). The core component of the RRR encoding model is a small set of temporal bases that are learned from data and shared across neurons to describe their time-varying responses (see Methods, Suppl. Fig. 1EF). The sharing of the temporal bases significantly reduces the number of parameters and mitigates the issue of overfitting. We used this model to fit the response of each neuron (13733 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. If they perform a correct wheel rotation, they are given a water reward. Stimuli are either shown left and 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 a 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), or decision-related (choice, outcome). After fitting the RRR model to predict trial-by-trial activity (Fig. 2b), each neuron is as-

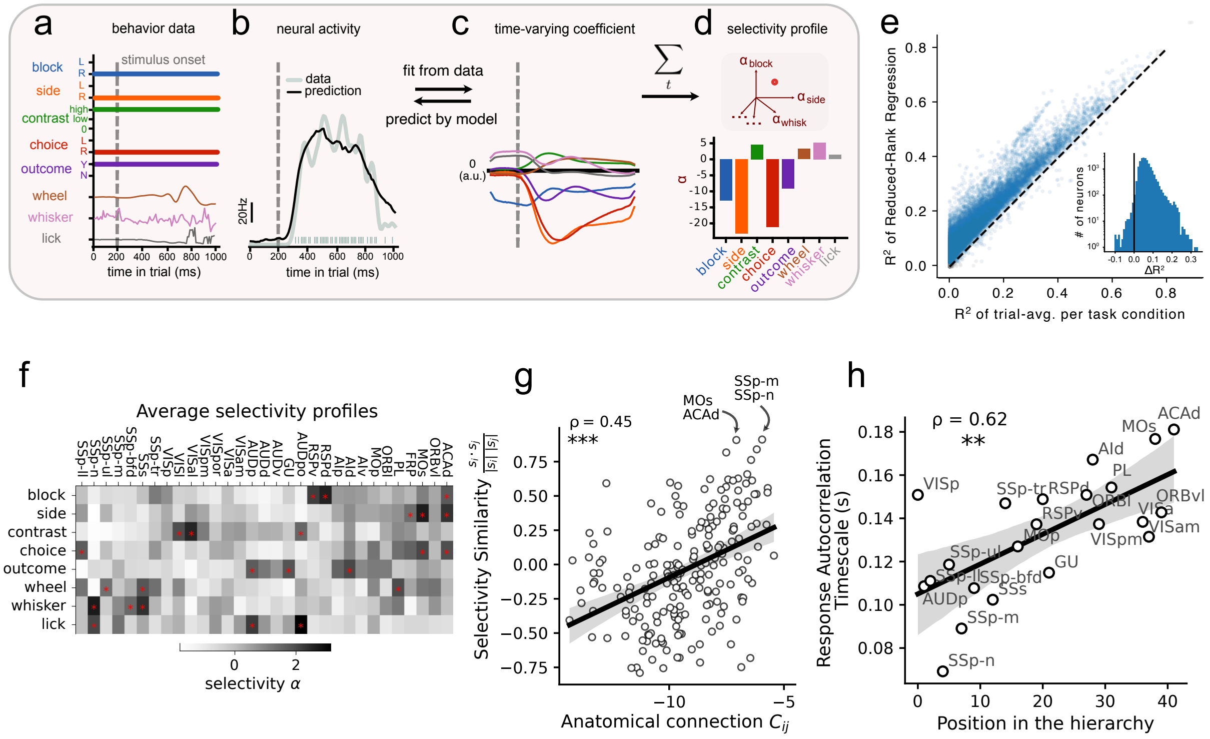


Fig. 2. The reduced-rank regression encoding model as an effective and meaningful tool to estimate single neurons' selectivity profiles. a-d Analysis pipeline for estimating single neurons' selectivity profiles. A linear encoding model is first fit to describe single neuron's temporal responses with respect to behavior data of interest (Methods). Selectivity (α) is computed as the sum of the magnitude of the coefficients across time. Behavior data and neural activity of an example trial are shown for an example neuron, along with the estimated coefficients and selectivity. **e** Goodness-of-fit (three-fold cross-validated R^2, Methods) achieved by our encoding model versus the trial-average estimate per task condition. The outperformance (ΔR^2) is shown in the inset. **f** Average (absolute) selectivity profiles for neurons in the analyzed cortical areas. Areas are first grouped by their modules and then sorted by their hierarchy positions. The selectivity values are normalized per input variable for better visualization. The top three selective areas are marked with red stars for each input variable. **g** Area-to-area functional similarities are strongly correlated with anatomical connectivity (Fig. 1-e). The pairwise functional similarity is measured as the cosine similarity between two selectivity profiles (f). **(h)** Correlation between the position on the hierarchy of single cortical areas with the average autocorrelation timescale of individual neural responses to task variables, estimated using the coefficients of the RRR (see Methods). (***$p < 0.001$, **$p < 0.01$, *$p < 0.05$)

sociated with 8 time-varying coefficients that describe how sensitive the analyzed neuron activity is to each variable in trial time (Fig. 2c). The time-varying coefficients are typically stereotyped across neurons and distinct across input variables (Suppl. Figure 3AB). We then took the sum of these coefficients in time as an estimate of the total effect on the neural responses (Method) and obtained an 8-dimensional selectivity vector for each neuron for further analysis (Fig. 2d, see also Suppl. Fig. 4 for examples of responses from strongly selective neurons).

We first checked the predictive performance of our RRR model against the average firing rate per condition (PSTH), finding a significant improvement across the whole population of neurons (mean $R^2 = 0.16$, mean $\Delta R^2 = 0.064$, Fig. 2e, see also Suppl. Fig. 5A-D for additional benchmark results of model performance and Suppl. Fig. 5E-H for visualization of example neurons). The goodness-of-fit is positively correlated with the mean firing rate and tuning to the behavior movement but not with the timescale of neural responses (Suppl. Fig. 6. To avoid our results being dominated by neurons that are not selective to any variable (10, 22), for the purpose of the following analyses, we only selected neurons whose individual ΔR^2 passed a minimal threshold ($\Delta R^2 > 0.015$, Methods). As discussed in Suppl. Fig. 7, our results are robust to different values of this threshold.

We then used the selectivity profiles of neurons within individual regions to test whether the different selectivity properties of cortical regions reflect the anatomical organization of the cortex. Our hypothesis is that regions that are strongly connected to each other would share similar functional properties that are captured by our selectivity analysis. To test this hypothesis, we computed the average selectivity vector for each cortical region to obtain a region-specific selectivity profile (Fig. 2f). 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 (15) and shown in Fig. 1e). Consistent with our hypothesis, we found a significant positive correlation between selectivity similarity and anatomical connectivity (Spearman correlation = 0.45, p<0.001, Fig. 2g), suggesting that the similarity of selectivity profiles estimated by our RRR model reflects the anatomical organization of cortical regions.

A hierarchical gradient of response timescales in the cortex

Previous and recent work has shown that neurons in regions higher up the cortical hierarchy exhibit progressively longer intrinsic timescales, defined as the autocorrelation time of the spontaneous firing activity (14, 23–27). Computational works have interpreted this result in terms of allowing the

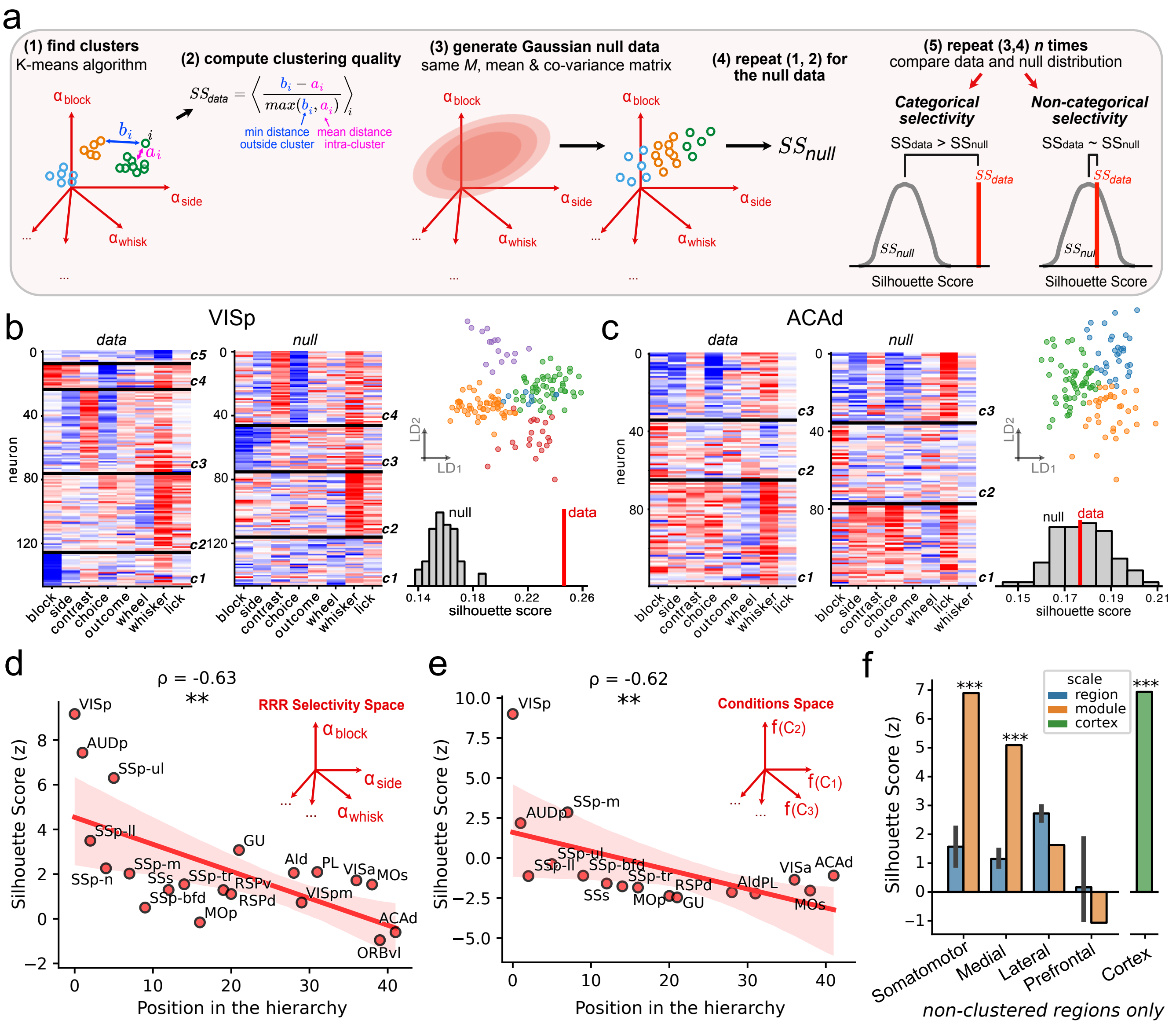


Fig. 3. The clustering quality of areas across hierarchy and across scale. **a** Schematic plot and characterization of categorical versus non-categorical population. For the categorical population, neurons are clustered into multiple (> 1) groups in terms of their selectivity profiles, which results in a silhouette score that is significantly larger than a random mixed null model. The null model is a Gaussian distribution with mean and covariance matrix matched to the data (see Methods and (5)). 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 within 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 the same way. The data silhouette score and its corresponding null distribution are shown on the right, together with a reduced-dimensionality visualization obtained using linear discriminant analysis (individual points are neurons, color refers to cluster labels). **d** For most cortical areas, the neuronal response profiles are very diverse, and the clustering structure is only present in primary sensory areas. The clustering quality, estimated by the z-scored silhouette score, decreased significantly along the cortical hierarchy. **e** Clustering results obtained by applying the pipeline in (a) to the conditions space instead of the selectivity space, compared to the position in the hierarchy as done in (d). **f** Clustering quality of modules as defined in (15) (see Fig. 1) 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.

integration of information over extended periods in cognitive areas to facilitate cognitive functions such as working memory (28). However, it is currently unclear whether cortical regions exhibit a similar gradient for the response time to task variables instead of spontaneous activity.

Thus, we wondered whether we could observe an increase in timescales along the cortical hierarchy defined by Harris et al. (15). Our RRR model allows us to estimate the autocorrelation timescale of neural responses to task variables using the fitted coefficients in time (note that this approach differs from previous methods using the autocorrelation of spontaneous activity; see Methods for details and Suppl. Fig. 8 for estimated autocorrelation and timescales of example neurons). Using this measure, we computed the average autocorrelation timescale τ_0 for all neurons within the cortical regions and compared it with the position of said region along the cortical hierarchy. We found a significant positive correlation between the position in the hierarchy and τ_0 (Spearman correlation = 0.62, p<0.01; Fig. 2h), providing evidence that in mice, similar to what was observed in terms of spontaneous activity, neurons in cognitive areas have longer activity autocorrelations compared to sensory neurons.

Neural selectivity is clustered in primary sensory areas and non-clustered elsewhere

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 ("random mixed" selectivity).

As noted above, our RRR model associates an 8-dimensional selectivity vector to each neuron. Thus, a population within a region can be visualized as a cloud of points in an 8-dimensional selectivity space (Fig. 3a; analogous to the red space in Fig. 1a). As a measure of clustering quality, we used the silhouette score (SS) (29) of clusters identified using k-means in this space. Given a neuron, the SS compares its mean distance to neurons within the same cluster with the distance to the nearest out-of-cluster point (Fig. 1a).

