bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sequential and efficient neural-population coding of complex task information

Sue Ann Koay[1], Stephan Y. Thiberge[2], Carlos D. Brody[1,3] *, David W. Tank[1,2] *

* Correspondence: brody@princeton.edu, dwtank@princeton.edu

Abstract

Recent work has highlighted that many types of variables are represented in each neocortical area. How can these many neural representations be organized together without interference, and coherently maintained/updated through time? We recorded from large neural populations in posterior cortices as mice performed a complex, dynamic task involving multiple interrelated variables. The neural encoding implied that correlated task variables were represented by uncorrelated modes in an information-coding subspace. We show via theory that this can enable optimal decoding directions to be insensitive to neural noise levels. Across posterior cortex, principles of efficient coding thus applied to task-specific information, with neural-population modes as the encoding unit. Remarkably, this encoding function was multiplexed with rapidly changing, sequential neural dynamics, yet reliably followed slow changes in task-variable correlations through time. We can explain this as due to a mathematical property of high-dimensional spaces that the brain might exploit as a temporal scaffold.

Introduction

Hypothesized neocortical functions such as predictive coding[1–3] and Bayesian inference[4,5] have emphasized that a crucial component of cortical computation is context: variables that indicate the external state of the world, as well as the internal state of the animal. Our work here, as well as several recent studies[6–10], have indeed found that many different variables are all represented in almost every region of the dorsal cortex. These variables range from sensory and motor, to internal and cognitive. However, the need to simultaneously represent many pieces of information in neural activity can also pose computational challenges for neural systems to overcome. We focus on three such challenges. One, how are multiple variables represented together without crosstalk or interference? Two, how can this neural information be read out if the multiple variables are interrelated? Three, do these representations also include temporal context, an important factor for episodic memory and behavior in general? To answer these questions, we examined the structure of neural population coding during a rich yet well-controlled task, where context-dependent sensory information guided a decision-making behavior.

[1] Princeton Neuroscience Institute, Princeton University; Princeton NJ 08544; USA

[2] Bezos Center for Neural Dynamics, Princeton University; Princeton NJ 08544; USA

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We recorded from large neural populations across the posterior cortex as mice performed a navigation-based visual evidence accumulation task[11,12], which required subjects to generate/utilize time-varying relationships between multiple visual, motor, cognitive, and memory-related task variables. All these dorsal cortical areas were implicated in mice's performance of the task[13], and here we wished to understand how the neurophysiology relates to behavior. Our analysis of neural data is based on conceptualizing the collective activity of neurons as a point in a high-dimensional neural state space, where each coordinate is the activity level of one neuron. It has been observed that in many scenarios, the neural state seemed confined to a lower-dimensional region of the neural state space, termed the "neural manifold"[14–25]. We analyzed the geometrical structure of the neural manifold by examining two types of state-space directions, defined by the neural *encoding* and *decoding* of the above-mentioned task variables as explained below.

To understand how multiple variables were represented together, we considered encoding directions along which the neural state changes if the task variables change. These encoding directions can be thought of as defining a transformation of behavioral information into a neural code, which can also transform relationships between neurally-represented variables. For example, using the same pattern of neural activity to encode two different variables creates interference, in that these two pieces of information cannot then be distinguished from the neural state. However, such an encoding scheme could also support a cognitive function, generalization, by indicating that the two variables are equivalent in the process of computing successively more complex features of the world. This illustrates that the relationships between neural representations can themselves contain extractable information about the expected structure of the world[26], which we quantified by examining decoding directions along which the neural state best discriminates task variables of interest. How information can be decoded must depend on how it has been encoded[27–29], and others have used this to propose encoding schemes based on single-neuron stimulus tuning/filtering properties that optimizes various decoding-based criteria[28–33]. Our approach differs in considering neural-state directions to be the basic encoding unit, and consequences of how neural noise and statistical correlations between task variables modify the relationship between decoding and encoding directions. Our main finding is that all examined posterior cortices had a remarkably consistent structure of encoding and decoding directions, which were multiplexed with sequential neural dynamics that indicated temporal context within the task. Correlated task variables were encoded by approximately uncorrelated neural modes, which supports theories of efficient coding[34,35], and theoretically enables optimal decoding directions to not depend on neural noise levels. The encoding directions varied rapidly in time as neurons were sequentially active in all areas, yet the *geometry* of encoding directions remained much more stable, and linear decoding of task variables could be performed over timescales much longer than that of neural dynamics.

Our findings have implications for longstanding theories of efficient coding. Much work on this subject has focused on how individual neurons in a population should exhibit statistically independent responses in order to represent sensory information with minimal redundancy[36–44], as well as how this function is modified by representational constraints and neural noise[28,29,32,45–49]. Our contribution is threefold. First, we extend

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these notions to neural-state-level encoding and decoding directions, which capture how neural populations coordinate to represent information as a whole. Second, we discovered that efficient coding may not only apply to early sensory information, but also applies in downstream neocortical regions to a set of external and internally-computed variables associated through a learned behavioral task. Third, we report that under dynamic task conditions and also time-varying neural representations, the neural population nevertheless maintained efficient coding of task information through time. We can explain this stability as a mathematical property of large (i.e. high-dimensional) populations of a phenomenon we call "multiplicative neural sequences", where neural responses approximately had the form $w(x)\ g(t)$, with $w(x)$ being a function only of task variables x, and $g(t)$ being a function only of time t. Our results thus link concepts of efficient coding with properties of computation in high-dimensional spaces, through an ethologically important question of how neocortical areas represent multiple interrelated variables to support a complex, dynamic behavior.

Results

We performed cellular-resolution two-photon imaging of six posterior cortical regions of 11 mice trained in the Accumulating-Towers task (Fig. 1a). These mice were from transgenic lines that express the calcium-sensitive fluorescent indicator GCaMP6f in cortical excitatory neurons (Methods), and participated in previously detailed behavioral shaping[11] and neural imaging procedures (Methods), as summarized below.

We trained water-restricted mice in a head-fixed virtual reality system[50] to navigate in a T-maze. As they ran down the stem of the maze, a series of transient, randomly located cues appeared along the right and left walls of the cue region corridor, followed by a delay region with no cues. Mice received a liquid reward for turning down the arm corresponding to the side with more cues, and experienced a longer time-out in the inter-trial-interval (ITI) otherwise. In agreement with previous work[11], all mice utilized multiple cues to make decisions (Fig. 1b). To facilitate comparison of data across animals, trials, and also the ITI, we resampled the behavioral and neural data according to a coordinate that measured progress through the trial ("time in the trial"; see Methods). In addition, we identified thirteen variables that spanned execution and psychophysics of the task: (1&2) the running tally #ipsi and #contra of cue-counts on the sides ipsilateral and contralateral to the recorded brain hemisphere; (3&4) the final tally of ipsi/contra cue counts from the previous trial; (5) the navigational choice to turn right or left; (6) the choice in the past trial; (7&8) whether the (past) trial was rewarded; (9) the virtual viewing angle θ (Fig. 1c); (10) the last value of θ in the past trial; and (11&12) treadmill velocities v_x and v_y; (13) y spatial location in the virtual T-maze.

To obtain neurophysiological data, we first identified the locations of visual areas per mouse using a retinotopic visual stimulation protocol (Fig. 1d; Methods). Then, while mice performed the task, we used two-photon imaging to record from either layers 2/3 or 5 from one of six areas (Supplementary Table 1): the primary visual cortex (V1), secondary visual areas (V2, including AM, PM, MMA, MMP[51]), or retrosplenial cortex (RSC). After

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correcting for brain motion, putative single neurons were identified using a demixing and deconvolution procedure[52]. Neural activities were estimated using fluorescence-to-baseline ratios, and only neurons with ≥ 0.1 transients per trial were selected for analysis. In total, we analyzed 8,759 neurons from 145 imaging sessions. All neural-population level analyses were performed on datasets of simultaneously recorded neurons.

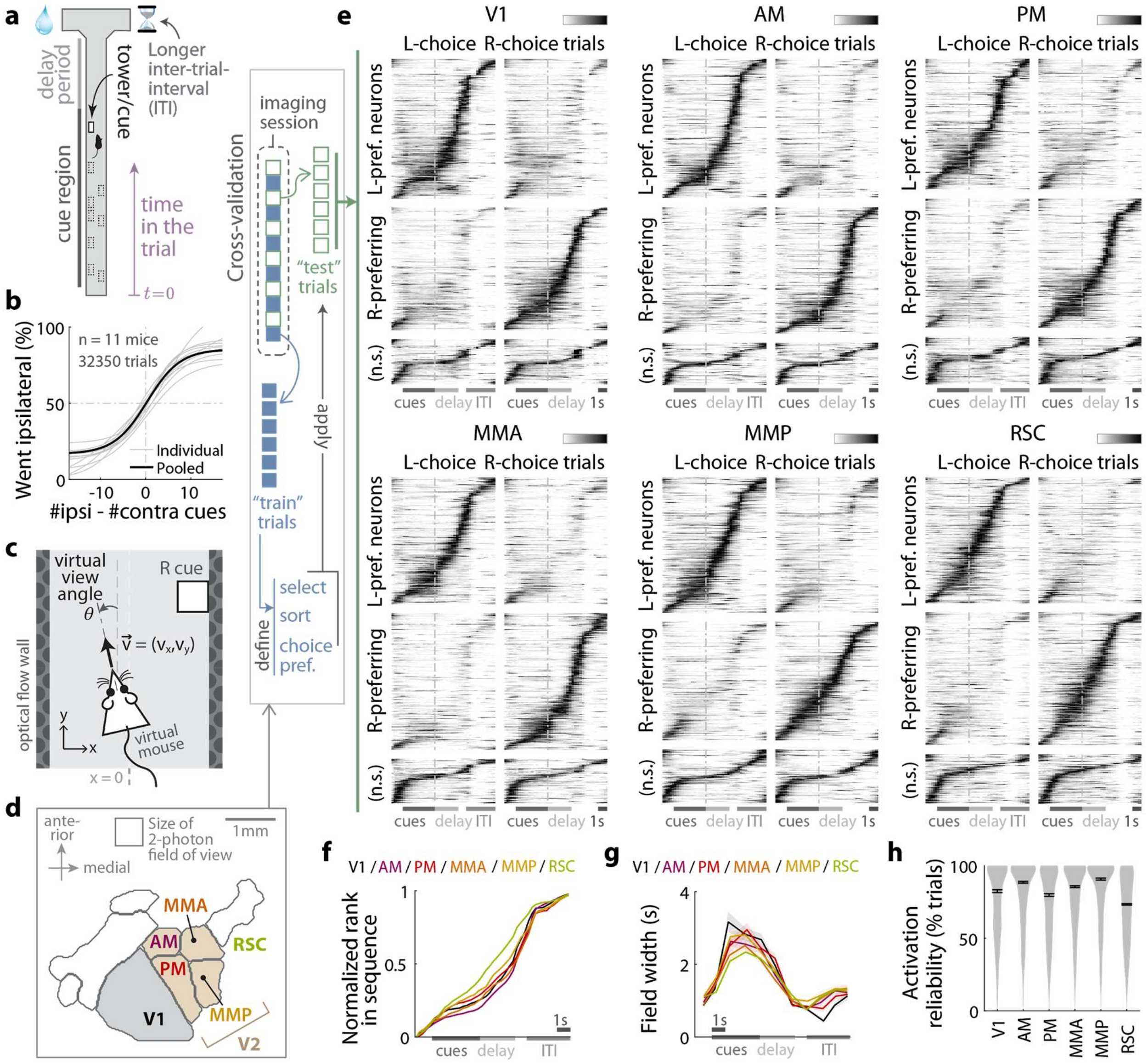


Figure 1. ***Neural populations across posterior cortex are sequentially active during the Accumulating-Towers task.*** **(a)** Layout of the virtual T-maze in an example left-rewarded trial. **(b)** Sigmoid curve fits to psychometric data for how frequently mice turned to the side ipsilateral to the recorded brain hemisphere, as a function of ipsilateral vs. contralateral cue counts. **(c)** Visual and motor task variables analyzed in this study. The virtual viewing angle θ determines the perspective of the virtual scene. $\vec{v}$ is the treadmill velocity. **(d)** Anatomical layout and labels for the six posterior cortical areas in this study. V2 refers to the collection of secondary visual areas. Visual area boundaries were functionally identified per mouse (Methods); shown here are average

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

boundaries for $n = 5$ mice. **(e)** Normalized and trial-averaged activity of neurons (rows), pooling data across sessions for the labeled cortical area. Neurons were divided into left-/right-choice preferring populations, and sorted by the peak activity times in correct preferred-choice trials. “n.s.”: neurons with no significant choice preference according to a t-test (sorted by peak activity averaged across all trials, see Methods). All sorting and normalization factors were computed using a set of training data, whereas these plots were made using the held-out set of testing data. Error trials were excluded in this analysis. **(f)** Rank (normalized to [0,1]) of sorted neurons vs. the peak activity time for that neuron. Data were pooled across sessions for a given area (colors). RSC is significantly different from other regions ($p \leq 10^{-3}$, K-S test). **(g)** Duration of activity fields vs. peak activity times. The activity field is defined as the span of time-points with activity at least half the height of the peak above baseline, in trial-averaged data. Data were pooled across sessions for a given area/layer. Line: Mean across neurons. Bands: S.E.M. **(h)** Distribution (kernel density estimate) of activation reliabilities for neurons in a given area, defined as the fraction of trials in which the neuron is significantly active within its putative activity field. Only neurons with $\geq 50\%$ reliability were shown in (e-g). See Supplementary Fig. 1 for more statistics. Error bars: S.E.M.

The neural state traverses an approximately time-ordered manifold in the course of a trial

Extending previous work[53], we show in a cross-validated sense (Methods) that neurons in all recorded areas were sequentially active vs. place/time in the trial, and could be divided into left- vs. right-choice-preferring subpopulations (Fig. 1e; see Supplementary Fig. 1 for statistics). Differences across areas were small, with RSC having more uniform tiling (Fig. 1f) and slightly more uniform field widths (Fig. 1g) of neuronal activities vs. time. Neurons were reliably active, i.e. in the majority of their preferred-choice trials (Fig. 1h), albeit a bit less so in RSC. These observations are compatible with previous findings of place/time-preferring (and choice-preferring) neurons in mouse cortex[53–58].

As individual neural activities could be ordered in time, we wondered if a similar concept could be applied to the neural manifold. We define “time order” for a manifold by analogy to the case of perfectly repeating sequential neural activity (Fig. 2a). In this idealized case, the neural state trajectory forms a ring-shaped manifold as it travels between the state-space axes of each neuron in the sequence, returning to the first neuron as the sequence repeats. More generally, the neural trajectory could pass through different state-space locations on different trials[54], and yet approximate a ring-like structure (Fig. 2b). At each timepoint in the trial the data across trials formed a point cloud in the neural-state space, and we show below that these point clouds constitute local regions of a neural manifold that can be ordered along a *single* time coordinate. This is what we call global time order for a manifold.

To measure the spread of the per-timepoint point clouds relative to time-related changes, we projected these point clouds onto axes related to the trial-average trajectory (Methods). Fig. 2c shows that the spread projected along the trial-average trajectory was a small fraction of the total length of the trajectory. Also, two point clouds for two distal timepoints had little overlap along the axis between the clouds (Fig. 2d). These results indicate that the neural state at any one timepoint in the trial occupied a relatively small, local region of the

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

neural manifold. If these regions can be ordered in time, they should be nearby (higher overlap) for nearby timepoints, but far from each other (low overlap) for any two distal timepoints (Fig. 2e). The matrix of point-cloud overlap scores for all pairs of timepoints is shown in Fig. 2f. Entries near the diagonal of this matrix correspond to nearby timepoints and entries far from the diagonal correspond to distal timepoints, so we expect sequential activity to correspond to an overlap matrix where high-valued entries should be close to the diagonal. Fig. 2f shows approximately such structure, albeit the overlap between point-clouds in the cue period tended to be high (cf. field widths of neurons in Fig. 1g). This could reflect reduced neural precision in keeping track of place/time along the stem of the T-maze away from boundaries[59,60], particularly since landmarks (cues) were randomly placed on every trial. Signatures of time-orderable structure were true for the neural manifold in all surveyed posterior cortical regions.

The above analyses concerned the Euclidean distance between neural states. If the activity levels of all neurons were to be scaled by the same amount between one timepoint and the next, this would also present as a large change in Euclidean distance, yet there would be no change in the identities of active neurons. To quantify whether there is a change in active neurons vs. time as expected of sequential activity, we examined the angular difference between the centers of the per-timepoint point clouds. As illustrated in Supplementary Fig. 2a, an overall scaling of neural activities generates zero angular difference, whereas a 90° difference is interpretable as a complete change in active neurons since activity levels are nonnegative (explained further in Supplementary Fig. 2b). Fig. 2g shows that at all timepoints, there was an above-chance rate of angular change and thus a sequence-like turnover in active neurons (see Supplementary Fig. 2c for all pairs of timepoints).

We note that the neural manifold may have additional structure related to behavioral factors, e.g. the clear choice-specificity of neural activities (Fig. 1e) imply choice-related bifurcations as illustrated in Supplementary Fig. 2f. We next quantify how multiple task variables were reflected in the local (i.e. per-timepoint) structure of the neural manifold.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

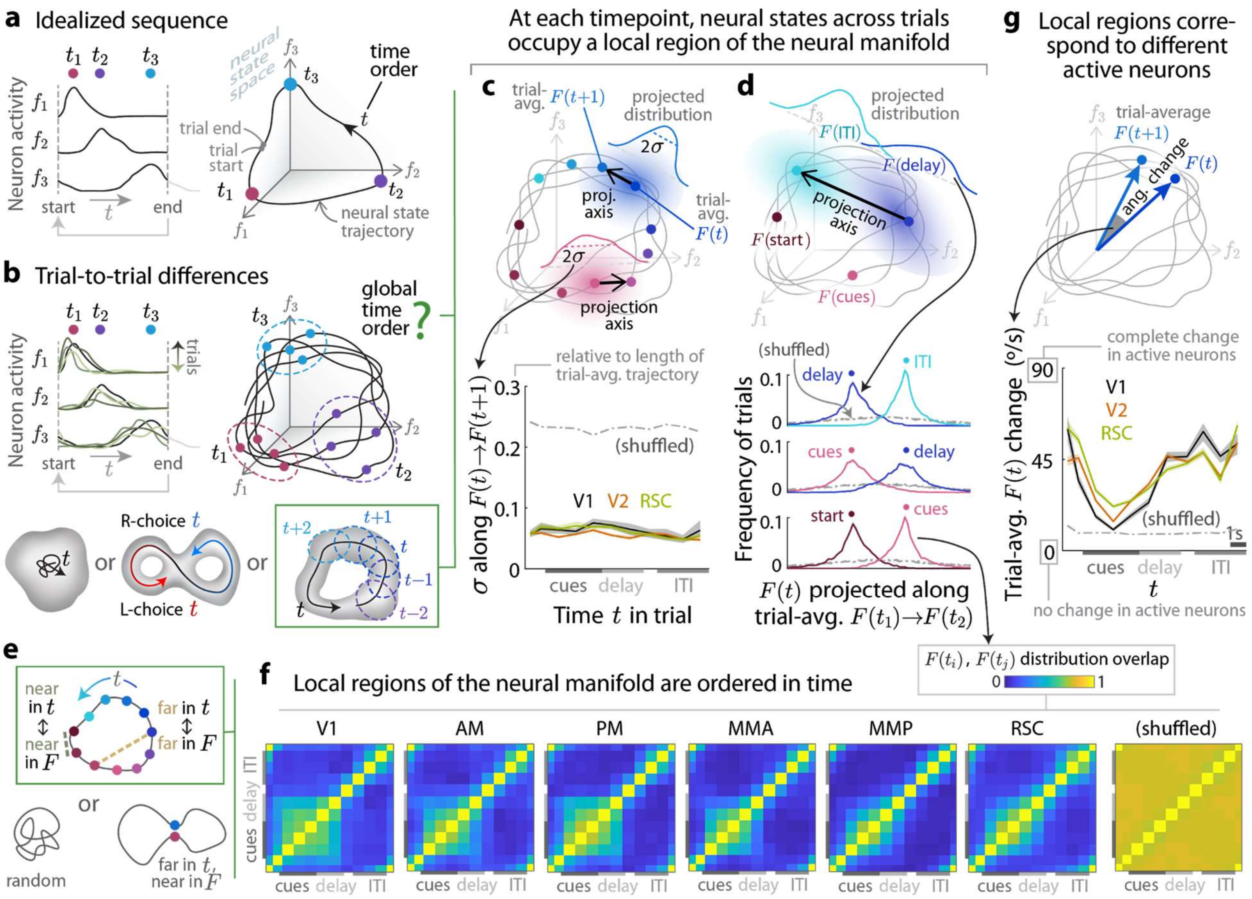


Figure 2. ***A single time coordinate can be used to order local regions of the neural manifold, regardless of trial-to-trial differences.*** **(a)** Illustration of perfectly repeating sequential activity (left), and the corresponding ring-shaped neural manifold (right). **(b)** Illustration of sequential activity with some trial-to-trial differences. For a given time in the trial, the neural data across trials constitutes a cloud of points in the neural state space (dashed ellipses). Point clouds are local if their spreads are small relative to the distance between their centers across time (bottom-right box). This is not true if the neural-state trajectory is a random walk through time (bottom-left). This is also not true if neural responses are non-random but depend strongly on behavioral conditions, leading to highly different neural-state locations at the same timepoint across different trial types (bottom-middle). **(c)** Normalized standard deviation of the neural-state point cloud for a given time t in the trial, projected onto the axis between centers of two adjacent clouds. These centers are the trial-average neural states $F(t)$ and $F(t + 1)$. The normalization factor is the total length of the trial-average trajectory. Shuffled: Pseudo-data with activity shuffled so as to preserve local temporal structure of neural activities, but destroying behavioral and neural population-level correlations. **(d)** Distribution of neural-state point clouds for two distal timepoints, projected onto the axis between the centers of those clouds. In the top illustration, the timepoints are at t_1 in the delay period vs. t_2 in the ITI. Shuffled: As in (c). **(e)** Illustration of how time-order means that regions nearby (far) in time should be nearby (far) in neural-state space (box), vs. alternative possibilities (bottom diagrams). **(f)** Overlap (Methods) between two projected distributions as in (d), for all possible pairs of timepoints. 0 means no overlap, and 1 means that the distributions are identical. Each plot was averaged across imaging sessions for the stated cortical areas. Shuffled: as in (c). **(g)** Angle between the centers of point-clouds at consecutive timepoints, divided by the interval between timepoints. Shuffled: as in (c). Band: S.E.M.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Multiple present- and past-trial task information can be uniquely decoded from all areas/layers

To investigate how neural populations could represent multiple variables, we first established that they contained information about thirteen variables that spanned visual, motor, cognitive, and trial history aspects of the task. For each variable, we trained a different linear decoder for each timepoint in the trial, using data that were z-scored per timepoint (totalling 11 timepoints ~ 1 second apart, see Methods). All thirteen variables could be decoded from all six areas (Fig. 3a-b; cross-validated and corrected for multiple comparisons, see Methods). Choice, reward, and past-trial quantities were most accurately decodable from the more medial areas MMP and RSC (Fig. 3c). In contrast, the more lateral (i.e. visual) areas had higher decoding performance for contralateral than ipsilateral cue counts (Fig. 3d). These inter-area differences are consistent with an anatomical gradient going from a more visual role for V1 towards a more cognitive/memory role for RSC[61], but areas were not sharply delineated by function[9].