Fig. 3a describes our pipeline to compute the clustering quality of a neural population (see also Methods): first, we use 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). The silhouette score, taken alone, is not indicative of the presence or absence of clustering, as different shapes of the cloud of points can yield widely different silhouette scores even in the absence of clustering (Suppl. Fig. 9). Thus, we compared the value found in the data with a null model that is sampled from a single Gaussian distribution (which is non-clustered by design) while preserving the mean, covariance structure, and the number of neurons of the original data (Fig. 3a step 3). This null model was inspired by the one used for the ePAIRS test in (5). By repeating many null model iterations, we can compare the SS_{data} value with a null model population $\{SS_{\text{null}}\}$. As a measure of clustering quality, we finally take the deviation of the data point from the null distribution, measured with the z-score.

We used this analysis to quantify the clustering structure of selectivity profiles of single neurons 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 mixed (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 on top). As shown in Fig. 3b, it appears that neurons are clustered based on the block prior (c1, c4), whisking power (c2, c5), and a combination of side, contrast, and choice (c3). See also Suppl. Fig. 10 for the clustering results of all cortical regions.

However, the visual aspect of these representations might be misleading, as these clusters might just reflect a heterogeneous selectivity across task variables, which can result in 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 with a null model distribution to obtain a meaningful quantification 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 whisking (c1) and a combination of stimulus side and choice (c3).

When applying this pipeline to all the cortical areas, we found that only three 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 a region in the primary somatosensory area (SSp-ul) (Suppl. Fig. 10). To test which variables contributed to the quality of clusters, we re-ran the clustering analysis by removing each variable individually and measured the drop in silhouette score caused by this removal for each cluster. As shown in Suppl. Fig. 11, 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, whisker, contrast and choice; AUDp: reward and whisker; SSp-ul: wheel and choice).

Since we observed categorical representations in primary sensory areas only, we hypothesized that this structure is inherited by primary sensory regions from the structure of their multi-modal input variables. We further hypothesize that, as information flows along the cortical hierarchy, this categorical structure might get mixed up within areas that inherit these different streams of information to create more flexible representations. If that is the case, we would expect the clustering quality to decrease with the position in the cortical hierarchy. Indeed, we observed that clustering quality, measured by the z-score of the silhouette score against the null model, decreased significantly with the position along the hierarchy (Spearman correlation: -0.63, p<0.01; Fig. 3d), suggesting that the neural code becomes more randomly mixed along the sensory-cognitive hierarchy. Moreover, the trend is consistent if we consider the specific temporal profiles of the time-varying selectivity coefficients on top of the sum across trial time (Spearman correlation: -0.67, p<0.01, Suppl. Fig. 3D).

As reasoned in the Methods, focusing on RRR model-based selectivity instead of simple condition-averages of neural activity has several advantages, including incorporation of non-instructed behavioral variables in the model. However, this approach has the potential drawback of making explicit assumptions about what variables are represented in the neural activity. For this reason, we repeated the same analysis in the conditions space, i.e., the space of trial-average firing rate of neurons in experimental conditions, as proposed in (5). To do so, we first need to define 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 Fig. 4b). Continuous variables (e.g., whisking motion) were discretized to binary values, to make them consistent with the other binary task variables (e.g., block prior). Consistent with the previous result, we observed a negative correlation between the hierarchy and the clustering quality in this 16-dimensional conditions space (Spearman correlation: -0.62, p<0.01; Fig. 3e). See Suppl. Fig. 12 for the clustering results of all cortical regions. Together, these results provide compelling evidence that areas become less categorical the higher they are positioned along the sensory-cognitive axis.

Clustering re-emerges on the mesoscale in lateral and somatomotor modules

The results so far indicate that single regions are typically non-categorical, with exceptions in the primary sensory areas. However, the cortex is organized on both the anatomical and functional levels. For example, Harris et al. (15) proposed a classification of groups of regions into larger *modules* based on their anatomical connectivity and previously known functional properties. We thus reasoned that neural populations that are not clustered within single areas might be so when pooled together in larger, module-scale populations, reflecting the functional specialization of single areas within cortical modules. To test this idea, we ran our clustering pipeline on neural populations created by pooling together neurons from the areas of the individual cortical modules as defined in (15): 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), we were restricted to four modules (somatomotor, medial, lateral, and prefrontal). After pooling the different regions together, we found that neurons were more categorical than the null model in the medial and somatomotor modules (somatomotor: $z \simeq 7$; medial: $z \simeq 5$; Fig. 3e, f), suggesting that areas within these modules are functionally specialized on a larger anatomical scale. Interestingly, the prefrontal and lateral modules were not found to be more categorical than the null model, suggesting that, at least for what concerns the current task, neural populations remained randomly mixed even at the larger module scale (Fig. 3f). Finally, when considering a single population spanning the whole cortex, neurons were found to be significantly more categorical than the null model, reflecting the higher-order organization on a cortical level (Fig. 3f). 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 highly sensitive to differences in the statistics of the neuronal response profiles.

The dimensionality of neural representations increases along the cortical hierarchy

Next, we investigated whether and how the response properties of single neurons are reflected in the population activity of cortical regions. One fundamental measure to characterize the structure of population representations is their embedding dimensionality (12). 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. It has been argued that high diversity in neural responses is required to achieve a high embedding dimensionality, which has important computational implications such as task flexibility (1, 2, 30) or high memory capacity (19). Thus, we sought to investigate how the embedding dimensionality of neural representations evolves along the hierarchy and whether it is related to the response properties of individual neurons within cortical areas.

Different measures of embedding dimensionality have been used in the literature to characterize representations in the activity space. One measure that is directly related to their geometrical shape in the activity space is the *PCA dimensionality*, which is the number of independent dimensions needed to linearly *embed* the structure of neural representations in the activity space. The PCA dimensionality of a set of patterns of neural activity can be summarized by the *participation ratio* (PR) (31), which describes how variance is distributed across the eigenvalue spectrum of the covariance matrix of the patterns (Fig. 4a). If a few eigenvalues dominate (suggesting that only a few dimensions explain most of the variance), the participation ratio will be low. Conversely, if the eigenvalues are more evenly distributed, the participation ratio will be higher, indicating that many dimensions are needed to capture the variability of neural responses.

To analyze the PCA dimensionality of the neural code in the different cortical regions, we defined the conditions (points in the representation space) as the set of combinations of values of the task-relevant variables (e.g., left stimulus & right choice & right block & high whisking power, etc.). However, we can not consider all 8 variables used in the RRR clustering analysis (Fig. 2a), as not all the combinations are well represented in the behavior. For example, when a stimulus is presented on the right-hand side of the screen during a "right" block, it is very uncommon for mice to rotate the wheel to the left. Thus, we selected four variables that span cognitive, sensory, and movement information while being well represented in the behavior: (1) stimulus side, (2) whisking power, (3) block, and (4) stimulus contrast. We binarized these variables (see Methods) to obtain 16 conditions and analyzed the participation ratio (PR) of their centroids (mean firing rate across trials) in the representation space (Fig. 4a, b).

We expect cognitive areas to encode relevant variables with a high-dimensional code, which allows for higher cognitive flexibility (1, 17). Conversely, there is evidence for both high and low-dimensional representations in sensory areas (32–34). Thus, we wondered whether dimensionality in specific cortical areas was related to their relative positioning on the cortical hierarchy. Our analysis revealed that, typically, primary sensory areas have lower PR compared to cognitive areas. Indeed, when comparing the PR of each area with its position on the cortical hierarchy, we found a positive correlation (Spearman corr: 0.75, p<0.001; Fig. 4c). Combined with the results of Fig. 3e, showing that clustering decreases with the hierarchy, this could imply that areas with more categorical representations exhibit a lower PCA dimensionality. This was indeed observed in the data, as we found a negative correlation between clustering and dimensionality in the coding architecture

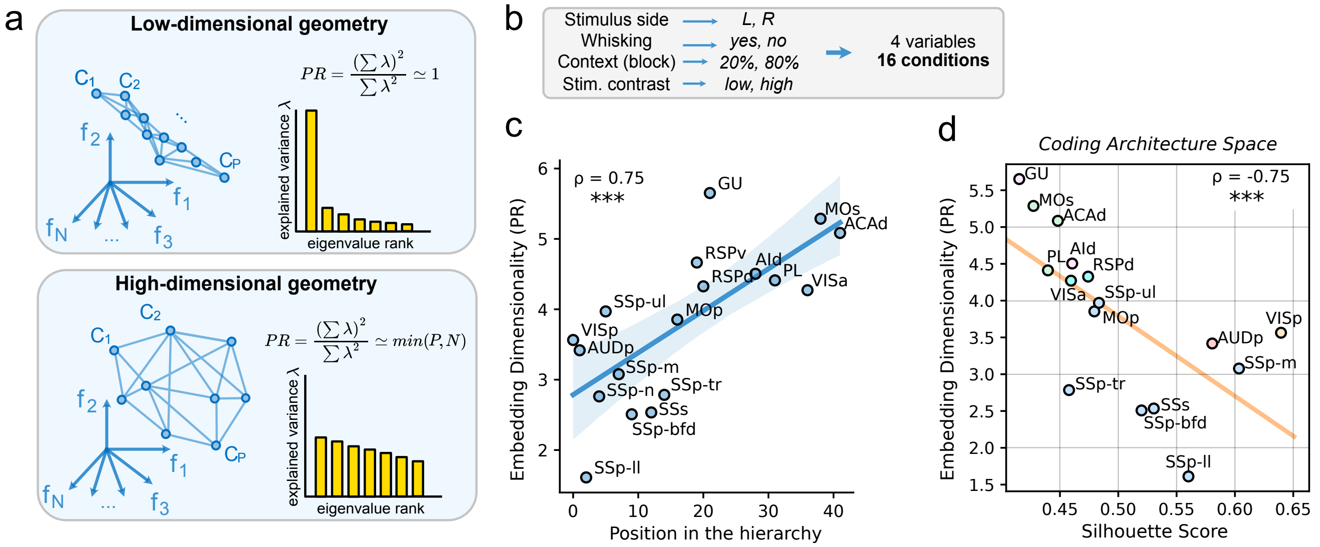


Fig. 4. The embedding dimensionality of neural representations increases along the cortical hierarchy (a) Schematic examples of two geometries with different participation ratios. In the top panel, one direction dominates the covariance structure of the neural data (one data point per condition); hence, the participation ratio (PR) is low. In the lower panel, the covariance structure is evenly distributed along different directions, yielding a high participation ratio. **(b)** To select conditions, we chose the largest set of combinations of motor, sensory, and cognitive variables that are well represented in the behavior. **(c)** The PCA dimensionality, estimated by the PR of the 16 chosen conditions, increased significantly along the cortical hierarchy. **(d)** The PCA dimensionality inversely correlated with the clustering quality measured in the space of the mean firing rates of the 16 conditions (data from Fig. 3e). *** p< 0.001.

space (Fig. 4d), suggesting the presence of a trade-off between clustering and dimensionality in the neural code.

A mathematical model relates the maximum PCA dimensionality to features of the selectivity clusters

One intuitive explanation of the inverse correlation between clustering and dimensionality can be drawn from the fact that the column rank and the row rank of a matrix are the same, and in the absence of noise, the rank is a measure of dimensionality. In other words, the dimensionality in the response profile space is the same as the dimensionality in the activity space. If neural responses are perfectly organized in k clusters, the rows of the activity matrix (Fig. 1a) will be highly correlated, lowering the rank of the activity matrix to the minimum between the number of conditions M and the number of clusters k. Since these "perfect" clusters are, effectively, k *mega-neurons*, the geometry in the representation space must also be k-dimensional (Fig. 5a). While this extreme case gives us an intuitive argument on why clustering in the response profiles is expected to lower the dimensionality of neural representations, it is unclear whether there is a quantitative relation between the two in more intermediate cases (Fig. 5b), such as those we observed in the data.

To understand the relation between response profile clustering and representation dimensionality, we developed a mathematical theory in which we were able to derive the participation ratio of a mixture of Gaussian clusters in the selectivity space (see Methods and Supplementary Material). As we assumed that the positions of these clusters are random (i.e., there is no additional structure besides clustering), the participation ratio we derived is an upper limit to the dimensionality; hence, we call this $PR_{\max}$. We found that $PR_{\max}$ can be expressed as a function of the number of conditions M, the dispersion of clusters σ (i.e., the diversity of the responses within each cluster), and the number of clusters k:

$$\mathrm{PR}_{max} = M\frac{k(1+\sigma^2)^2}{1+M+k(1+\sigma^2)^2} \quad . \qquad [1]$$

From this mathematical form, we can immediately draw a few conclusions. First, 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:

$$\mathrm{PR}_{\max} \xrightarrow[k\to\infty]{} M \qquad \mathrm{PR}_{\max} \xrightarrow[M\to\infty]{} k \quad , \qquad [2]$$

However, things become more nuanced when k, M, and σ are finite and non-zero. First, as shown in Fig. 5c, when k and M are kept fixed, the maximum dimensionality decreases with the clustering quality (expressed as the average silhouette score of a population of Gaussian clusters with given k, M, and σ). This is compatible with the negative correlation we observed in the data, suggesting that representations take the maximum available dimensionality given their respective clustering properties (more on this below). Second, if we fix the quality of clusters and the number of conditions (Fig. 5c, 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) (36). 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).

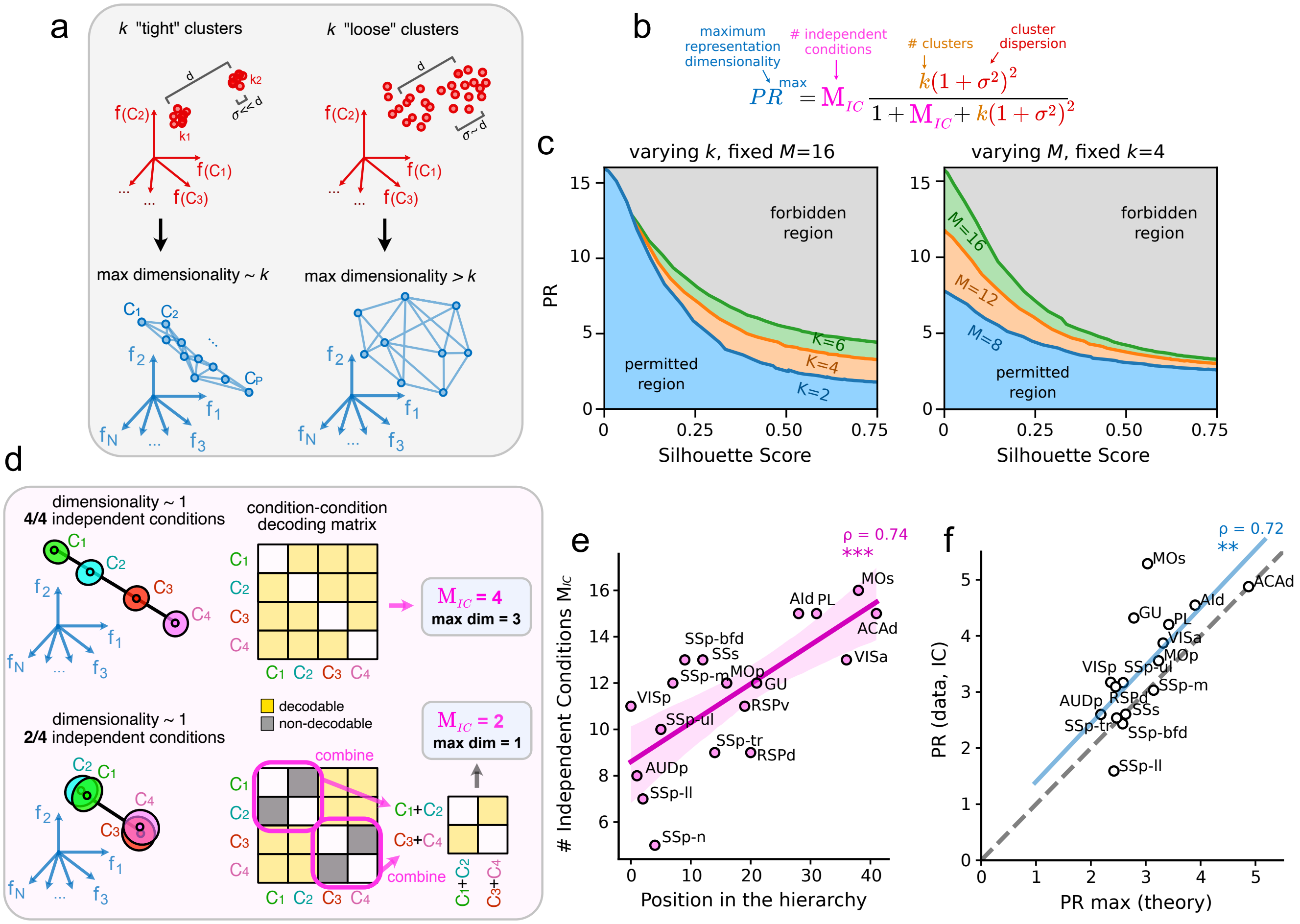


Fig. 5. A mathematical model relates the PCA dimensionality to features of the clusters in the selectivity space (a) Schematic example on how good clustering limits the maximum dimensionality in the representation space. **(b)** Analytical relationship between the features of a collection of clusters (multi-modal Gaussian) in the selectivity space (number of clusters k, number of conditions M, dispersion of the clusters σ) and their maximum PR in the representation space. Here, P is the number of independent conditions, i.e., those conditions out of the 16 total that are separable from each other in the representation space (35). **(c)** Visualization of the relationship in (b) varying the number of clusters k (left panel) and the number of independent conditions (right panel). The clustering quality was converted to a silhouette score to aid comparison with the data. **(d)** Schematic of the algorithm to find the number of independent conditions in the data, with two examples of a low-dimensional geometry with $IC = 4$ (top) and $IC = 2$ (bottom). See Methods and Suppl. Fig. 13. **(e)** The number of independent conditions was found to vary across areas and increase along the hierarchy. **(f)** The theoretical value for max PR obtained by the relation in (b) is predictive of the PR of independent conditions in the data.

Revealing the independent conditions encoded in the representation space

We then wanted to assess whether our Gaussian theory was indeed descriptive of the data. This requires knowing, for each area, the three parameters M, k, and σ. While the number of clusters is given by the k-means algorithm and the dispersion σ can easily be measured in the selectivity space (see Methods), it is less trivial to measure M, the number of encoded conditions. One could be tempted to use $M = 16$, the total number of conditions defined from task variables, but this is not necessarily the number of *independent* conditions that are encoded in the neural activity of a specific area. For example, if an area only represents the stimulus side, the effective number of independent conditions, which we will call M_{IC}, is $M_{IC} = 2$ (left and right). In this case, the maximum representation dimensionality would be PR $= 2$, much lower than the total number of labeled conditions (16). This point is explained in Fig. 5d: on the top panel, we have a low-dimensional geometry with $M_{IC} = 4$ independent conditions. The max dimensionality here is $M-1 = 3$; however, the measured dimensionality is lower (PR $\simeq 1$), indicating a strongly constrained structure in the activity space (conditions are indeed positioned in a straight line). The case depicted in the bottom panel is characterized by a similar dimensionality, PR $\simeq 1$; however, here, the neural code only represents two *independent* conditions (green+blue clouds vs. red+purple clouds); hence, the observed PR should not be surprising, as it is the maximum achievable by the neural code.

To measure the number of independent conditions, we developed 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 population activity vectors (Fig. 5d, see Methods). In brief, the algorithm computes the decoding performance for all the pairs of conditions (1 vs 1) in a putative set of conditions (initially, the 16 behavioral labels). If two conditions are not decodable from each other, they are grouped together and given a new label. The algorithm iterates the 1-vs-1 decoding for the new set of labels and ends when all individual conditions

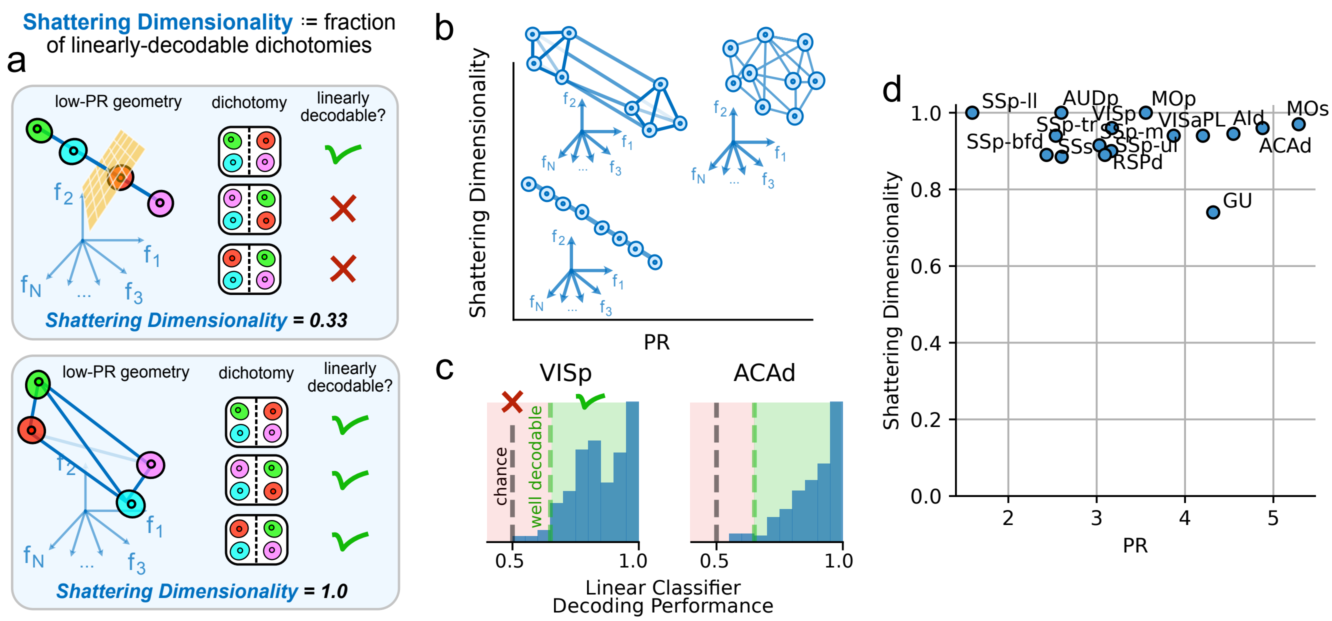


Fig. 6. The shattering dimensionality of neural representations is close to maximal across all regions (a) Definition of Shattering Dimensionality (SD) and schematic representations of two geometries with similar Participation Ratio (PR) and different SD. In this schematic case, both geometries represent $M = 4$ independent conditions. However, in the top panel, only one dichotomy of these conditions is linearly separable (green-blue vs. red-purple), defining an SD value of $1/3$. In the bottom case, all dichotomies are linearly separable, defining 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 performances 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, we used a threshold of decoding performance to define whether a dichotomy was well-decodable (green dashed line). The SD was then defined as the fraction of well-decodable random dichotomies using a cross-validated linear classifier (see Methods). **(d)** Using this definition of SD, all regions were found to represent their independent conditions with close to maximal dimensionality.

are decodable against each other. The size of the final set is taken as the number of independent conditions M_{IC} (see Methods and Suppl. Fig. 13 for a step-to-step example of this iterative process). When we applied this algorithm to the data, we found that 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). Strikingly, the number of independent conditions M_{IC} increased significantly along the cortical hierarchy (Spearman corr: 0.74, p<0.001; Fig. 5e). Thus, cortical regions encode an increasing number of combinations of task variables as they move higher up the cortical hierarchy.

Once we have the value of M_{IC}, we are able to test the predictive power of our analytical theory. As the theory assumes that all the M conditions are independent, we first re-computed the participation ratio of the subset of independent conditions found using our iterative algorithm, called PR$_{IC}$. This measure was found to be strongly correlated with the initial measure of PR (Spearman corr: 0.91, p<0.001; Suppl. Fig. 14). Consistently, PR$_{IC}$ was also found to correlate inversely with clustering (Spearman corr: -0.66, p<0.01; Suppl. Fig. 14). As suggested by the theory, PR$_{IC}$ was also significantly correlated with the number of independent conditions M_{IC} (Spearman corr: 0.81, p<0.001; Suppl. Fig. 14). Using the form in Eq. 1, we compared the value of PR$_{IC}$ found in the data with that predicted by the theory as a function of M_{IC}, k, and σ. Strikingly, we found that this relation was not only correlated (Spearman corr: 0.72, p<0.01; Fig. 5f), but also quantitatively predictive (linear regression: slope=1.03 ± 0.28 SE; intercept=0.37 ± 0.84 SE). Together, these results validate our new theoretical framework, which relates response clustering to the representation dimensionality of the neural code.

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

So far, we have shown that the task conditions (i.e., combinations of values of task variables) are represented in different areas with a PCA dimensionality (measured by the participation ratio, PR) that increases along the hierarchy (Fig. 4c), driven by a decreasing clustering quality (Fig. 3) and an increasing number of independent conditions (Fig. 5).

As shown in Fig. 5a, for a given number of conditions, the PCA dimensionality is a good indicator of whether conditions are randomly spread in the activity space (high PR) or whether there are privileged axes of encoding that unequally contribute to the variance (low PR). However, taken alone, the PR does not discriminate between geometries that can have important differences in their computational properties: in Fig. 6a, we show two geometries with similar PR (low PCA dimensionality) but different properties in terms of separability and flexibility. In the first one (top panel), conditions are highly organized along a single coding direction in the activity space. This arrangement is disruptive for flexibility: out of the three possible 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 (hence the low PR), but conditions are distributed in a more unstructured way. This organization differs from the one above in that here, all possible dichotomies of conditions are linearly separable (Fig. 6a), assuming that the noise is smaller than the shortest distances. One measure of dimensionality that directly quantifies the coding flexibility of a particular geometry is the "shattering dimensionality" (SD) (1, 8), 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, accounting for both heterogeneity of coding direction as well as coding flexibility (Fig. 6b).