We also asked how decoding depended on two aspects of the neural code: (1) time dependence of neural activities; and (2) neuron identities vs. activity levels. For (1), we compared the above per-timepoint decoders to when a single, time-independent decoder was trained per variable, treating all timepoints as if they were additional trials. Time-independent decoders performed moderately worse for most variables (Fig. 3e, Supplementary Fig. 3b), with choice decoding being the least affected. Notably, y location in the T-maze could not be decoded well using a single linear decoder. This is expected from the strongly nonlinear place/time tuning of neurons (Fig. 1e), and is the reason why we separately constructed linear decoding models as a function of time in the trial. For (2), we performed a similar comparison but to per-timepoint decoders based on binarized neural data (neurons “on” if above noise, or “off”). Most variables could still be decoded nearly as well (Fig. 3f, details vs. time in Supplementary Fig. 3b), suggesting that a combinatorial coding scheme[62–65] underlies their decoding performance. Together, these findings allude to a neural code that depended little on the precise activity levels of neurons, but more on their identities out of a time-dependent subset of participating neurons.

Lastly, we address whether decoding truly reflected unique neural information about all thirteen variables. Fig. 3g shows the number of significantly decodable variables to be around ten out of thirteen at any one timepoint. However, others have noted that task variables such as choice and view angle are highly interrelated, and can thus be confounded as explanations of neural activity[56]. To detect this, we consider each decoder as specified by a decoding direction in the neural state space (illustrated in Fig. 3f-left). If there is only neural representation of x_1, but another variable x_2 can be indirectly decoded using information about x_1, then the x_1 and x_2 decoding directions should be collinear. More generally to account for multi-way relationships, we computed the angle of a given decoding direction w.r.t. the subspace spanned by all other decoding directions, and looked for statistically near-zero angles a.k.a. degenerate directions (Methods; all recordings had ≥ 13 neurons). This leave-one-out angle was smallest for choice and view angle decoders

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Fig. 3h). Nevertheless, so long as a variable could be decoded with above-chance performance at a given timepoint, its decoding direction at that timepoint was not degenerate (see Supplementary Fig. 4 for method validations). We thus conclude that the neural activity in all areas contained unique information about all variables as counted in Fig. 3g.

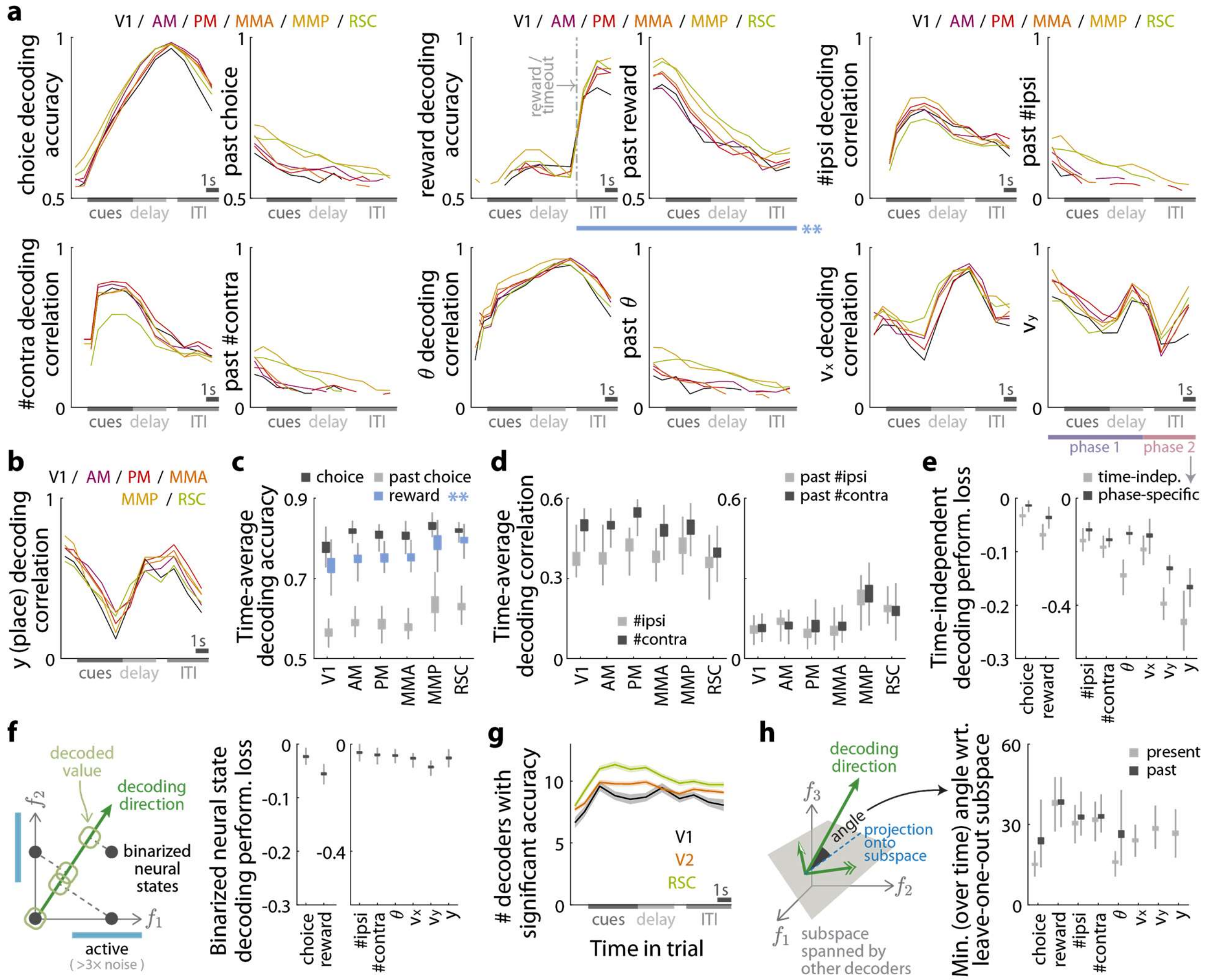


Figure 3. ***Thirteen visual, motor, cognitive, and memory-related variables can be significantly decoded from all areas, with related but non-degenerate directions.*** **(a-b)** Cross-validated performance for decoding thirteen task variables (individual plots), vs. time in the trial. For each variable, a different linear decoder was trained per timepoint. For categorical variables, the performance measure is the classification accuracy. For continuous variables, the performance measure is the correlation between decoded and actual variable values. Performance was averaged across imaging sessions for a given area (lines), with points blanked out if < 25% of sessions had significant decoding performance vs. a permutation test ($p \leq 0.48$ post correction for multiple comparisons; see Methods). **(c-d)** Time-averaged summary of decoding performances in (a), vs. cortical area. As the reward outcome was only known to the mouse at the end of the trial, the reward decoding accuracy was summarized as a time-average including all the indicated timepoints (blue horizontal bar) in (a). Error bars: std. dev. across sessions. Rectangles: Median and S.E.M. **(e)**

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Time-average difference in decoding performance for two alternative decoding methods, relative to the per-timepoint decoders in (a). The "time-independent" decoding method (light gray points) used a single decoder for all timepoints, treating data at different timepoints like additional trials. The "phase-specific" decoding method (dark gray points) used two decoders for two phases of the trial, being all timepoints before/after the turn respectively. For comparability, performance differences are shown on a different scale for categorical (chance level is 0.5 accuracy) vs. continuous variables (chance level is 0 correlation). Error bars: std. dev. across sessions. Rectangles: Median and S.E.M. See Supplementary Fig. 3b-c for performance of all decoders vs. time. **(f)** Same as (e), but for per-timepoint decoders that used neural data that was set to 1 if significantly above noise for a given neuron, and 0 otherwise. Left plot: Illustration of how a combinatorial code of 1-bit neural activities can have high performance when decoding a continuous variable. The projected distance along an optimized decoding direction is the value that is read out by to predict a task variable of interest, and can be sensitive to up to 2^n possible neural states if there are n neurons. **(g)** Number out of thirteen variables that could be significantly decoded at a given timepoint in the trial, with significance defined vs. a permutation test as in (a-b). Lines: Mean across sessions. Bands: S.E.M. **(h)** Angle between a given decoding direction and the subspace spanned by all other decoding directions. As the directions change vs. time in the trial, the minimum angle over time is shown here as a summary. Error bars: std. dev. across sessions. Rectangles: Median and S.E.M.

How decoding directions depend on encoding directions, and potentially also on task-variable correlations

Our decoding studies indicated that most of the neural information about task variables could be extracted from the identities of active neurons, disregarding the detailed activity levels of individual neurons (decoding from binarized neural states in Fig. 3f). This result highlights two ideas. One, as multiple neurons are required to implement an identity-based code (at least for continuous variables, see Fig. 3f-left), the basic unit of neural encoding may involve a cell assembly as proposed by Hebb[66]. Two, although individual neurons can have complex and heterogeneous responses to task variables, such details can be mostly irrelevant when decoding these variables from the collective activities of neurons. Guided by these two ideas, we set out to understand the *effective* structure of the neural code, in the sense of focusing on only the aspect of neural encoding that affects decoding of task variables. We did this by considering how decoding directions depend on the (unknown) neural encoding and (known) statistics of task variables[28,29,67]. We give an intuitive explanation of this relationship below, and refer the interested reader to a more complete derivation in the Methods.

Fig. 4a illustrates how at a fixed timepoint in the trial, or equivalently a local region of the neural manifold, trials with different evidence levels may tend to occupy different neural-state-space locations. In keeping with the above ideas, we hypothesized that what matters is a population-level summary of this neural response: the evidence "encoding direction", defined to be the direction of strongest change in neural state due to a change in evidence levels. In general, there is one encoding direction (per timepoint) for each of the task variables, and together they form an information-coding subspace of the neural-state space. Estimating these encoding directions from data corresponds to solving the system of

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

equations $\mathbf{F} \approx \mathbf{X}\mathbf{W}_{\mathrm{enc}}$, where the rows of $\mathbf{F}$ are neural states at each trial, each column of $\mathbf{X}$ corresponds to (z-scored) values of one task variable across trials, and the rows of $\mathbf{W}_{\mathrm{enc}}$ are the encoding directions (Fig. 4b; Eq. 2). As before, we defined the decoding direction for a particular task variable x as the direction along which the neural state best discriminates x. This corresponds to solving a different but related system of equations $\mathbf{X} \approx \mathbf{F}\mathbf{W}_{\mathrm{dec}}$, where now the columns of $\mathbf{W}_{\mathrm{dec}}$ are the decoding directions (Fig. 4c; Eq. 3). We ask: what are the expected values of these decoding directions, assuming that the brain responds noisily to task variables according to the above linear encoding scheme? The solution is[28,29]: $\mathbf{W}_{\mathrm{dec}} \approx \sigma^{-2}\mathbf{W}_{\mathrm{enc}}^{\top}\mathbf{C}\left(\mathbf{I} + \phi_{\mathrm{enc}}\mathbf{C}\right)^{-1}$ (Eq. 8 in the Methods; see Eq. 7 for correlated noise). Here, σ^{-1} can be thought of as the neural signal-to-noise ratio (SNR), $\mathbf{C} \equiv \mathbf{X}^{\top}\mathbf{X}$ is the task-variable correlation matrix, and $\phi_{\mathrm{enc}} \equiv \sigma^{-2}\mathbf{W}_{\mathrm{enc}}\mathbf{W}_{\mathrm{enc}}^{\top}$ is the matrix of dot products between pairs of encoding directions.