We thus measured the shattering dimensionality of the M_{IC} independent conditions represented in the population activity of the different regions in the cortex. We used a threshold of decoding performance to quantify the fraction of dichotomies that are well decodable with a linear classifier (min performance = 0.65). 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 the maximum value of 1 ($SD \gtrapprox 0.9$, Fig. 6d, see also Suppl. Fig. 15). 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.

Discussion

The brain has a clear anatomical organization, and not too surprisingly, we observe that it is reflected by 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 can be related to each other: as the diversity of responses increases, the dimensionality can be significantly higher. Remarkably, the dimensionality is as high as it can be in all the brain areas when compared to the maximal dimensionality determined by the number of independent conditions. 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 and the geometrical structure are related to each other, and this relation can help us understand the computational implications of the neuronal response properties. All the analyses that 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 (2, 4, 37). Flexibility is often defined in terms of the number of input-output functions that can be implemented by a simple linear readout (8). High-dimensional representations allow a network to be highly flexible. 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 (38), which is rather limited for low-dimensional memories (19, 39). 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 in order to increase the dimensionality of the concatenated input/recurrent state (4, 30).

Fortunately, increasing the dimensionality is relatively easy as non-linear random projections already do a surprisingly good job, both in RNNs (30, 40–42) and feed-forward networks (31, 43, 44). 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 (8). 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 (6). 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, in order to achieve the maximal shattering dimensionality, the neuronal responses need to be diverse enough, which often means that they exhibit mixed selectivity (2, 4). Our analysis of the response profiles revealed that they are indeed very diverse: 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 response 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 (45), highlighting a biological complexity that is especially marked for inhibitory neurons (46). 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 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 and well separated. A recent article from the Harris-Carandini lab (47) 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 (48), and eventually also transcriptomic/anatomical identity.

Clustering of response profiles in other species. The recent article by the Kanwisher group (16) 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). Consistently 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 of 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 in 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 (49–51). 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 (8, 19, 21, 34) 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 (16, 52).

A second type of modularity has been characterized in artificial neural networks that are trained to perform multiple tasks (53, 54) or operate in different contexts (12, 55, 56). In this case, it is possible to observe modularity, even when no additional constraints are imposed on metabolic costs and a number of connections (56). 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 (56), and even enable some form of compositionality in which the dynamics of sub-circuits can be reused in different tasks (54).

Limitations of our study. Given all these computational advantages of modular structures, 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 (57). However, there are also other possible explanations that are related to the major limitation of our study: the IBL task is relatively simple, and the animals that are recorded 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. (55) showed that, in RNNs, simple tasks do not require clusters while complex tasks do. Similar considerations apply to other theoretical studies (54, 56). 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.

Author Contributions : L. Posani and S. Fusi conceived the project; S. Wang developed the reduced-rank regression model with guidance from L. Paninski; L. Posani developed the dimensionality analysis and independent conditions algorithm with guidance from S. Fusi; L. Posani and S. Wang analyzed the data; L. Posani and S. Muscinelli developed the analytical theory; all authors interpreted the results and wrote the paper.

Competing interests : Authors have no competing interests to report.

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

ACKNOWLEDGMENTS. We thank Ramon Nogueira for the insightful discussions on the specialization of neural selectivity. The authors are grateful 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.

References

1. M Rigotti, et al., The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
2. S Fusi, EK Miller, M Rigotti, Why neurons mix: high dimensionality for higher cognition. *Curr. opinion neurobiology* **37**, 66–74 (2016).
3. MT Kaufman, et al., The implications of categorical and category-free mixed selectivity on representational geometries. *Curr. opinion neurobiology* **77**, 102644 (2022).
4. KM Tye, et al., Mixed selectivity: Cellular computations for complexity. *Neuron* (2024).
5. J Hirokawa, A Vaughan, P Masset, T Ott, A Kepecs, Frontal cortex neuron types categorically encode single decision variables. *Nature* **576**, 446–451 (2019).
6. CJ Cueva, et al., Low-dimensional dynamics for working memory and time encoding. *Proc. Natl. Acad. Sci.* **117**, 23021–23032 (2020).
7. L Duncker, M Sahani, Dynamics on the manifold: Identifying computational dynamical activity from neural population recordings. *Curr. opinion neurobiology* **70**, 163–170 (2021).
8. S Bernardi, et al., The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967 (2020).
9. S Chung, LF Abbott, Neural population geometry: An approach for understanding biological and artificial neural networks. *Curr. opinion neurobiology* **70**, 137–144 (2021).
10. IB Laboratory, et al., A brain-wide map of neural activity during complex behaviour. *Biorxiv* pp. 2023–07 (2023).
11. M Jazayeri, S Ostojic, Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Curr. opinion neurobiology* **70**, 113–120 (2021).
12. S Ostojic, S Fusi, Computational role of structure in neural activity and connectivity. *Trends Cogn. Sci.* (2024).
13. D Badre, M D'esposito, Is the rostro-caudal axis of the frontal lobe hierarchical? *Nat. reviews neuroscience* **10**, 659–669 (2009).
14. JD Murray, et al., A hierarchy of intrinsic timescales across primate cortex. *Nat. neuroscience* **17**, 1661–1663 (2014).
15. JA Harris, et al., Hierarchical organization of cortical and thalamic connectivity. *Nature* **575**, 195–202 (2019).
16. M Khosla, AH Williams, J McDermott, N Kanwisher, Privileged representational axes in biological and artificial neural networks. *bioRxiv* pp. 2024–06 (2024).
17. D Raposo, MT Kaufman, AK Churchland, A category-free neural population supports evolving demands during decision-making. *Nat. neuroscience* **17**, 1784–1792 (2014).
18. DL Hocker, CD Brody, C Savin, CM Constantinople, Subpopulations of neurons in lofc encode previous and current rewards at time of choice. *Elife* **10**, e70129 (2021).
19. LM Boyle, L Posani, S Irfan, SA Siegelbaum, S Fusi, Tuned geometries of hippocampal representations meet the computational demands of social memory. *Neuron* **112**, 1358–1371 (2024).
20. PK O'Neill, et al., The representational geometry of emotional states in basolateral amygdala. *bioRxiv* pp. 2023–09 (2023).
21. HS Courellis, et al., Abstract representations emerge in human hippocampal neurons during inference. *Nature* pp. 1–9 (2024).
22. NA Steinmetz, P Zatka-Haas, M Carandini, KD Harris, Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576**, 266–273 (2019).
23. CA Runyan, E Piasini, S Panzeri, CD Harvey, Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).
24. JH Siegle, et al., Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**, 86–92 (2021).
25. L Rudelt, et al., Signatures of hierarchical temporal processing in the mouse visual system. *PLOS Comput. Biol.* **20**, e1012355 (2024).
26. M Song, et al., Hierarchical gradients of multiple timescales in the mammalian forebrain. *bioRxiv* pp. 2023–05 (2023).
27. R Zeraati, Y Shi, A Levina, T Engel, , et al., A census of neural timescales across the mouse brain in *Bernstein Conference 2024.* (2024).
28. JF Mejias, XJ Wang, Mechanisms of distributed working memory in a large-scale network of macaque neocortex. *Elife* **11**, e72136 (2022).
29. PJ Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. computational applied mathematics* **20**, 53–65 (1987).
30. M Rigotti, DBD Rubin, XJ Wang, S Fusi, Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Front. computational neuroscience* **4**, 24 (2010).
31. A Litwin-Kumar, KD Harris, R Axel, H Sompolinsky, L Abbott, Optimal degrees of synaptic connectivity. *Neuron* **93**, 1153–1164 (2017).
32. C Stringer, M Pachitariu, N Steinmetz, M Carandini, KD Harris, High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).
33. R Nogueira, CC Rodgers, RM Bruno, S Fusi, The geometry of cortical representations of touch in rodents. *Nat. Neurosci.* **26**, 239–250 (2023).
34. L Chang, DY Tsao, The code for facial identity in the primate brain. *Cell* **169**, 1013–1028 (2017).
35. P Gao, S Ganguli, On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr. opinion neurobiology* **32**, 148–155 (2015).
36. C Langdon, TA Engel, Embedding dimension of neural manifolds and the structure of mixed selectivity. *In preparation.* (2024).
37. WJ Johnston, JM Fine, SBM Yoo, RB Ebitz, BY Hayden, Semi-orthogonal subspaces for value mediate a binding and generalization trade-off. *Nat. Neurosci.* pp. 1–13 (2024).
38. JJ Hopfield, Neural networks and physical systems with emergent collective computational abilities. *Proc. national academy sciences* **79**, 2554–2558 (1982).
39. DJ Amit, *Modeling brain function: The world of attractor neural networks.* (Cambridge university press), (1989).
40. H Jaeger, H Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *science* **304**, 78–80 (2004).
41. W Maass, T Natschläger, H Markram, Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation* **14**, 2531–2560 (2002).
42. DV Buonomano, W Maass, State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.* **10**, 113–125 (2009).
43. O Barak, M Rigotti, S Fusi, The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. *J. Neurosci.* **33**, 3844–3856 (2013).
44. B Babadi, H Sompolinsky, Sparseness and expansion in sensory representations. *Neuron* **83**, 1213–1226 (2014).
45. M Zhang, et al., Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* **624**, 343–354 (2023).
46. H Markram, et al., Interneurons of the neocortical inhibitory system. *Nat. reviews neuroscience* **5**, 793–807 (2004).
47. S Bugeon, et al., A transcriptomic axis predicts state modulation of cortical interneurons. *Nature* **607**, 330–338 (2022).
48. H Yu, et al., In vivo cell-type and brain region classification via multimodal contrastive learning. *bioRxiv* pp. 2024–11 (2024).
49. B Long, CP Yu, T Konkle, Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl. Acad. Sci.* **115**, E9015–E9024 (2018).
50. FR Doshi, T Konkle, Cortical topographic motifs emerge in a self-organized map of object space. *Sci. Adv.* **9**, eade8187 (2023).
51. JS Prince, GA Alvarez, T Konkle, Contrastive learning explains the emergence and function of visual category-selective regions. *Sci. Adv.* **10**, eadl1776 (2024).
52. JC Whittington, W Dorrell, S Ganguli, TE Behrens, Disentanglement with biological constraints: A theory of functional cell types. *arXiv preprint arXiv:2210.01768* (2022).
53. GR Yang, MR Joglekar, HF Song, WT Newsome, XJ Wang, Task representations in neural networks trained to perform many cognitive tasks. *Nat. neuroscience* **22**, 297–306 (2019).
54. LN Driscoll, K Shenoy, D Sussillo, Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nat. Neurosci.* **27**, 1349–1363 (2024).
55. A Dubreuil, A Valente, M Beiran, F Mastrogiuseppe, S Ostojic, The role of population structure in computations through neural dynamics. *Nat. neuroscience* **25**, 783–794 (2022).
56. WJ Johnston, S Fusi, Modular representations emerge in neural networks trained to perform context-dependent tasks. *bioRxiv* pp. 2024–09 (2024).
57. ME Raichle, DA Gusnard, Appraising the brain's energy budget. *Proc. Natl. Acad. Sci.* **99**, 10237–10239 (2002).
58. AH Williams, et al., Discovering precise temporal patterns in large-scale neural recordings through robust and interpretable time warping. *Neuron* **105**, 246–259 (2020).
59. AJ Izenman, Reduced-rank regression for the multivariate linear model. *J. multivariate analysis* **5**, 248–264 (1975).
60. IB Laboratory, et al., Reproducibility of in-vivo electrophysiological measurements in mice. *bioRxiv* pp. 2022–05 (2022).
61. L Posani, Decodanda: a python package for decoding and geometrical analysis of neural activity. *In preparation. Available on Github: https://www.github.com/lposani/decodanda* (2024).
62. F Pedregosa, et al., Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).
63. C Bron, J Kerbosch, Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM* **16**, 575–577 (1973).
64. E Batty, et al., Multilayer recurrent network models of primate retinal ganglion cell responses in *International Conference on Learning Representations.* (2022).
65. SA Cadena, et al., Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology* **15**, e1006897 (2019).
66. L McIntosh, N Maheswaranathan, A Nayebi, S Ganguli, S Baccus, Deep learning models of the retinal response to natural scenes. *Adv. neural information processing systems* **29** (2016).

Materials and Methods

Data structure

We used the International Brain Laboratory (IBL) public data release (10). For each experiment session (153 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 IBL's open API. The following section (1) details the steps for preprocessing this 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 are not encoding any relevant variable, including those neurons that were recorded with a low signal-to-noise ratio.

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 notations 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 (10). 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 (58) 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):

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

Since the squared error between the preprocessed data and model predictions was used in the loss function for optimizing the model (Sec 1), the applied normalization prevented biases due to inherent differences in activity scales and ensured that the predictions of 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. 1-A) 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 of $p(\text{left}) = 0.5$ 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 (Suppl. Fig. 1-A).

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. The distribution of optimal shift for each movement variable is displayed in Suppl. Fig. 2. Last, significant differences were observed in the variance of input values across trials, both for different input variables and for 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. 1-A.

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. 1-E for an example decomposition. Sharing the temporal bases across neurons and input variables significantly reduces the number of parameters (Suppl. Fig. 1-F). 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.