The decoding directions depend on encoding directions in a way that resembles how sensory systems are thought to adapt to environmental stimulus vs. noise statistics[68]. Intuitively, the factors that affect decoding are: (1) unrelated neural fluctuations, which can be mistaken for changes in task-variables and so should be subtracted; and (2) statistical correlations between task-variables, which can be exploited as they contain indirect information. We point out some limiting cases below that show how optimal decoding directions switch from prioritizing (1) to prioritizing (2) depending on the neural SNR, which suggests how we can quantitatively compare this theory to data.

First, whether there is just one encoded variable x_1 (Fig. 4d), or multiple unrelated variables encoded along orthogonal directions (Fig. 4e), the optimal decoding direction for x_1 is parallel to the x_1 encoding direction. However when another variable $x_2$is encoded along a non-orthogonal direction, neural state changes along the x_1 encoding direction can be in part due to x_2. To isolate a purely x_1 signal, the optimal x_1 decoding direction should subtract any x_2 dependence by having a negative component along the x_2 encoding direction (Fig. 4f). Mathematically, this means that at high SNR, $\mathbf{W}_{\mathrm{dec}}$ approximates the pseudo-inverse of $\mathbf{W}_{\mathrm{enc}}$ (Eq. 12). However, this kind of subtraction becomes detrimental if neurons have too much independent noise. Instead, averaging responses across neurons can reduce noise, provided that the responses being averaged are compatible. Mathematically, this means that at low SNR $\mathbf{W}_{\mathrm{dec}} \approx \sigma^{-2}\mathbf{W}_{\mathrm{enc}}^{\top}\mathbf{C}$, i.e. decoding directions are a weighted sum of encoding directions, with weights being the statistical correlation between the encoded variable and the variable to be decoded (Fig. 4g).

Lastly, to understand the relative structure of the neural code, we propose to examine the dot products between pairs of decoding directions, $\phi_{\mathrm{dec}} \equiv \mathbf{W}_{\mathrm{dec}}^{\top}\mathbf{W}_{\mathrm{dec}}$. This has a geometrical interpretation in that the dot product of two vectors $\vec{w}^{(1)}$ and $\vec{w}^{(2)}$ is proportional to the cosine of the angle between these vectors: $\vec{w}^{(1)} \cdot \vec{w}^{(2)} / |\vec{w}^{(1)}||\vec{w}^{(2)}| = \cos\angle(\vec{w}^{(1)}, \vec{w}^{(2)})$. Eq. 9 shows that ϕ_{dec} is only a function of ϕ_{enc}, σ^{-1} and $\mathbf{C}$, i.e. as illustrated in Fig. 4h, the angles between pairs of decoding directions (abbreviated as "decoding angles" $\propto \phi_{\mathrm{dec}}$) are related to the angles between pairs of encoding directions ("encoding angles" $\propto \phi_{\mathrm{enc}}$) in a way that depends on $\sigma^{-1}$and $\mathbf{C}$. With increasing SNR ($\sigma^{-1} \to \infty$) the decoding angles become progressively less dependent on $\mathbf{C}$, eventually reaching the high-SNR limit discussed above where there is no dependence.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

How decoding angles depend on $\mathbf{C}$ can therefore tell us about the unknown neural SNR, and this is what we look into next.

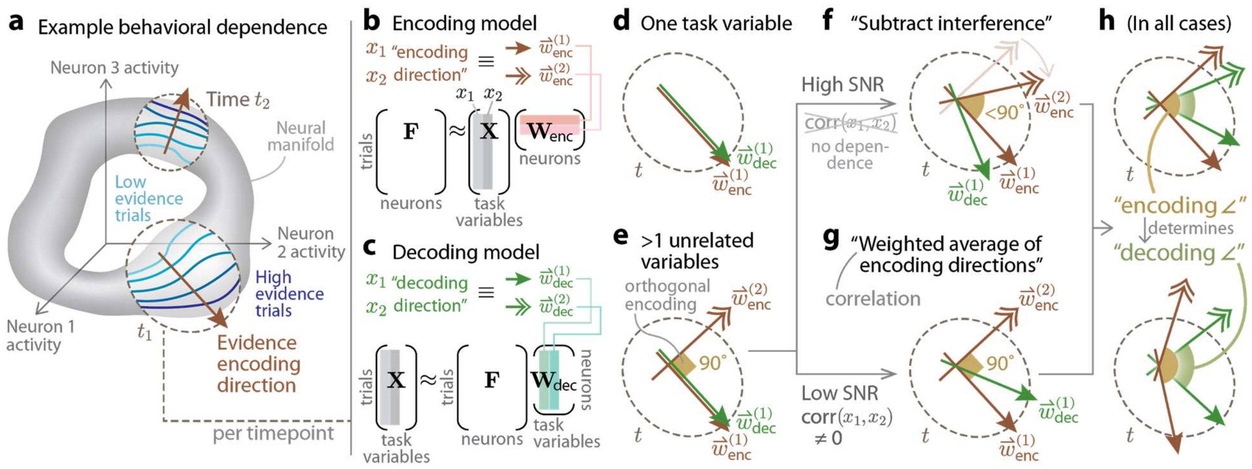


Figure 4. ***At high neural signal-to-noise (SNR), optimal decoding directions cancel out the effects of neural encoding, but at low SNR, optimal decoding directions are a weighted sum of encoding directions.*** **(a)** Illustration of how the neural manifold may have substructure related to a task variable such as evidence. For a fixed time within the trial (insets), the neural state for trials with different evidence levels occupy different sub-regions of the manifold (colored lines), and the gradient is called the evidence encoding direction. **(b)** Illustration of the system of equations used to define encoding directions, at a fixed time in the trial. $\mathbf{F}$ is a trial-by-neuron matrix of neural activities, $\mathbf{X}$ is a trial-by-variable matrix of task variable values, and $\mathbf{W}_{\text{enc}}$ is a variable-by-neuron matrix, the rows of which are the encoding directions. **(c)** Illustration of the system of equations used to define decoding directions. $\mathbf{F}$ and $\mathbf{X}$ are as in (b), and $\mathbf{W}_{\text{dec}}$ is a neuron-by-variable matrix, the columns of which are the decoding directions. **(d)** If there is only one encoded variable x_1, then the decoding direction for x_1 should be parallel to the encoding direction for x_1. **(e)** If another unrelated variable x_2 is encoded along an orthogonal direction, the optimal x_1 decoding direction is still along the x_1 encoding direction. **(f)** If the encoding directions are not orthogonal, but neural noise levels are sufficiently low (high SNR), then the optimal decoding direction for x_1 should aim to subtract fluctuations along the x_1 encoding direction that are driven by x_2, and thus has a negative component along the x_2 encoding direction. **(g)** If instead neural noise levels are high (low SNR), then the optimal x_1 decoding direction should be a weighted sum of the x_1 and x_2 encoding directions. The optimal weight is the correlation of the variable to be decoded (e.g. x_1) to the neurally encoded information. **(h)** Regardless of SNR, the decoding angles are a function of encoding angles and have no explicit dependence on the encoding directions (but see Eq. 7 for when there is correlated neural noise).

Angles between pairs of decoding directions track the correlation level between task variables

In order to not overfit to noise, for all the following analyses we restricted the set of task variables to nine out of thirteen, i.e. dropping past-trial quantities for which there was little-to-no significant neural information in many datasets. For each timepoint in the trial,

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

we computed the matrix of decoding angles for these nine variables, and compared this to the task-variable correlation matrix $\mathbf{C}$ that was computed using the behavioral data across trials. We favored such an angular measure (as opposed to dot products) as it does not explicitly depend on the number nor activity scale of neurons, and can therefore be directly compared across brain areas and animals.

Fig. 5a shows examples of how the angle between two decoding directions matched closely the correlation coefficient for those two decoded variables. Confirming this, the values of these decoding angles were very well-predicted by the first two powers of the task-variable correlation matrix $\mathbf{C}$ (Fig. 5a-insets). In fact, for most pairs of task variables the decoding angles depended mostly on just the first power of $\mathbf{C}$ (see Supplementary Fig. 5a for regression coefficients that show sharply decreasing dependence on higher powers of $\mathbf{C}$, and Supplementary Fig. 5b for goodness-of-fit). This close correspondence was true for all timepoints in the trial (Fig. 5b-d), as well as across brain regions (Fig. 5e), with better correspondence for imaging sessions with more recorded neurons (Fig. 5f). The correspondence also improved with decoding accuracy (Supplementary Fig. 6a), and was preserved under cross-validation, where angles were computed between decoding directions estimated using two different halves of trials (Supplementary Fig. 7a). These checks argue against a noise-induced bias[69] (see Methods for further discussion).

We were surprised by the above dominant and highly precise dependence of decoding angles on $\mathbf{C}$. This matches neither high- nor low-SNR limits of the theory that we have previously discussed without assuming a particular form for the encoding directions, $\mathbf{W}_{\text{enc}}$. However, as explained below, our decoding observations imply that no matter the SNR, there is only one experimentally favored answer for the matrix of encoding angles, i.e. $\phi_{\text{enc}} \propto \mathbf{W}_{\text{enc}}\mathbf{W}_{\text{enc}}^{\top} \propto \mathbf{C}^{-1}$.

To recap, at low SNR the decoding directions are predicted to be $\mathbf{W}_{\text{dec}} \approx \sigma^{-2}\mathbf{W}_{\text{enc}}^{\top}\mathbf{C}$, resulting in decoding angles proportional to $\phi_{\text{dec}} \equiv \mathbf{W}_{\text{dec}}^{\top}\mathbf{W}_{\text{dec}} \propto \mathbf{C}\phi_{\text{enc}}\mathbf{C}$. If the encoding angles are $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$, this cancels out one factor of $\mathbf{C}$ to yield decoding angles $\phi_{\text{dec}} \propto \mathbf{C}$, as observed. On the other hand, at high SNR the decoding directions $\mathbf{W}_{\text{dec}}$ approximates the pseudo-inverse of $\mathbf{W}_{\text{enc}}$, which gives decoding angles $\phi_{\text{dec}} \propto (\mathbf{W}_{\text{enc}}\mathbf{W}_{\text{enc}}^{\top})^{-1} \propto \phi_{\text{enc}}^{-1}$ (Eq. 13). The same choice of $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ means that $\phi_{\text{dec}} \propto (\mathbf{C}^{-1})^{-1} = \mathbf{C}$, which is also as observed. We discussed limiting cases here to provide intuition of how a specific form of encoding angles was strongly implied by the model-agnostic observations of $\phi_{\text{dec}} \propto \mathbf{C}$. The $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ hypothesis can also simply be plugged in to the full theory (Eq. 9) to show that in the general case, this predicts decoding directions $\mathbf{W}_{\text{dec}} \propto \mathbf{W}_{\text{enc}}^{\top}\mathbf{C}$ (from Eq. 8) and decoding angles $\phi_{\text{dec}} \propto \mathbf{C}$ (Eq. 15). Intriguingly, this means that $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ is a rather special encoding scheme that results in optimal decoding directions and angles having no dependence on SNR.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

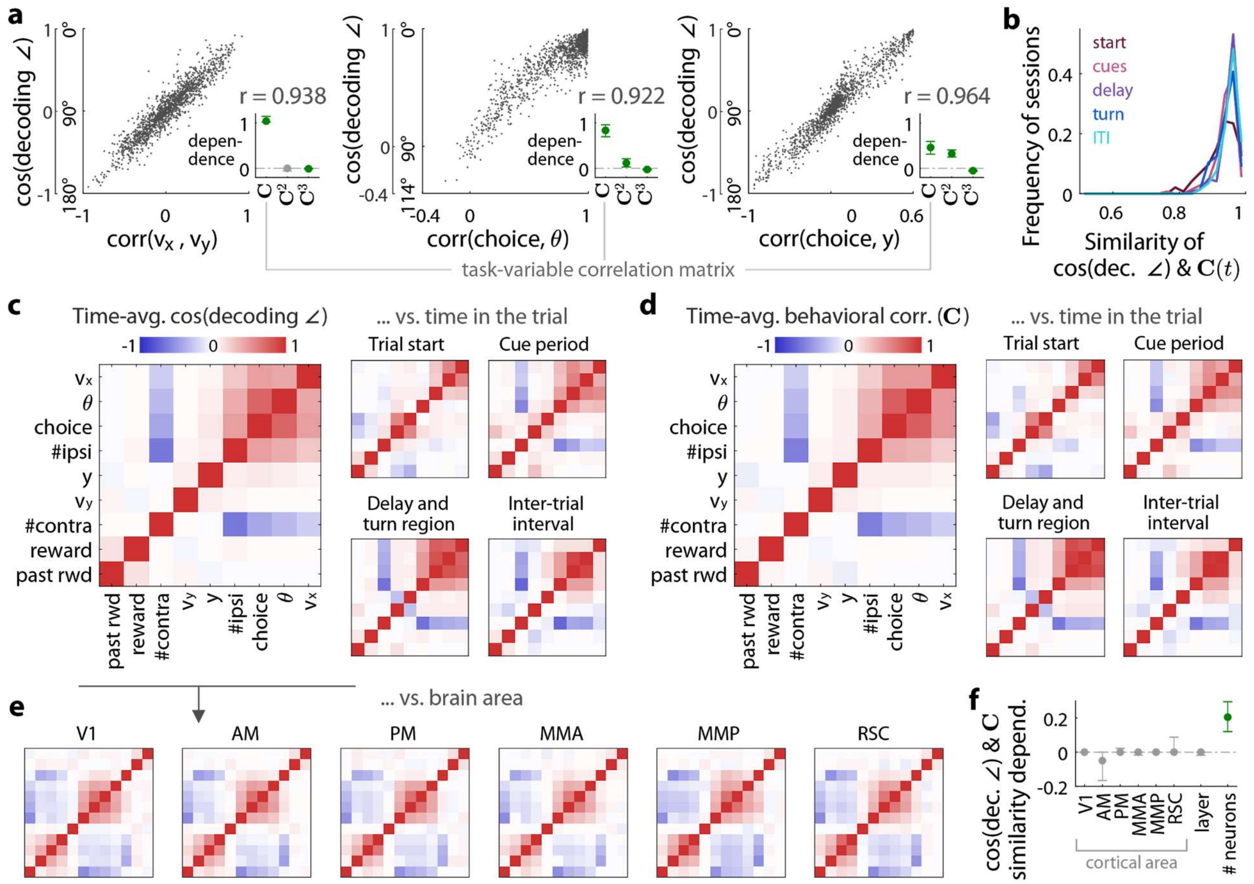


Figure 5. ***Angles between pairs of optimal decoding directions precisely track the time-dependent behavioral correlation between that pair of decoded variables.*** **(a)** Scatter plot of the cosine angle between pairs of decoding directions vs. the behavioral correlation between the decoded pair of task variables. Each data point corresponds to a timepoint within a recording session, all sessions included. Per variable, a different decoder was trained per timepoint, and task-variable correlations were computed using the behavioral data across trials but at that timepoint in the trial. Insets: coefficients from an L1-regularized linear regression model for explaining the dependency of the cosine decoding angles on various powers of the behavioral correlation matrix $\mathbf{C}$. The goodness-of-fit of these models are shown as Pearson's correlation coefficient r. See Supplementary Fig. 5 for all pairs of variables. **(b)** Distribution of similarity scores (Pearson's correlation) for how well the cosine decoding angles matched the behavioral $\mathbf{C}$. One score was computed per imaging session, using as data points all pairs of task variables (i.e. the upper-triangular elements of the matrices in (c) and (d)) and all timepoints within the indicated periods in the trial (colored lines) that had significant decoding performances for both variables ($p < 0.05$ post correction for multiple comparisons). See Supplementary Fig. 7a-b for cross-validation and permutation tests. **(c)** Matrix of cosine angles between all pairs of decoding directions. The order of variables was determined using hierarchical clustering on the time-average behavioral correlation matrix in (d). The left plot shows the time- and session-averaged cosine angles, whereas the four plots on the right were averaged over imaging sessions but for various time periods in the trial. **(d)** Same format as (c) but for the behavioral correlation matrix. **(e)** Time-average cosine decoding angles as in (c), but for datasets in various posterior cortical regions from lateral (V1) to medial (RSC). **(f)** Coefficients from an L1-regularized linear regression model for how strongly the similarity scores in (b) depended on factors like the brain area/layer and the number of recorded neurons. Error bars: 95% C.I. Significant factors are indicated in green.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Angles between pairs of encoding directions approximate a decorrelation operation

Although we arrived at the encoding angles $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ hypothesis in a data-driven way, it is particularly interesting because it corresponds to a neural transformation of behavioral information that "undoes" the correlation $\mathbf{C} \equiv \mathbf{X}^{\top}\mathbf{X}$ between variables. We derive this mathematically in the Methods, and sketch the idea and implications here. Fig. 6a illustrates the joint distribution of two correlated task variables. Assuming that these variables are linearly encoded, $\mathbf{F} = \mathbf{X}\mathbf{W}_{\text{enc}}$, then the signal covariance of the neural data is $\mathbf{F}^{\top}\mathbf{F} = \mathbf{W}_{\text{enc}}^{\top}\mathbf{C}\mathbf{W}_{\text{enc}}$. If there are exactly as many neurons as encoded task variables, we can choose encoding directions to be $\mathbf{W}_{\text{enc}} = \mathbf{W}_{\text{enc}}^{\top} = \mathbf{C}^{-1/2}$, corresponding to encoding angles $\phi_{\text{enc}} \propto \mathbf{W}_{\text{enc}}^{\top}\mathbf{W}_{\text{enc}} = \mathbf{C}^{-1}$. This encoding scheme results in uncorrelated neural activities because then $\mathbf{F}^{\top}\mathbf{F} = \mathbf{C}^{-1/2}\,\mathbf{C}\,\mathbf{C}^{-1/2} = \mathbf{I}$ (Fig. 6b). However, if there are more neurons than encoded variables, the information-coding subspace spanned by encoding directions can be arbitrarily oriented in the neural state space (Fig. 6c). We show in the Methods that only this lower-dimensional subspace of neural modes is relevant to optimal linear decoding of task information, and encoding angles $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ are sufficient to ensure that these neural modes are statistically uncorrelated. Notably, even though the neural modes are uncorrelated, individual neurons still can have a variety of nonzero signal correlations (Fig. 6d). Discovering encoding angles $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ can thus have potential implications on theories of efficient coding and brain function.

To check the above in the experimental data, we fit a set of linear regression models to explain each neuron's activity in terms of the nine task variables (one regression model per timepoint and per neuron, see Methods). These models estimate the encoding directions $\mathbf{W}_{\text{enc}}$, as previously explained (Fig. 4b). About 50% of variance in single-neuron activities was unexplainable (Supplementary Fig. 8a), and the performance of these regression models was moderate (Supplementary Fig. 8b). We thus expected results involving encoding models to be less precise than what we obtained for decoding models. Nevertheless, we found fair agreement between the estimated encoding angles and the $\mathbf{C}^{-1}$ hypothesis (Fig. 6e, Fig. 6f vs. g; see Supplementary Fig. 8c for raw data for pairs of task variables). As for decoding angles, this finding held for all posterior cortical areas (Fig. 6h), and the agreement improved for imaging sessions with more neurons (Fig. 6i). Again arguing against a noise-induced bias, the correspondence between encoding angles and $\mathbf{C}^{-1}$ improved with decoding accuracy (Supplementary Fig. 6b), and was highly positively correlated with the structure of cross-validated encoding angles (Supplementary Fig. 7c).

In sum, both our observations for decoding and encoding angles point to the brain's encoding scheme as one that decorrelates task information. As hypothesized in Fig. 6c-d, Supplementary Fig. 8d-g shows that signal correlations for neural modes than span the encoding subspace had smaller magnitudes and were centered around zero, compared to signal correlations between individual neurons which were large and with a prevalence of positive values (Supplementary Fig. 9b,e).