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:

$$\mathcal{L}\left(\boldsymbol{V}, \{\boldsymbol{U}_n^v\}_{n,v}\right) = \sum_n \left(\sum_k \sum_t \left(y_n(k,t) - \hat{y}_n(k,t)\right)^2 + \lambda \sum_v \sum_t \beta_n^v(t)^2 \right). \tag{6}$$

Minimizing this particular loss function is straightforward as a closed-form solution exists (59). 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 $\hat{fr}_n(k,t)$ * 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 $\hat{y}_n(k,t) = \sum_v \beta_n^v(t)\, x^v(k,t)$ (Suppl. Fig. 1-B) is induced by the input variables and varied trial-by-trial. Non-selective modulation, captured by the mean time-varying response $\mu_n(t) = \frac{\sum_k fr_n(k,t)}{K_n}$ (Suppl. Fig. 1-D) is locked to the key events of the trial (stimulus onset in this case) and does not vary trial-by-trial. See Suppl. Fig. 5-GH for examples of two modulation types. Both types play a significant role in modulating neuronal responses. In this work, we focus mostly on the neuronal responses selectively modulated by the 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:

$$\hat{y}_n^{\text{null}}(k,t) = 0, \ \forall k,t,n. \tag{7}$$

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

The outperformance, ΔR^2, defined as:

$$\Delta R^2(\text{model}) = R^2(\text{model}) - R^2(\text{null}), \tag{8}$$

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. 5-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, Suppl. Fig.10, Suppl. Fig.11), we took the sum of the coefficients across time as a measure of the total selectivity of neuron n to input variable v.

$$\alpha_n^v = \sum_t \beta_n^v(t) \tag{9}$$

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. 4 for examples of responses from strongly selective neurons). In the selectivity analysis (Fig.2fg), 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 activity $\boldsymbol{z}_n \in \mathbb{R}^{K\times T}$, where $\boldsymbol{z}_n$ could represent the neural responses to task variables $\hat{\boldsymbol{y}}_n$ (used in Fig. 2h), or the total neural activity $\boldsymbol{y}_n$ (used in Suppl. Fig. 6B), 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

$$c_n(i) = \frac{\sum_k \sum_t z_n(k,t) z_n(k,t+i)}{K}, \ i \geq 0,$$

with z_n sequences being zero-padded where necessary. 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. See Suppl. Fig. 8 for the estimated autocorrelation and timescales of two example neurons (i.e., the neurons shown in Suppl. Fig. 5-GH). The timescale of brain area a (Fig. 2h) is further determined by averaging the values over all the selectively modulated neurons within this area.

Table 1. Notation

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 9, Figure 3bcdf, referred to as clustering analysis in the selectivity space) or the average response in each experimental condition (16 conditions as described in the main text, Figure 3e, referred to as clustering analysis in the mean firing rate 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 5, neural responses are very noisy, and simple averaging may be non-satisfactory (Figure 2e). The selectivity estimated from the encoding model, in comparison, provides a more reliable account of the task-driven variance in the neural responses.

We summarize our clustering pipeline below.

1. Check whether there are more than 50 neurons and only continue if so.
2. Given the response 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 $< \frac{b_i - a_i}{\max(b_i, a_i)} >_i$ 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 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 ($> 90\%$). (See also (47).) 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.

When clustering in the space of mean firing rate, two more 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 response profile is large, we decreased the dimensionality using principal component analysis (60). In Figure 3e, we decrease the dimensionality to 2, but the exact dimensionality does not impact our results (see Suppl. Figure 12 where we decrease the dimensionality to 4).

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 then defined a collection of "conditioned trial" population activity vectors $\{\mathbf{f}_c^k\}$:

$$f_c^k(i) = \frac{\sum_t \delta_t^k(c)\ f_t^k(i)}{\sum_t \delta_t^k(c)} \quad , \tag{10}$$

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 $P = 5$ 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 (31). 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:

$$\mathbf{f}_c = \left\langle \mathbf{f}_c^k \right\rangle_k \quad , \tag{11}$$

To compute the PR for the set of centroids, we first calculated their covariance matrix. We then performed Principal Component Analysis (PCA) on this covariance matrix to obtain its eigenvalues, λ_i. The PR is defined as:

$$\mathrm{PR} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2} \quad , \tag{12}$$

where λ_i are the eigenvalues 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 (61) (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 (62). 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 (61). 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 14, 13, 15.

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. 13.

1. First, we decoded individual condition trial population vectors $\mathbf{f}_c^k$ 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. 13) 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 (63) 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 Suppl. Fig. 13).
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. 13).
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) (1, 8) 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 $\mathbf{f}_c^k$ 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 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 SD was then defined as the fraction of well-decodable dichotomies: $\mathrm{SD} := \frac{1}{n}\sum_d \theta(\varphi_d - \varphi_{\min]})$, where θ is the Heaviside step function. In the present analyses, we used the same threshold of the independent-conditions analysis ($\varphi_{\min} = 0.666$).

The distributions of decoding performances for all the analyzed regions are shown in Suppl. Fig. 15.

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

$$\boldsymbol{x}^\mu = \boldsymbol{z}^\mu + \boldsymbol{\eta}^\mu \quad , \tag{13}$$

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

$$\mathrm{PR} = \frac{\mathrm{Tr}(\mathbf{C})^2}{\mathrm{Tr}\mathbf{C}^2} = \frac{N\langle C_{ii}\rangle^2}{\langle C_{ii}^2\rangle + (N-1)\langle C_{ij}^2\rangle} \quad , \tag{14}$$

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

$$\mathrm{PR} = N\frac{\langle C_{ii}^h + C_{ii}^c + 2C_{ii}^{ch}\rangle^2}{\langle (C_{ii}^h + C_{ii}^c + 2C_{ii}^{ch})^2\rangle + (N-1)\langle (C_{ij}^h + C_{ij}^c + 2C_{ij}^{ch})^2\rangle} \quad . \tag{15}$$

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:

$$\begin{aligned}
C^c_{ii} &= \frac{1}{M}\sum_{\mu=1}^{M}(z_i^\mu)^2 \quad \Rightarrow \quad \langle C^c_{ii}\rangle = 1\,, \quad \langle (C^c_{ii})^2\rangle = \frac{M+2}{M} \\
C^h_{ii} &= \frac{1}{M}\sum_{\mu=1}^{M}(\eta_i^\mu)^2 \quad \Rightarrow \quad \langle C^h_{ii}\rangle = \sigma^2\,, \quad \langle (C^h_{ii})^2\rangle = \frac{M+2}{M}\sigma^4 \\
C^{ch}_{ii} &= \frac{1}{M}\sum_{\mu=1}^{M} z_i^\mu \eta_i^\mu \quad \Rightarrow \quad \langle C^{ch}_{ii}\rangle = 0\,, \quad \langle (C^{ch}_{ii})^2\rangle = \frac{1}{M}\sigma^2 \quad ,
\end{aligned} \tag{16}$$

and

$$\langle C^c_{ii} C^h_{ii}\rangle = \sigma^2, \quad \langle C^c_{ii} C^{ch}_{ii}\rangle = 0, \quad \langle C^h_{ii} C^{ch}_{ii}\rangle = 0 \quad . \tag{17}$$

For the off-diagonal elements, we have

$$\begin{aligned}
C^c_{ij} &= \frac{1}{M}\sum_{\mu=1}^{M} z_i^\mu z_j^\mu \quad \Rightarrow \quad \langle C^c_{ij}\rangle = \frac{1}{k}\,, \quad \langle (C^c_{ij})^2\rangle = \frac{1}{M} + \frac{1}{kM} + \frac{1}{k} \\
C^h_{ij} &= \frac{1}{M}\sum_{\mu=1}^{M} \eta_i^\mu \eta_j^\mu \quad \Rightarrow \quad \langle C^h_{ij}\rangle = 0\,, \quad \langle (C^h_{ij})^2\rangle = \frac{1}{M}\sigma^4 \\
C^{ch}_{ij} &= \frac{1}{M}\sum_{\mu=1}^{M} z_i^\mu \eta_j^\mu \quad \Rightarrow \quad \langle C^{ch}_{ij}\rangle = 0\,, \quad \langle (C^{ch}_{ij})^2\rangle = \frac{1}{M}\sigma^2 \quad ,
\end{aligned} \tag{18}$$

and

$$\langle C^c_{ij} C^h_{ij}\rangle = 0, \quad \langle C^c_{ij} C^{ch}_{ij}\rangle = 0, \quad \langle C^h_{ij} C^{ch}_{ij}\rangle = 0 \quad . \tag{19}$$

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:

$$\begin{aligned}
\langle (C^c_{ij})^2\rangle &= \frac{1}{M^2}\sum_{\mu,\nu=1}^{M} \langle z_i^\mu z_j^\mu z_i^\nu z_j^\nu\rangle \\
&= \frac{1}{M^2}\sum_{\mu=1}^{M} \langle (z_i^\mu)^2 (z_j^\mu)^2\rangle + \frac{1}{M^2}\langle z_i^\mu z_j^\mu\rangle\langle z_i^\nu z_j^\nu\rangle
\end{aligned} \tag{20}$$

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

$$\begin{aligned}
\langle (C^c_{ij})^2\rangle &= \frac{1}{kM^2}\sum_{\mu=1}^{M}\langle (z_i^\mu)^4\rangle + \frac{1}{M^2}\left(1-\frac{1}{k}\right)\sum_{\mu=1}^{M}\langle (z_i^\mu)^2\rangle^2 + \frac{1}{kM^2}\sum_{\mu\neq\nu}^{M}\langle (z_i^\mu)^2\rangle^2 \\
&= \frac{3}{kM} + \frac{1}{M} - \frac{1}{kM} + \frac{1}{kM}(M-1) \\
&= \frac{1}{M} + \frac{1}{kM} + \frac{1}{k} \quad .
\end{aligned} \tag{21}$$

The other terms can be computed following the same steps.

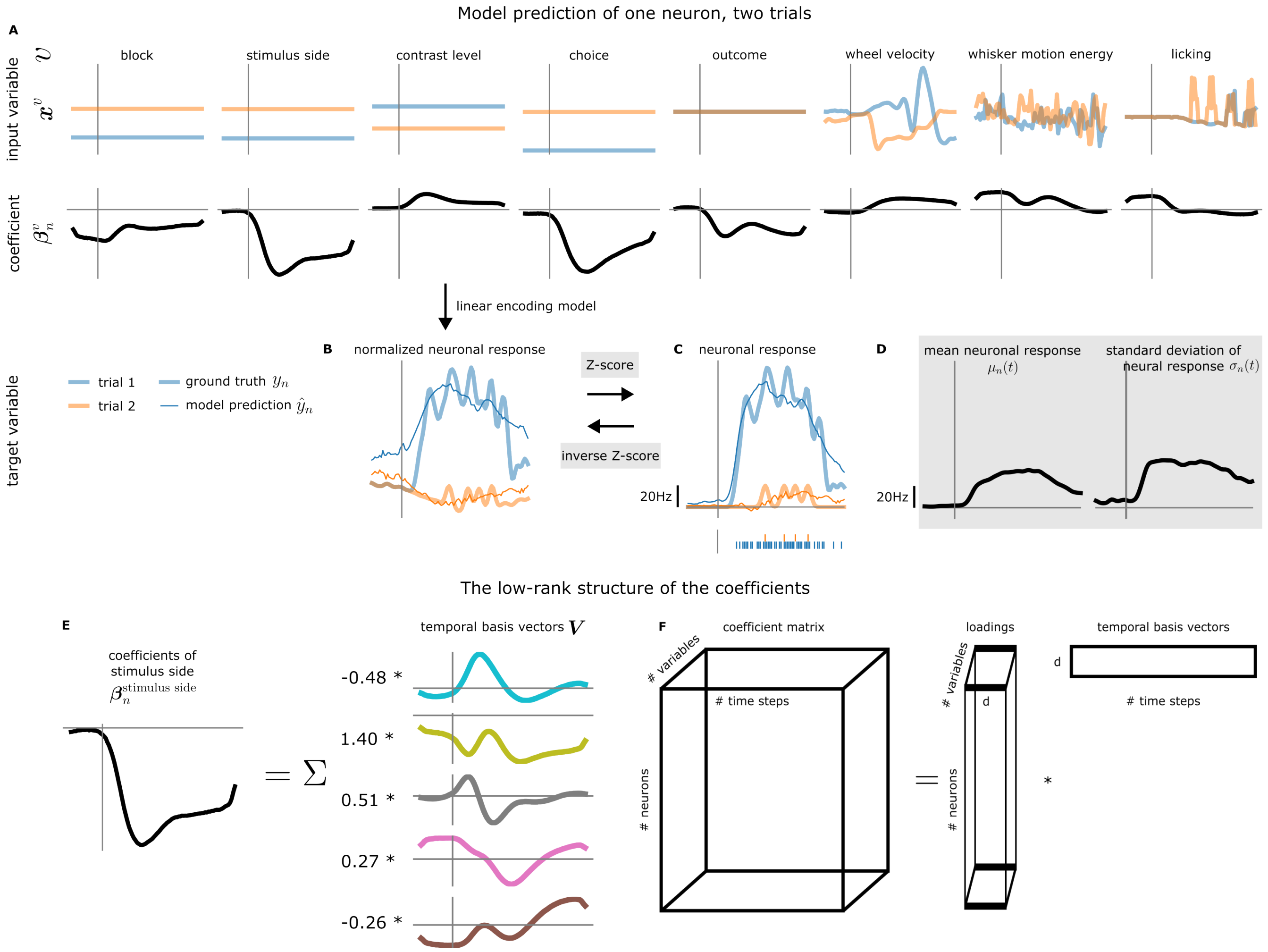
Given that k, M are finite, we can approximate the PR for large N as:

$$\mathrm{PR} \simeq \frac{\langle C^h_{ii} + C^c_{ii} + 2C^{ch}_{ii}\rangle^2}{\langle (C^h_{ij} + C^c_{ij} + 2C^{ch}_{ij})^2\rangle} \quad . \tag{22}$$

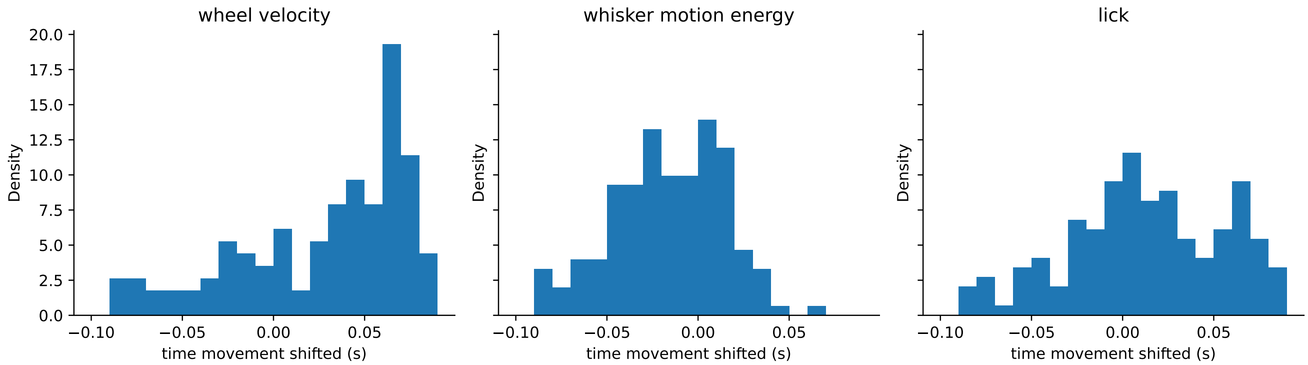
Expanding the square and using the results above for the first and second moments of the covariance matrices, we get to our final expression:

$$\mathrm{PR} \;\simeq\; \frac{\left(1+\sigma^2\right)^2}{\frac{1}{k} + \frac{1}{kM} + \frac{1}{M}\left(1+\sigma^2\right)^2} \;=\; M\frac{k\left(1+\sigma^2\right)^2}{1 + M + k\left(1+\sigma^2\right)^2} \tag{23}$$

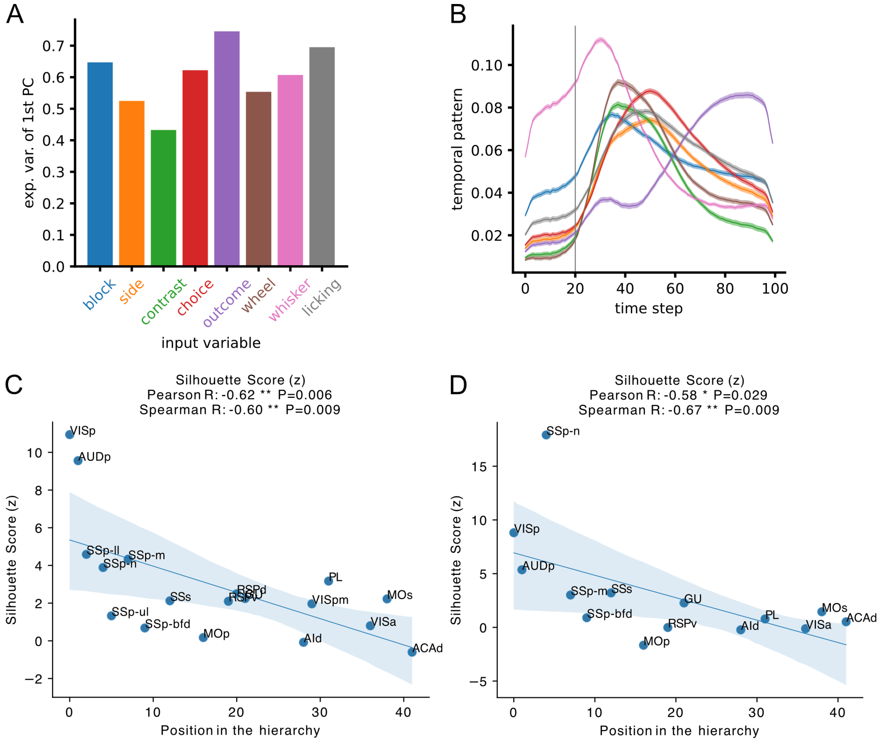
1. Supplementary Figures



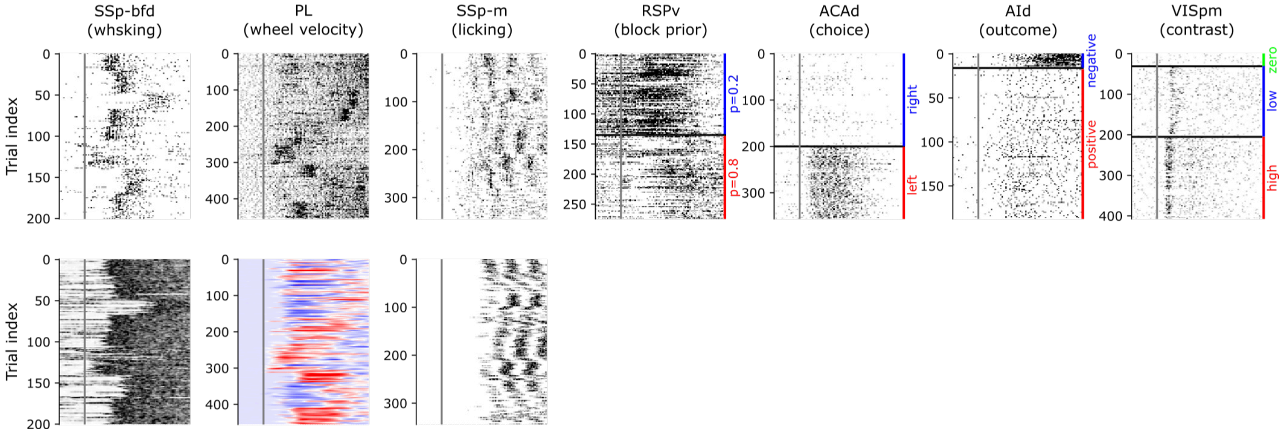
Supplementary Figure 1. Visualization of the RRR encoding model. A-D) Illustration of how the RRR encoding model predicts the neuronal responses using two example trials (blue and orange lines): The model weights the input variables x^v with the corresponding coefficients β_n^v (A) and sums the values over all the input variables to get the prediction $\hat{y}_n$ (B). The prediction (thin and opaque lines) overlays the data (thick and transparent lines). Furthermore, the prediction can be mapped back to the original units of firing rate (C) by applying the inverse Z-score. The mean and standard deviation used in the inverse Z-score (D) are computed during the Z-score step in the target variable preprocessing. 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.



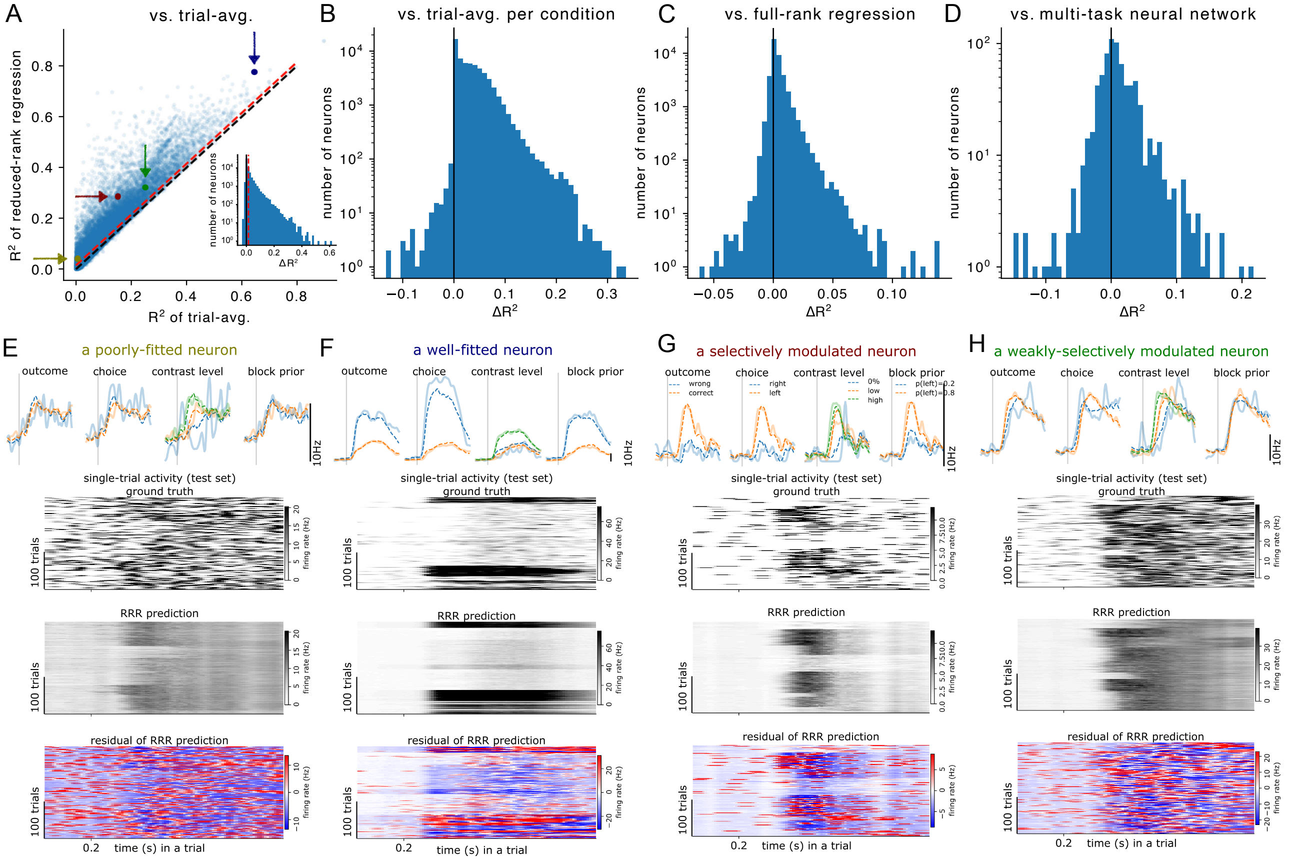
Supplementary Figure 2. Distribution of the optimal time shift across all the sessions. Positive shift indicates that neuronal activity is correlated with future movement.



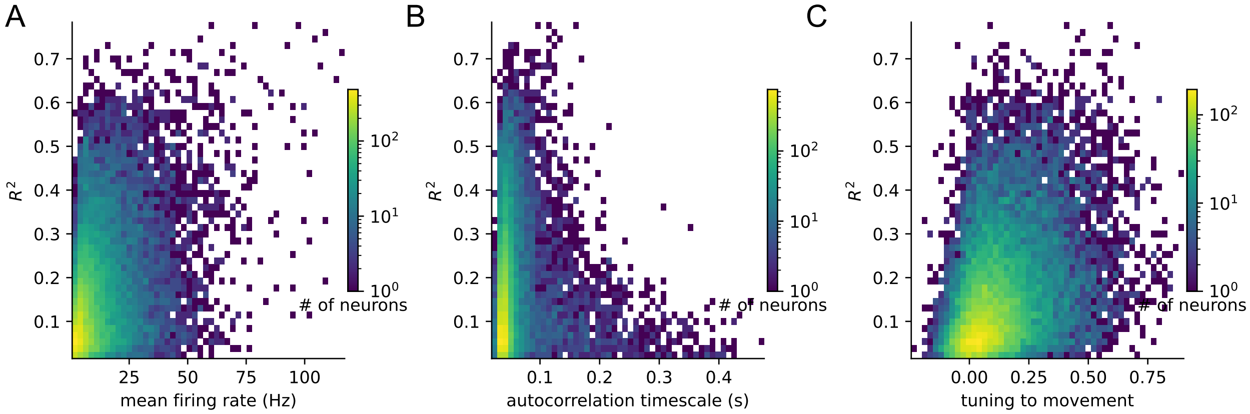
Supplementary Figure 3. The clustering results remain largely consistent regardless of whether we consider temporal patterns. A) The variance in all neurons' time-varying coefficients explained by the first principal component (PC) is high, suggesting that the time-varying coefficients $\boldsymbol{\beta}_n^v$ across neurons largely share a common temporal pattern. B) The common temporal pattern (i.e., first PC) for each input variable. C) Clustering analysis result based on estimating the selectivity of input variables using the projection onto the first PC. D) Clustering analysis result based on estimating the selectivity of input variables as the projection onto the first n PCs. Here, n (typically 2 or 3) is chosen as the number of PCs required to explain over 80% of the variance in all neurons' time-varying coefficients.