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

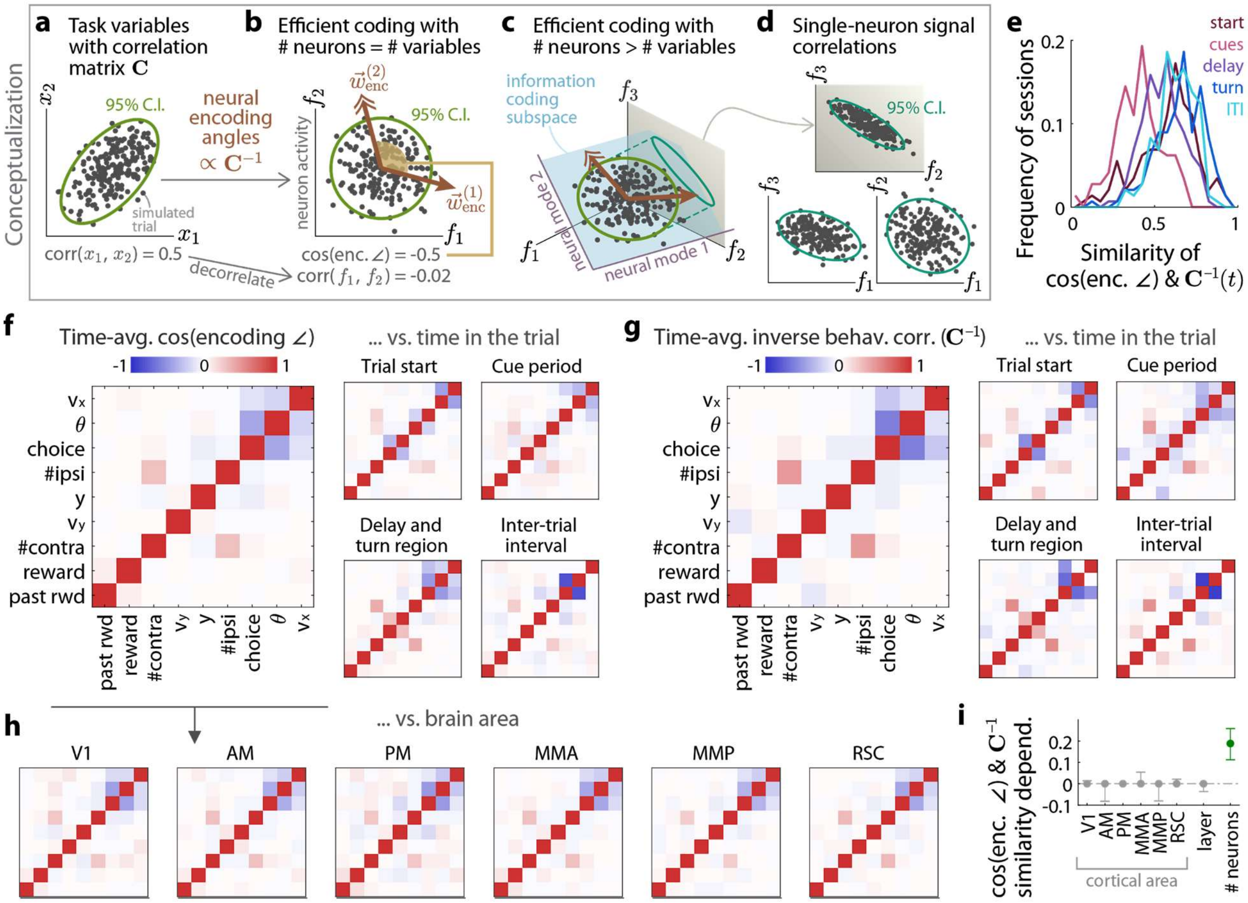


Figure 6. ***Pairwise angles between encoding directions are compatible with a decorrelation operation, suggesting that correlated task variables are represented by uncorrelated neural modes.*** **(a-d)** Illustration of a neural encoding scheme that decorrelates the simulated neural response to correlated task variables, as follows. **(a)** Simulated distribution of two correlated task variables, x_1 and x_2, i.e. the 95% C.I. is an ellipse. **(b)** Simulated responses of two neurons that (noiselessly) encode the task variables in (a), with angle between encoding directions (brown arrows) proportional to $\mathbf{C}^{-1}$. The neural activities f_1 and f_2 are statistically uncorrelated, i.e. the 95% C.I. is a circle. **(c)** As in (b), but with three neurons encoding the two task variables in (a). The neural activities lie within a 2-dimensional information coding subspace (blue plane) spanned by the encoding directions (brown arrows), and the neural modes that define this subspace are uncorrelated (light green 95% C.I. is a circle). **(d)** The same simulated data in (c), but plotted for various pairs of neural axes. These pairs of neurons have nonzero signal correlations (95% C.I. are ellipses). **(e)** Distribution of similarity scores (Pearson's correlation) for how well the cosine encoding angles matched the cosine angles between columns of the inverse behavioral correlation matrix, $\mathbf{C}^{-1}$. One score was computed per imaging session, using as data points all pairs of task variables (i.e. the upper-triangular elements of the matrices in (f) and (g)) and all timepoints within the indicated periods in the trial (colored lines) that had significant decoding performances for both variables ($p < 0.05$ post correction for multiple comparisons). See Supplementary Fig. 7c-d for cross-validation and permutation tests. **(f)** Matrix of cosine angles between all possible pairs of encoding directions, with the same order of variables as in Fig. 5c. The left plot shows the time- and session-averaged cosine angles, whereas the four plots on the right were averaged over imaging sessions but for various time periods in the trial. **(g)** Same format as (f) but for cosine angles between columns of the inverse behavioral correlation matrix. **(h)** Time-average cosine encoding

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

angles as in (f), but for datasets in various posterior cortical regions from lateral (V1) to medial (RSC). **(i)** Coefficients from an L1-regularized linear regression model for how strongly the similarity scores in (e) depended on factors like the brain area/layer and the number of recorded neurons. Error bars: 95% C.I. Significant factors are indicated in green.

Angles between encoding directions followed slow changes in inverse task-variable correlations despite rapid changes in encoding directions

So far we have performed all analyses independently per timepoint in the trial, but which findings are actually time-dependent and which are not? First, the scale of some task variables depended on time in the trial, and the corresponding encoding weights of neurons across the population *inversely* followed this scale[70] (Supplementary Fig. 10a; this motivated our use of z-scored variables in all analyses). This can be interpreted as the neural-population encoding performing more than decorrelation because it also transforms the time-dependent variance of variables to 1 (called a "whitening" operation). Second, correlations between task variables changed slowly as a function of time in the trial, and the encoding angles tracked the corresponding time-variations in the inverse task-variable correlation matrix $\mathbf{C}^{-1}(t)$ fairly well (Fig. 7a). Third, these changes in angles *between* encoding directions were small compared to changes in the encoding directions themselves, which was $> 50°/s$ at all timepoints (Fig. 7b). Supplementary Fig. 11f shows that for an example pair of task variables (choice and view angle θ), the rate of change of encoding angles was always about 5 times smaller (68% C.I.) than the rate of change of encoding directions. The same holds for all pairs of task variables (Fig. 7c; 68% C.I. rate of change of encoding angles $>$ 4 times smaller than the rate of change of encoding directions).

How can a neural population implement slow changes in encoding angles despite fast time-variations in encoding directions? We hypothesize that this can happen when neural activities have factorizable task-variable vs. time dependencies. As illustrated in Fig. 7d, the activity of each neuron i is hypothesized to have the form $f_i(t) = \mu_i(t) + g_i(t)\,\vec{u}_i \cdot \vec{x}$. In this multiplicative time-modulation model, $\vec{u}_i$ are time-independent encoding weights for task variables $\vec{x}$, which are multiplied by a characteristic time-modulation function $g_i(t)$ for that neuron; $\mu_i(t)$ is the mean activity across trials, which does not depend on task variables. Fig. 7e shows that this simple model predicted single-neuron activities almost as well as the per-timepoint encoding models used throughout this article, where encoding weights can all be different at every timepoint, $f_i(t) = \mu_i(t) + \vec{w}_i(t) \cdot \vec{x}$. Moreover, it can explain our observations of sequential neural activity and time-varying encoding directions, because the set of active neurons and therefore nonzero coordinates of the encoding directions change with time (Supplementary Fig. 10b-c shows that a model with additive time-modulations but no time-dependence of task variable responses does not explain time-varying encoding directions, and also fits more poorly to the neural data). If the time-variations of (say) two encoding directions are *not* somehow synchronized, then the angle between these directions could exhibit time-variations caused by---and therefore ballpark as fast as---the changes in encoding directions. Intriguingly, we instead found that the

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

encoding angles tracked the slowly changing $\mathbf{C}^{-1}(t)$ (Fig. 7a), and this change was much slower than the change in encoding directions (Fig. 7c, Supplementary Fig. 11f).

We can understand the stability of encoding angles by starting from how each neuron was well-characterized, in the multiplicative model, by time-independent encoding weights $\vec{u}_i$ (Fig. 7d). Across the neural population (Fig. 8a), these weights can be thought of as underlying set of time-independent encoding directions $\mathbf{U}_{\text{enc}}$ in the high-dimensional neural state space (Fig. 8b-top). We refer to the angles between these directions as given by $\psi_{\text{enc}} \equiv \mathbf{U}_{\text{enc}}\mathbf{U}_{\text{enc}}^{\top}$. As illustrated in Fig. 8b-bottom, at every timepoint the observed encoding directions $\mathbf{W}_{\text{enc}}(t)$ are approximately a projection of this underlying encoding structure onto a low-dimensional subspace of active neurons. If the underlying $\mathbf{U}_{\text{enc}}$ directions are randomly oriented in the neural state space, the observed encoding angles $\phi_{\text{enc}}(t)$ can approximate ψ_{enc} because random projections in a high-dimensional space are likely to preserve relative distances between points (a constructive proof of the Johnson-Lindenstrauss theorem[71]), and thus also angles (Supplementary Fig. 12a). We show via simulations in Supplementary Fig. 12b-g that for sufficiently large populations of sequentially active neurons, this $\phi_{\text{enc}}(t) \approx \psi_{\text{enc}}$ approximation becomes highly precise regardless of the exact shape of each neuron's time-modulation function. Sequential dynamics can therefore be functionally equivalent to each neuron having a binary time-modulation function that simply determines whether it is "on" or "off" (Fig. 8c), and where "on" means that the neuron has an effectively time-independent task-variable response.

Lastly, the observed slow time-variations in $\phi_{\text{enc}}(t)$ are still consistent with a constant $\mathbf{U}_{\text{enc}}$. This is because different subsets of neurons are active at different timepoints in the trial, so a different subset of columns (neurons) of $\mathbf{U}_{\text{enc}}$ contribute to ϕ_{enc} at the start of the trial than, say, the end of the trial. If we order the columns of $\mathbf{U}_{\text{enc}}$ by the preferred activation times of the corresponding neurons, then relationships between rows (weights for task variables) that differ systematically across columns (time-ordered neurons) can be observed as a time-varying $\phi_{\text{enc}}(t)$. Supplementary Fig. 12h-k describes a simulation where we constructed an example of such a systematic structure in $\mathbf{U}_{\text{enc}}$, and shows that again for large, sequentially active neural populations, the "observed" $\phi_{\text{enc}}(t)$ converges to the designed truth even with random time-modulations per neuron. Geometrically, this kind of systematic structure means that $\mathbf{U}_{\text{enc}}$ is not fully randomly oriented in the high-dimensional neural state space (see Supplementary Fig. 12l for a geometrical explanation). In this way, large sequentially active neural populations can use changes in encoding weights *across* neurons to allow $\phi_{\text{enc}}(t)$ to stably follow $\mathbf{C}^{-1}(t)$ even with constant encoding weights per neuron.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

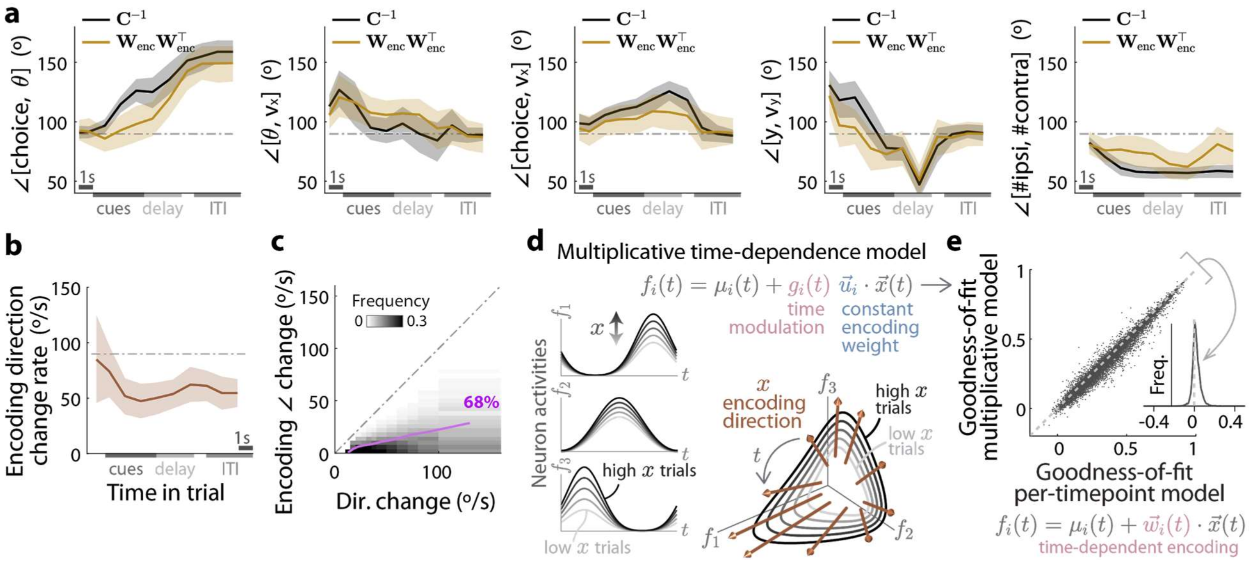


Figure 7. ***Encoding angles follow slow time-variations in inverse task-variable correlations, despite rapid changes in encoding directions.*** **(a)** For a given pair of task variables (individual plots), the value of the corresponding entry in the inverse task-correlation matrix $\mathbf{C}^{-1}$ as a function of time (light brown line), compared to the encoding angle vs. time (black line). Lines: mean across imaging sessions. Band: 68% C.I. **(b)** Angular difference in encoding directions between two consecutive timepoints, divided by time interval, and as a function of time in the trial. All variables had a similar timecourse (data not shown) and were pooled together for simplicity. Lines: mean across imaging sessions and task variables. Band: 68% C.I. **(c)** Rate of change of the angle between encoding angles (absolute value), vs. the rate of change of encoding directions (cf. (b); each x-coordinate is the average of the change in x_1 and the change in x_2 encoding directions, where x_1 and x_2 are the variables for which the encoding angle was calculated). Each data point corresponds to one pair of task variables and one timepoint in an imaging session, i.e. pooling data for all variables. See Supplementary Fig. 11f for example of one pair of task variables (choice and view angle). Lines: 68% C.I. of encoding-angle change (y-coordinate), calculated in bins of the direction change rate. This 68% C.I. rate of encoding angle changes were > 4 times less than the rate of direction changes. **(d)** Three simulated neurons with linear dependencies on a task variable x, multiplied by nonlinear time-modulation functions. The encoding directions changed in time to point towards the more active and thus more behaviorally responsive neuron at that time (right plot). **(e)** Cross-validated goodness-of-fit scores for the multiplicative time-dependence model in (d) vs. a model with fully flexible, time-dependent task-variable encoding weights. Each point is a score for a single neuron, being Pearson's correlation of the predicted vs. actual neural activity. Inset: distribution of differences in scores for the per-timepoint vs. multiplicative models. The per-timepoint model performed statistically better ($p = 4.6 \times 10^{-10}$, Wilcoxon rank-sum test), but the two scores were close. See Supplementary Fig. 10b-c for how a model with additive time-modulations and *no* time-dependence of task variable responses performed qualitatively worse than the multiplicatively time-modulated response model.