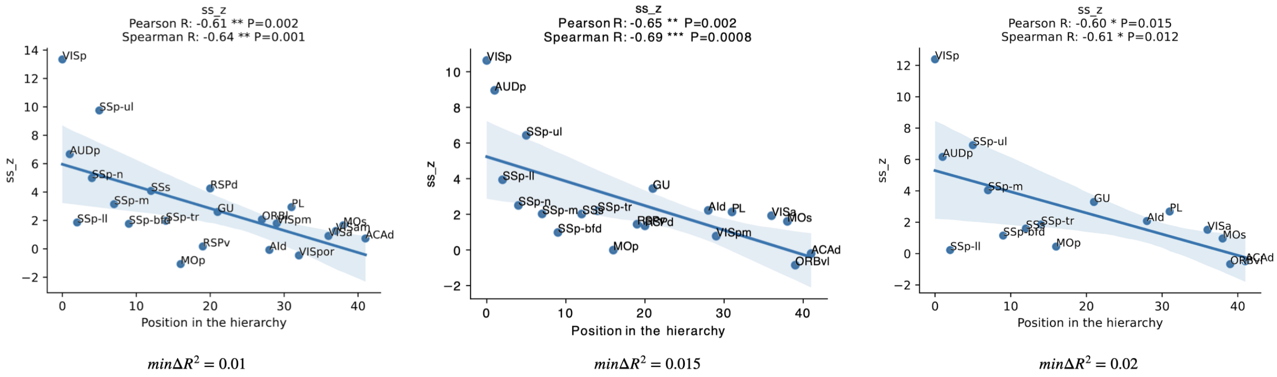
Supplementary Figure 4. Neurons with high selectivity are strongly modulated by the corresponding input variable. Examples of neurons with strong selectivity to each input variable are displayed: the first row shows neuronal activity, while the second row illustrates the associated behavioral movements.



Supplementary Figure 5. The reduced-rank regression (RRR) encoding model effectively overcomes the dominant "noise" in the single-trial activity and efficiently captures task-related variance. A) Goodness-of-fit achieved by the RRR encoding model. Results for all the neurons are shown. The black line indicates when the performance of the two methods is the same, and the red line indicates when RRR is outperforming the trial-average estimate by 0.015, the threshold further used for neuron inclusion. B-D) Comparison of the performance between the RRR method and three baseline methods. $\Delta R^2 = R^2(RRR) - R^2(\text{baseline method})$. B) The trial-average estimate per condition is a non-parametric method that predicts activity based on the average across all trials of the same task condition. C) The full-rank regression model follows the same format as the linear encoding model (Eq 4) but without constraining the coefficient matrix to be low-rank. D) The "multi-task" neural network (as in (60), adapted from (64–66)) shares the encoding of the input variables across neurons in a complicated nonlinear way that involves several feed-forward networks and one recurrent network. E-H) Results were visualized for four representative neurons (corresponding to the four colored dots in A). When plotting the PSTH, the RRR predictions (dashed, thin, thick lines) were overlaid on the data (solid, thick, transparent lines). Colors correspond to different task conditions. For example, the blue and orange outcomes refer to a water reward (+1.) and negative feedback (-1.). The grey vertical lines and 0.2 s represent the onset of the stimulus. When plotting the single-trial activity, the X-axis represents the time in a trial, and the Y-axis represents the trial index. Trials were first clustered by spectral clustering and then sorted by the cluster labels so that trials with similar patterns sit nearby. E) An example of poorly-fitted neurons. A great portion of the variance in neuron responses is not captured by the RRR encoding model and is not apparently correlated with the task. F) An example of well-fitted neurons. A great portion of the variance in neuron responses is driven by the task and reliably captured by the RRR encoding model. G) An example of selectively modulated neurons, characterized by a large ΔR^2 (relative to the R^2). This example neuron showed significant differences under different task conditions, which were clearly visible from both the PSTH and single-trial activity. H) An example of weakly-selectively modulated neurons, characterized by a large R^2 and small ΔR^2 (relative to the R^2). In contrast to the G), no strong differences among PSTHs of different task conditions were observed. However, a consistent step pattern was observed for all the PSTHs shortly after the stimulus onset, indicating its response was mostly modulated by the onset of the task.



Supplementary Figure 6. The goodness-of-fit (cross-validated R^2) is correlated with the mean firing rate and tuning to the behavior movement but not with the timescale of neural responses. The timescale is approximated by the time at which the autocorrelation function of the total neural activity first reaches half its peak value (Methods). Tuning to behavioral movement is measured by the Pearson correlation between the behavioral L^2 distance and the response L^2 distance across all trial pairs. A high correlation suggests that neuronal responses are similar for trial pairs with similar behavioral movements and different for trial pairs with different behavioral movements. The Pearson correlations for the mean firing rate (A), the timescale (B), and tuning to the behavior movement (C) are 0.38, 0.04, and 0.34, respectively.

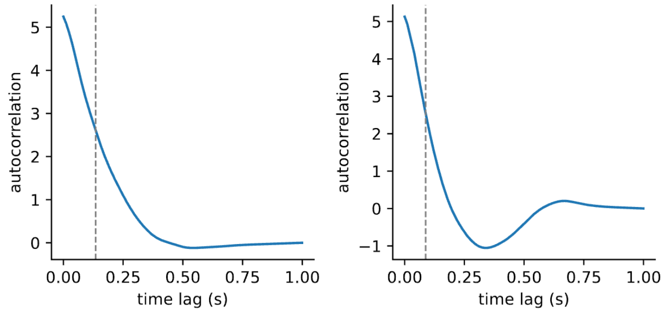


$min\Delta R^2 = 0.01$

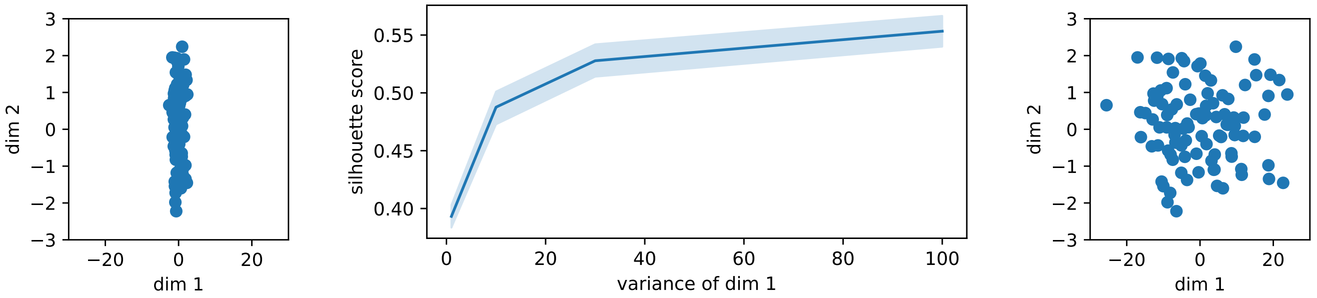
$min\Delta R^2 = 0.015$

$min\Delta R^2 = 0.02$

Supplementary Figure 7. The exact threshold of min ΔR^2 is not crucial for the clustering results. Clustering results were obtained by applying the same analysis as in Fig. 3d, but with varying min ΔR^2 criteria for neuron inclusion.



Supplementary Figure 8. The autocorrelation function of neural responses to task variables is generally smooth. The autocorrelation function of two example neurons' responses estimated by the RRR model (Suppl. Fig. 5-GH) is shown, along with the estimated timescales (the gray vertical lines, Methods).

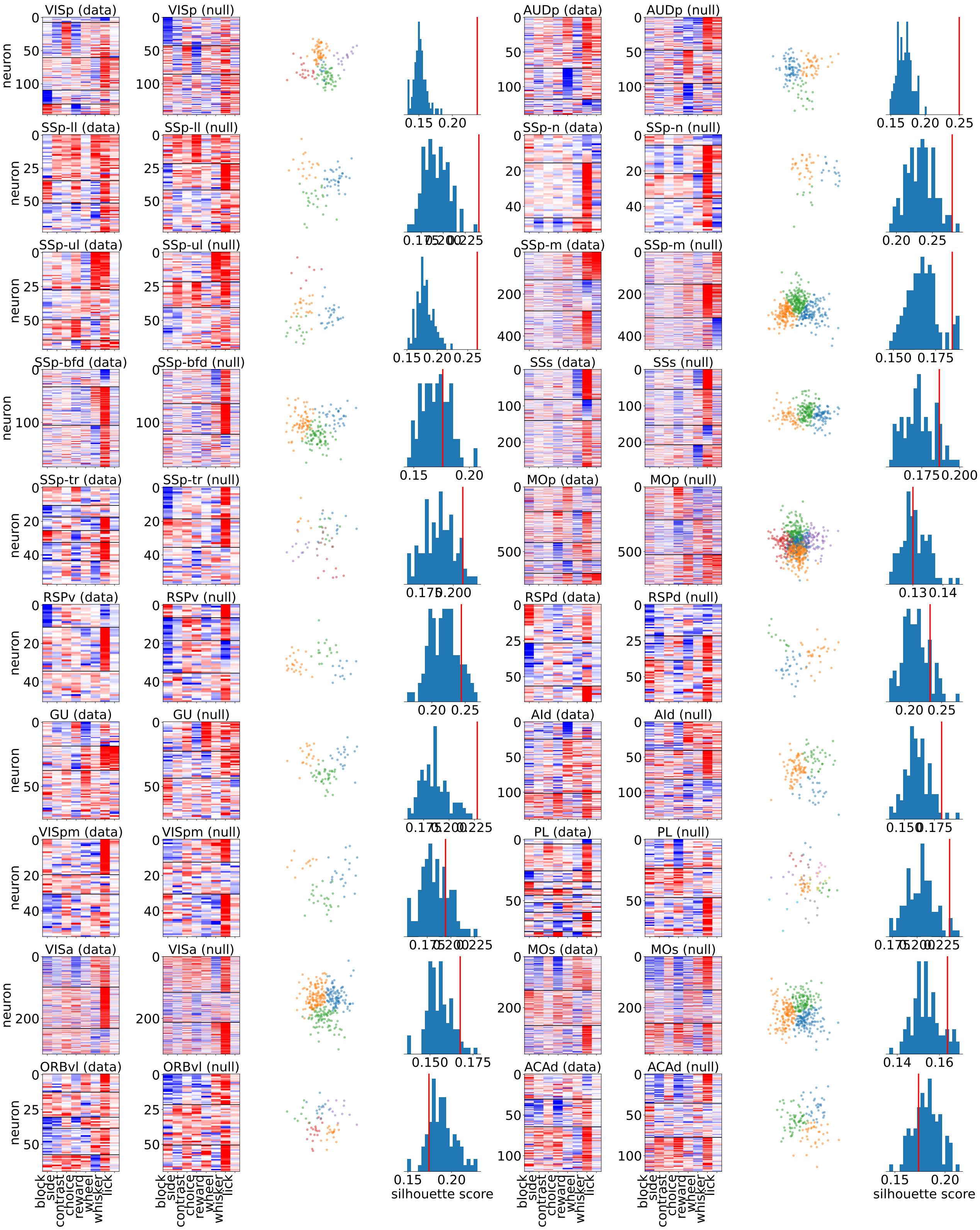


(a) An example dataset of $\sigma_1^2 = 1$.

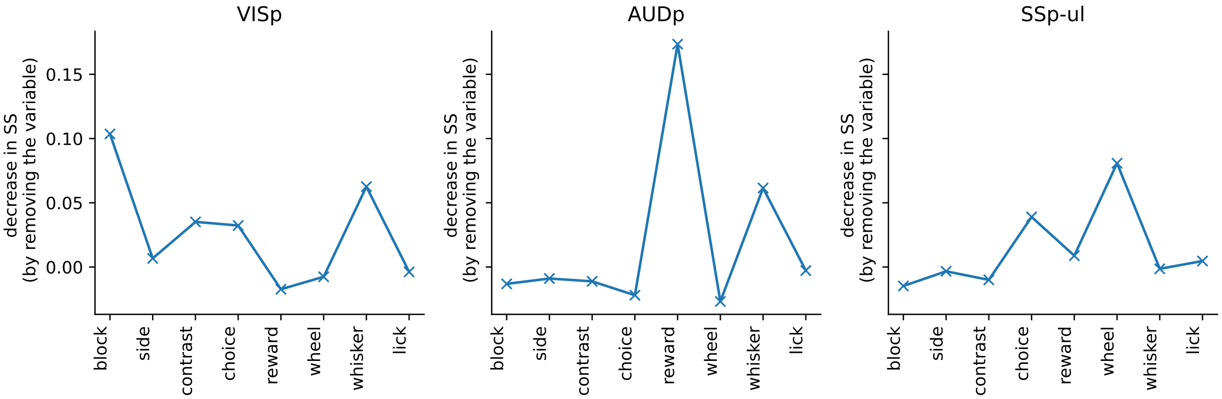
(b) Silhouette scores of datasets with different shapes of the cloud of points.

(c) An example dataset of $\sigma_1^2 = 100$.