Discussion

In this work, we described some geometrical structures of neural-population activity across posterior cortical areas as mice performed a complex, dynamic task. How were neural representations of the many task-related variables organized relative to each other and maintained/updated through time? We answered in three parts. First, neurons were

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequentially active vs. place/time in the trial (Fig. 1e), and in fact had time preferences independent of behavioral factors (Fig. 7d-e), corresponding to a neural manifold with global time order (Fig. 2). Second, thirteen visual, motor, cognitive, and memory-related task variables could be uniquely decoded from all areas/layers (Fig. 3), with some anatomical differences but no evident areal specialization[9]. Third, encoding directions varied rapidly across time (Fig. 7b), yet in all areas/layers and at all timepoints, the angles between pairs of decoding/encoding directions ("decoding/encoding angles") respectively resembled the correlation (Fig. 5) and inverse correlation (Fig. 6, Fig. 7a) matrices of the task variables. This supports the hypothesis that the brain's encoding scheme performs decorrelation of the correlated task information. Below, we discuss some implications of our findings in regards to the three questions posed in the Introduction: on how neural populations simultaneously encode multiple variables, consequences of the encoding scheme on decoding, and how this neural code is dynamically coordinated and represents temporal context.

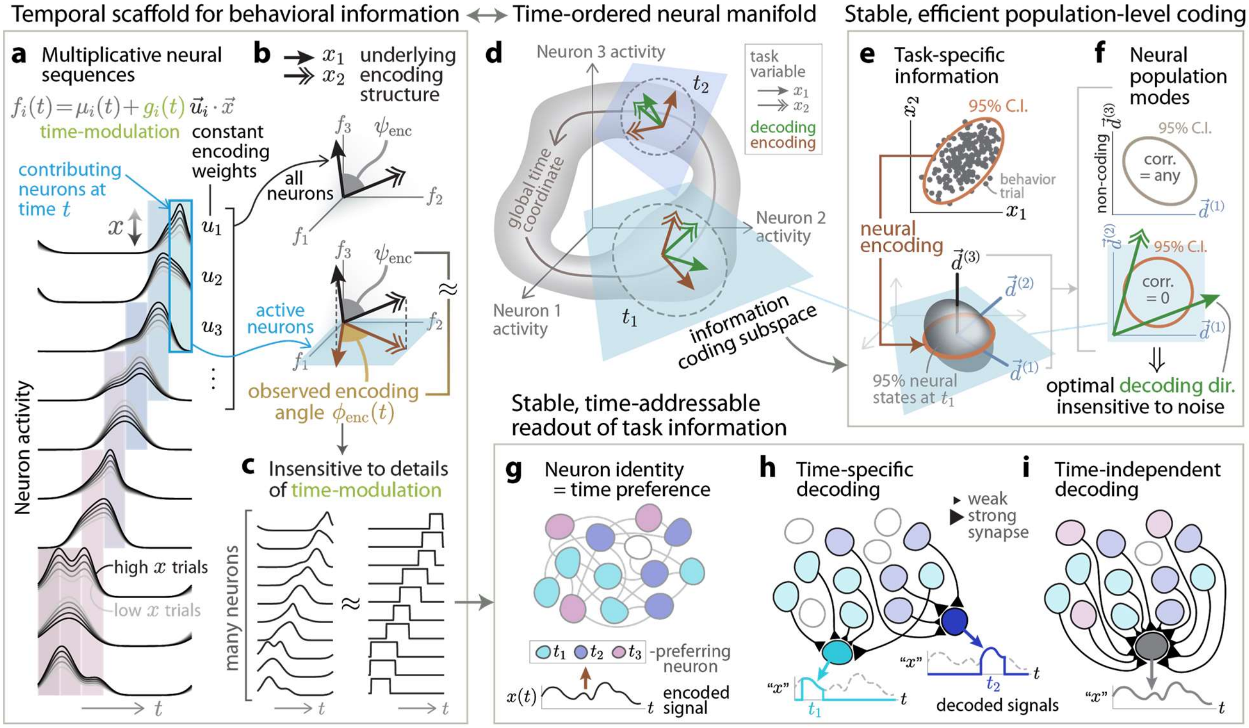

Temporal scaffold for behavioral information ↔ Time-ordered neural manifold
Stable, efficient population-level coding
a Multiplicative neural sequences
$f_i(t) = \mu_i(t) + g_i(t)\,\vec{u}_i \cdot \vec{x}$
time-modulation
constant encoding weights
contributing neurons at time t
x
u_1
u_2
u_3
high x trials
low x trials
Neuron activity
t
b x_1 x_2 underlying encoding structure
f_3 ψ_{enc}
all neurons
f_1 f_2
active neurons
observed encoding angle $\phi_{\text{enc}}(t)$
≈
c Insensitive to details of time-modulation
many neurons
≈
t
d Neuron 3 activity
task variable
x_1 x_2
decoding
encoding
t_2
global time coordinate
Neuron 2 activity
Neuron 1 activity
t_1
information coding subspace
e Task-specific information
x_2
95% C.I.
behavior trial
x_1
neural encoding
$\vec{d}^{(3)}$
$\vec{d}^{(2)}$
$\vec{d}^{(1)}$
95% neural states at t_1
f Neural population modes
non-coding $\vec{d}^{(3)}$
95% C.I.
corr. = any
$\vec{d}^{(1)}$
$\vec{d}^{(2)}$
95% C.I.
corr. = 0
$\vec{d}^{(1)}$
⇓
optimal decoding dir. insensitive to noise
Stable, time-addressable readout of task information
g Neuron identity = time preference
t_1 t_2 t_3 -preferring neuron
$x(t)$ encoded signal
t
h Time-specific decoding
weak
strong
synapse
"x"
t_1
t_2
t
decoded signals
i Time-independent decoding
"x"
t

Figure 8. ***Conceptual summary: multiplicative neural sequences implement stable and efficient neural-population coding of task-specific variables*.** **(a)** Neurons were sequentially active, with approximately constant task-variable encoding weights $\vec{u}_i$ multiplied by behavior-independent time-modulation functions $g_i(t)$. Vertical colored bands indicate neurons that contribute significantly to the x encoding direction. **(b)** At each timepoint, the observed encoding angles approximates a projection of a hypothesized underlying encoding structure ψ_{enc} onto a low-dimensional subspace of active neurons. The underlying encoding directions (black arrows) correspond to the set of constant encoding weights in (a), and ψ_{enc} is the matrix of dot products between these directions. In high dimensions (many neurons), the observed encoding angle $\phi_{\text{enc}}(t)$ is likely to be nearly equal to the underlying ψ_{enc}. **(c)** With many sequentially active neurons, the projection effect in (b) can be insensitive to details of single-neuron time-modulation functions (left traces), i.e. these time-modulations can be functionally equivalent to simple on/off functions (right

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

traces). See simulations in Supplementary Fig. 12. **(d)** Local regions of the neural manifold could be ordered by time in the trial, and at each timepoint/local region there was a different information-coding subspace spanned by encoding directions (brown arrows). Theoretically, optimal decoding directions (green arrows) should lie within this information-coding subspace. **(e)** At each timepoint, task variables were correlated across trials (top plot), and part of the variance in neural states could be explained as dependence on these task variables (bottom plot). **(f)** The distribution of neural states in (e) was approximately uncorrelated within the information-coding subspace (bottom plot), but can have arbitrary statistics along non-coding directions orthogonal to this subspace (top plot). Here "non-coding" directions refer to how optimal decoding directions should have no components in these directions. **(g)** Since the time-modulations of neurons in (a) do not depend on behavioral factors, neuron identities can be used to select for neurons with particular time preferences. **(h)** Neuron identities in (g) can be used to selectively and stably read out task information at specific timepoints in the trial, i.e. a simple weighted sum of upstream neural activities (static synapses) can produce a decoded signal that is undistorted by time-modulations of neurons. **(i)** The union of synaptic weights for the time-specific readouts in (h) can be used to read out task information stably through time.

How *should* the brain encode/decode information? Theories of efficient coding propose that to minimize resource usage, an efficient code should utilize statistically independent neural representations[34,35]. Our results support these theories, but with three intriguing distinctions. One, we did not observe that individual neurons have statistically independent responses (Supplementary Fig. 9), so in a strict sense our findings differ from being a fully efficient code, as well as the focus of a large body of related work[36–44,68]. Rather, we argued that what is relevant for decoding is an information-coding subspace spanned by encoding directions (Fig. 6c, Fig. 8d). The encoding angles that characterized this subspace approximated the inverse task-variable correlation matrix $\mathbf{C}^{-1}$ (Fig. 6e). We showed mathematically that finding encoding angles $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ implies that the neural representation utilizes uncorrelated modes within the information-coding subspace (Fig. 8e-f). This information-coding subspace can be arbitrarily oriented in a high-dimensional neural state space, leading to signal correlations between individual neurons (Fig. 6d; Supplementary Fig. 9b,e) even when the information-coding modes have much reduced signal correlations (Supplementary Fig. 8e,g). There can also be many non-coding modes (i.e. that do not affect optimal linear decoding) with arbitrary correlations (Fig. 8f-top). Two, even though we started with a theoretical formulation of how optimal decoding directions should change to account for neural noise levels[28,29], the theory illuminated that our experimental observation of $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ falls into a special case where decoding directions do not depend on neural noise levels (assuming that noise correlations are small, see Supplementary Fig. 9c,f). Speculatively, the brain could utilize a $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ encoding scheme to ensure that an optimized readout circuitry would remain optimal even if neural noise levels were to fluctuate. Three, we report observing $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ for a mix of external and internally computed variables, which were (only) interrelated through a learned behavioral task. One possibility is that all our findings here arose from a learning process that optimized computational utility of the neural code, more so than resource-based criteria. For example, it has previously been pointed out that removing expected statistical

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

regularities from the neural representation permits easy detection of unexpected associations, as is relevant for survival[26].

As emphasized in the design of our task, another major computational challenge that brains face is that the information they process is not static, but changes in time depending on the environment and the animal's own behavior. Despite the dynamic and nonlinear nature of neural responses, we could accurately decode almost all task variables with time-independent, linear decoders (Fig. 3e, Supplementary Fig. 3b). Except for y location in the T-maze, for which the neural representation seemed highly nonlinear (cf. sequential place/time preferences of neurons in Fig. 1e), Supplementary Fig. 3c shows that the performance of decoders trained on seconds-long temporal phases of the task reached levels comparable to per-timepoint decoders. This is compatible with previous reports of long-timescale structure in neural state transitions in posterior parietal cortex[72] and long-timescale order in single-trial neural state trajectories in premotor cortex[73]. Our work extends these studies by discovering that across posterior cortices, it was specifically (but not necessarily only) the information encoding geometry that was stable, up to tracking slow changes in task variable correlations in a way that suggested a decorrelation function throughout time in the trial. Strikingly, this efficient-coding-like function did not seem to be implemented in a static way by dedicated neurons, but instead involved a sequence of different neurons across time. These sequential activity patterns corresponded to neural manifolds with a ring-like geometry[74], and could be related to previous reports of low-dimensional manifolds corresponding to rotational neural dynamics[75]. We introduced a quantitative measure for the notion of sequentiality at the neural-population level, i.e. the rate of angular change in the neural state. This measure showed that the set of active neurons changed continuously with time in the trial, albeit most slowly in the cue period (particularly for V1 but less so for RSC, see Fig. 2f and Supplementary Fig. 2c).