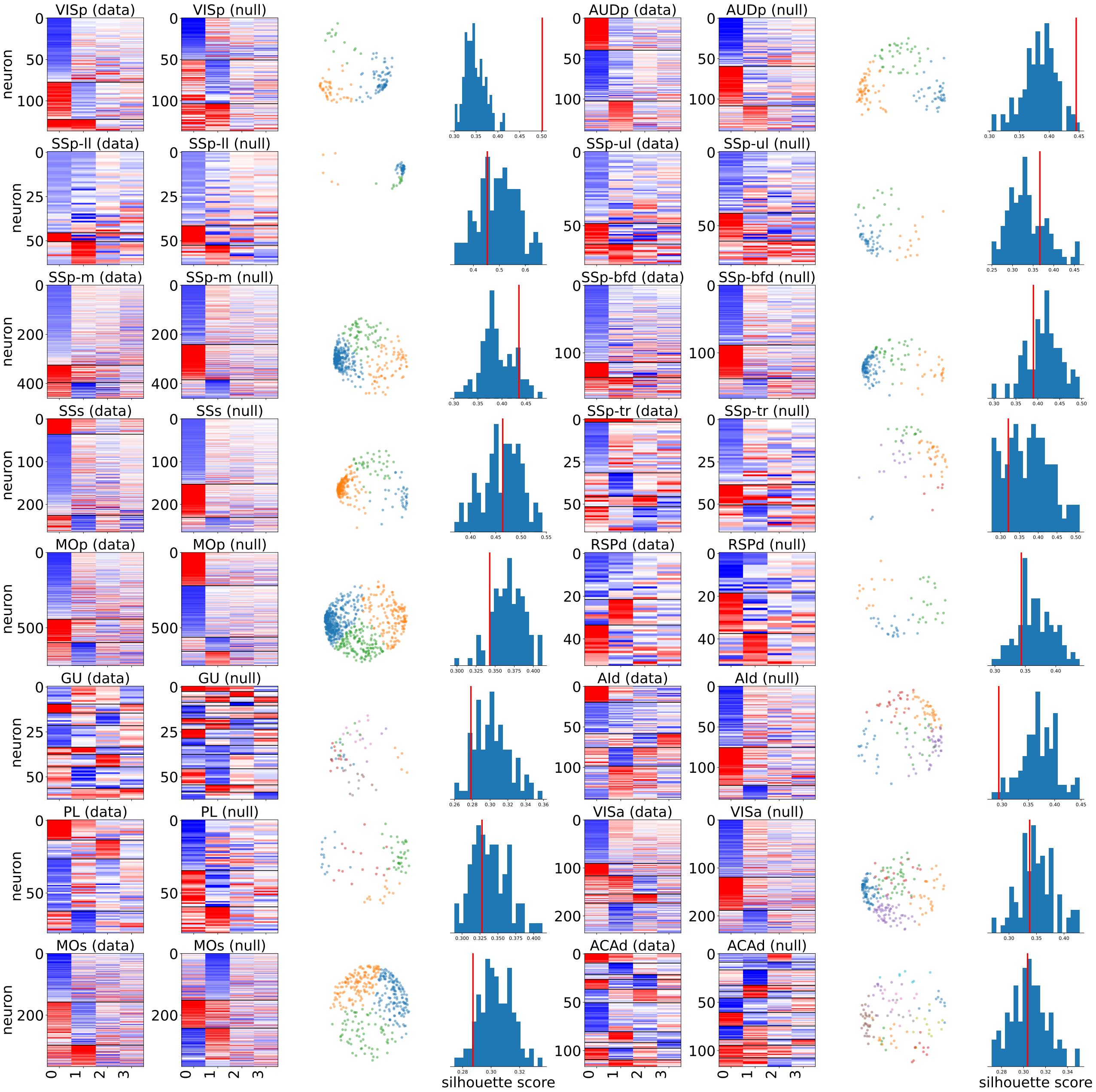
Supplementary Figure 9. The silhouette score, taken alone, is not indicative of the presence or absence of clustering, as different shapes of the cloud of points can yield widely different silhouette scores even in the absence of clustering. The covariance matrix of the data samples was set as $diag(\sigma_1^2, \sigma_2^2)$, where σ_1 varied while σ_2 was set to 1. Each dataset consisted of 100 samples and for each σ_1 we randomly sampled 10 different datasets.



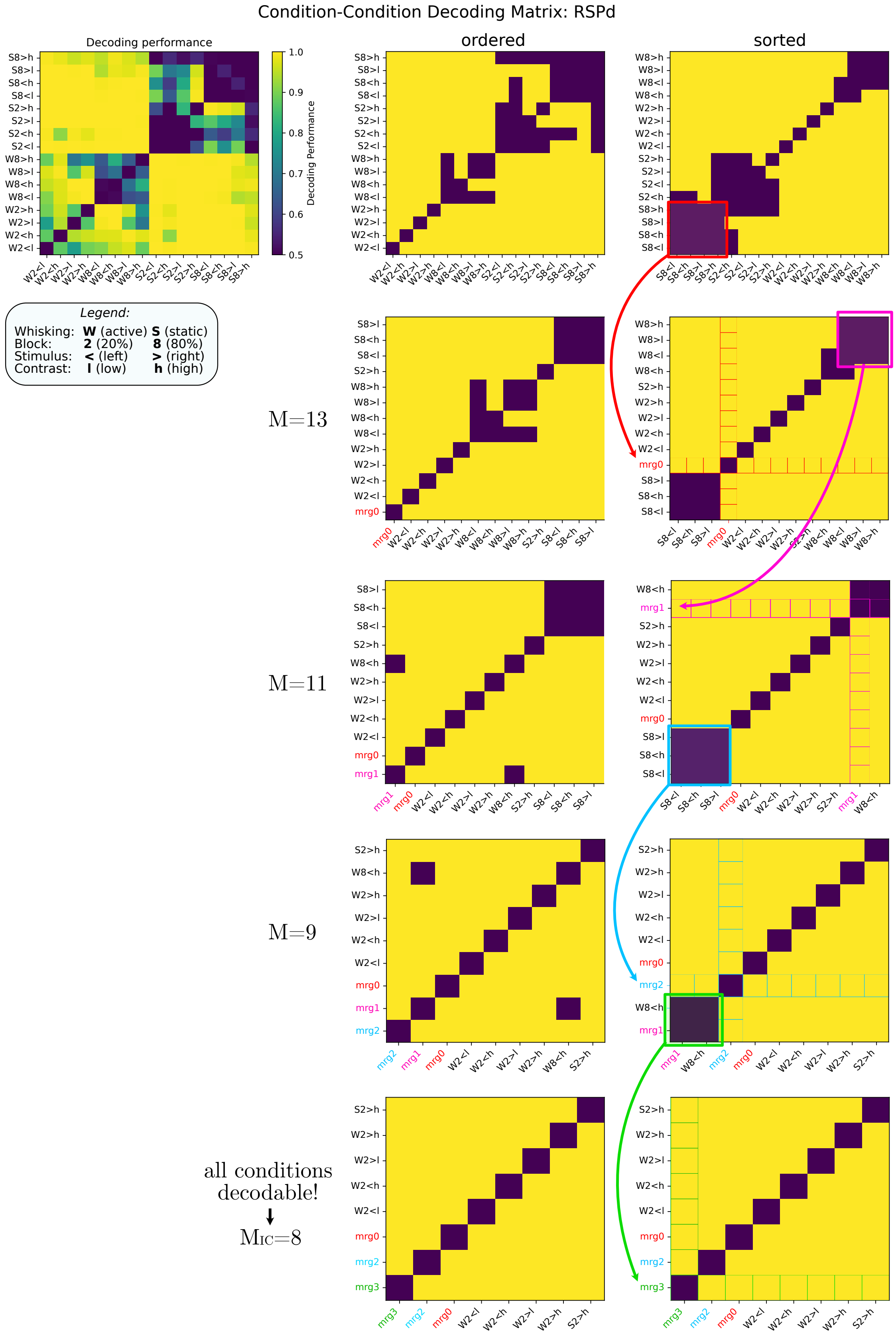
Supplementary Figure 10. Clustering results of all cortical regions in the selectivity space.



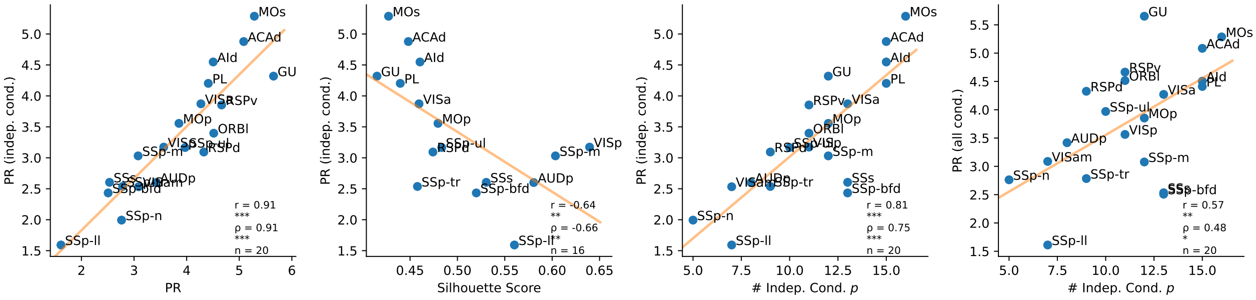
Supplementary Figure 11. Key variables influencing clustering quality varied across brain areas and were often multi-modal, encompassing cognitive, movement, and sensory domains. Variable importance was measured by the reduction in the silhouette score when removing each input variable. (VISp: block, whisker, contrast and choice; AUDp: reward and whisker; SSp-ul: wheel and choice).



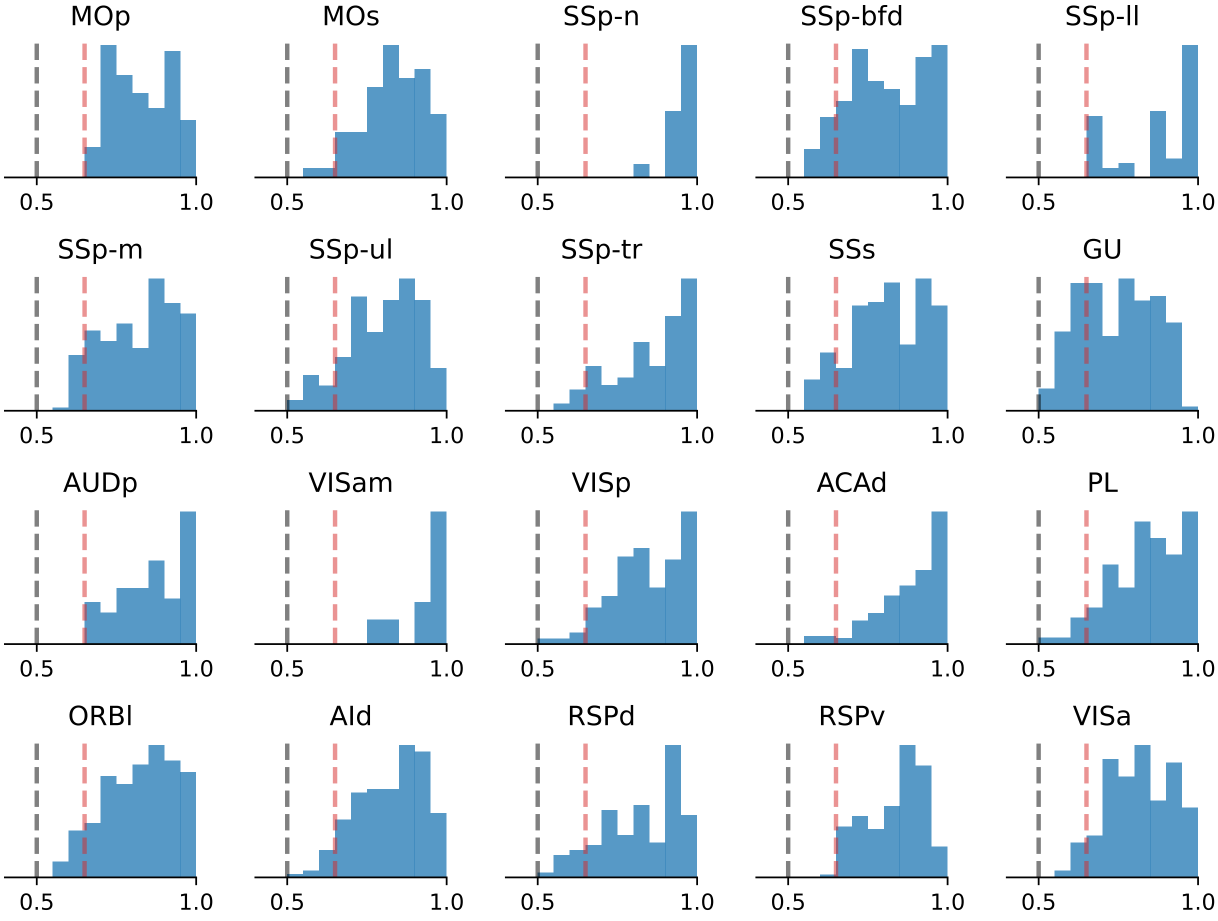
Supplementary Figure 12. Clustering results of all cortical regions in the mean firing rate space.



Supplementary Figure 13. Example of all the individual steps of the independent conditions algorithm applied to the population activity of RSPd.



Supplementary Figure 14. Relation between the number of independent conditions and other geometrical or clustering quantities across cortical areas.



Supplementary Figure 15. Distribution of decoding performance of $n = 200$ random dichotomies of the independent conditions for all the regions used in Fig. 6.