Our observations of sequential dynamics in all areas are amongst a growing number of such for the mouse neocortex[53–58]. These phenomena are reminiscent of place[76] or time[77–79] cells in the hippocampus, which are also known to jointly encode a variety of other spatial and nonspatial factors. An interesting idea that has arisen in that subfield concerns how sequential activity could act as a temporal scaffold upon which arbitrary information can be imprinted, i.e. multiplexing this information with timestamps to indicate *when* they occurred[77,80–82]. How can such multiplexing be designed so that information can be read out without confounding the timestamp with the imprinted information? We point out a simple design inspired by what we call "multiplicative neural sequences" in our data (Fig. 8a), where each neuron's response to task variables was well described as a product of two functions, $w(x)\ g(t)$. $w(x)$ is a behavioral response function that depends on task variables x but does not depend on time, and $g(t)$ is a time-modulation function that depends only on time t in the trial. In other words, the nominally high-dimensional neural population activity of all surveyed cortical areas can be parsimoniously described by a low-dimensional set of multiplicative factors. This type of factorizable neural responses seems intriguingly ubiquitous, as similar findings have been reported for mouse prefrontal cortex and nonhuman-primate motor cortex[83]. As each neuron has a characteristic time-preference (Fig. 8g), computing a weighted sum of activities of neurons with time preferences around t allows a readout circuit to detect that a behavioral signal has

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

occurred specifically at time around t (Fig. 8h). What is particularly interesting is that for large, sequentially active neural populations, such a time-specific readout (with static synapses) can be stable in the sense of having no further dependence on neural time-modulations, even if these time-modulations are randomly constructed. We hypothesized that this is because the neural time-modulations effectively project $w(x)$ onto a low-dimensional subspace corresponding to the active neurons (Fig. 8b), and this kind of projection is likely to preserve geometrical properties of $w(x)$ according to the Johnson-Lindenstrauss theorem[71] (and extensions[84]). We showed via simulations (Supplementary Fig. 12) that neural time-modulations can be quite arbitrary and still be functionally equivalent to simple on/off functions (Fig. 8c). As each neuron has effectively time-independent behavioral responses when "on", task information can also be decoded in a time-independent way by reading out neurons with a range of time preferences (Fig. 8i), compatible with the high performance we observed with time-independent decoders (Fig. 3e, Supplementary Fig. 3b,c). The multiplicative neural sequences observed in our data thus suggest a form of time-behavior multiplexing that enables simple, stable readout of time-specific behavioral information. Furthermore, by having systematic differences in encoding weights across neurons with different time preferences, the neural population can implement a decorrelation operation that tracks changes in behavioral correlations across time in the trial (Fig. 7). We suggest that the above computational properties of multiplicative neural sequences underlies efficient coding by neural modes across posterior cortex, and we propose that they could in general be a useful design principle for temporal scaffolds.

Online Methods

Animals

All procedures were approved by the Institutional Animal Care and Use Committee at Princeton University and were performed in accordance with the Guide for the Care and Use of Laboratory Animals[85]. We used 11 mice aged 2-16 months of both genders, and from two transgenic strains that express the calcium-sensitive fluorescent indicator GCamp6f[86] in excitatory neurons of the neocortex. 6 mice were of the Thy1-GP5.3[87] strain (Jackson Laboratories, stock #028280), and 5 were crosses of the Ai93-D;CaMKIIα-tTA[88] and Emx1-IRES-Cre[89] strains (Jackson Laboratories, stocks #024108 and #005628). All the data analyzed in this work were from fully-trained mice as described in the following sections.

Surgery

Young adult mice (2-3 months of age) underwent aseptic stereotaxic surgery to implant an optical cranial window and a custom lightweight titanium headplate under isoflurane anesthesia (2.5% for induction, 1-1.5% for maintenance). Mice received one pre-operative dose of meloxicam subcutaneously for analgesia (1 mg/kg) and another one 24 h later, as well as peri-operative intraperitoneal injection of sterile saline (0.5cc, body-temperature) and dexamethasone (2–5 mg/kg). Body temperature was maintained throughout the procedure using a homeothermic control system (Harvard Apparatus). After asepsis, the

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

skull was exposed and the periosteum removed using sterile cotton swabs. A 5mm diameter craniotomy approximately centered over the parietal bone was made using a pneumatic drill. The cranial window implant consisted of a 5mm diameter round #1 thickness glass coverslip bonded to a steel ring (0.5mm thickness, 5mm diameter) using a UV-curing optical adhesive. The steel ring was glued to the skull with cyanoacrylate adhesive. Lastly, a titanium headplate was attached to the cranium using dental cement (Metabond, Parkell).

Behavioral task

After at least three days of post-operative recovery, mice were started on water restriction and the Accumulating-Towers training protocol[11], summarized here. Mice received 1-2mL of water per day, or more in case of clinical signs of dehydration or body mass falling below 80% of the pre-operative value. Behavioral training started with mice being head-fixed on an 8-inch Styrofoam® ball suspended by compressed air, and ball movements were measured with optical flow sensors. The VR environment was projected onto a custom-built Styrofoam® toroidal screen and the virtual environment was generated by a computer running the Matlab (Mathworks) based software ViRMEn[90], plus custom code.

For historical reasons, 3 out of 11 mice were trained on mazes that were longer (30cm pre-cue region + 250cm cue region + 100-150cm delay region) than the rest of the cohort (30cm pre-cue region + 200cm cue region + 100cm delay region). In VR, as the mouse navigated down the stem of the maze, tall, high-contrast visual cues appeared along either wall of the cue region when the mouse arrived within 10cm of a predetermined cue location. These locations were drawn randomly per trial according to a spatial Poisson process with 12cm refractory period between consecutive cues on the same wall side. Cues were made to disappear after 200ms. The mean number of majority:minority cues was 8.5:2.5 for the 250cm cue region maze and 7.7:2.3 for the 200cm cue region maze. Mice were rewarded with $\geq 4\mu L$ of a sweet liquid reward (10% diluted condensed milk, or 15% sucrose) for turning down the arm on the side with the majority number of cues. Correct trials were followed by a 3s-long inter-trial-interval (ITI), whereas error trials were followed by a loud sound and an additional 9s time-out period. To discourage a tendency of mice to systematically turn to one side, we used a de-biasing algorithm that adjusts the probabilities of sampling right- vs. left-rewarded trials[11]. Per session, we computed the percent of correct choices using a sliding window of 100 trials and included the dataset for analysis if the maximum performance was $\geq$ 65%.

Functional identification of visual areas

We adapted methods[51,91,92] to functionally delineate the primary and secondary visual areas using widefield imaging of calcium activity paired with presentation of retinotopic stimuli to awake and passively running mice. We used custom-built, tandem-lens widefield macroscopes consisting of a back-to-back objective system[93] connected through a filter box holding a dichroic mirror and emission filter. One-photon excitation was provided using a blue (470nm) LED (Luxeon star) and the returning green fluorescence was bandpass-filtered at 525 nm (Semrock) before reaching a sCMOS camera (Qimaging, or Hamamatsu). The LED delivered about 2-2.5mW/cm^2 of power at the focal plane, while the camera was

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

configured for 20-30Hz frame rate and about 5-10μm spatial resolution. Visual stimuli were displayed on either a 32" AMVA LED monitor (BenQ BL3200PT), or the same custom Styrofoam® toroidal screen as for the VR rigs. The screens were placed to span most of the visual hemifield on the side contralateral to the mouse's optical window implant. The space between the headplate and the objective was covered using a custom made cone of opaque material.

The software used to generate the retinotopic stimuli and coordinate the stimulus with the widefield imaging acquisition was a customized version of the ISI package[94] and utilized the Psychophysics Toolbox[95]. Mice were presented[51] with a 20° wide bar with a full-contrast checkerboard texture (25° squares) that inverted in polarity at 12 Hz, and drifted slowly (9°/s) across the extent of the screen in either of four cardinal directions. Each sweep direction was repeated 15 times, totaling four consecutive blocks with a pause in between. Retinotopic maps were computed similarly to previous work[92] with some customization that improved the robustness of the algorithms for preparations with low signal-to-noise ratios (SNR). Boundaries between the primary and secondary visual areas were detected using a gradient-inversion-based algorithm[91], again with some changes to improve stability for a diverse range of SNR.

Two-photon imaging during VR-based behavior

The virtual reality plus two-photon scanning microscopy rig used in these experiments follow a previous design[50]. The microscope was designed to minimally obscure the $\sim 270°$ horizontal and $\sim 80°$ vertical span of the toroidal VR screen, and also to isolate the collection of fluorescence photons from the brain from the VR visual display. Two-photon illumination was provided by a Ti:Sapphire laser (Chameleon Vision II, Coherent) operating at 920nm wavelength, and fluorescence signals were acquired using a 40x 0.8 NA objective (Nikon) and GaAsP PMTs (Hamamatsu) after passing through a bandpass filter (542/50, Semrock). The amount of laser power at the objective used ranged from ~40-150mW. The region between the base of the objective lens and the headplate was shielded from external sources of light using a black rubber tube. Horizontal scans of the laser were performed using a resonant galvanometer (Thorlabs), resulting in a frame acquisition rate of 30Hz and configured for a field of view of approximately $500 \times 500\mu m$ in size. Microscope control and image acquisition were performed using the ScanImage software[96]. Data related to the VR-based behavior were recorded using custom Matlab-based software embedded in the ViRMEn engine loop, and synchronized with the fluorescence imaging frames using the I2C digital serial bus communication capabilities of ScanImage. A single field of view at a fixed cortical depth and location relative to the functional visual area maps was continuously imaged throughout the 1-1.5 hour behavioral session. The vasculature pattern at the surface of the brain was used to locate a two-photon imaging field of view (FOV) of interest.

Identification of putative neurons

All imaging data were first corrected for rigid brain motion by using the Open Source Computer Vision (OpenCV) software library function $cv::matchTemplate$. Fluorescence timecourses corresponding to individual neurons were then extracted using a deconvolution and demixing procedure that utilizes the Constrained Non-negative Matrix

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Factorization algorithm (CNMF[52]). A custom, Matlab Image Processing Toolbox (Mathworks) based algorithm was used to construct initial hypotheses for the neuron shapes in a data-driven way. In brief, the 3D fluorescence movie was binarized to mark significantly active pixels, then connected components of this binary movie were found. Each of these components arose from a hypothetical neuron, but a neuron could have contributed to multiple components. A shape-based matching procedure was used to remove duplicates before using these as input to CNMF. The "finalized" components from CNMF were then selected post-hoc to identify those that resembled neural somata, using a multivariate classifier with a manual vetting step.

Time and task variables

As mice appeared to learn stereotyped running patterns[11], their x, y position in the T-maze were highly correlated with time and the eventual right/left-turn choice. To compare data across mice and trials of uneven durations, we resampled the neural and behavioral data according to a time-like coordinate that measured progression through epochs of the task. In this procedure, the time-traces of behavioral variables and neural activities were averaged within each time-bin, but no other smoothing was applied. Different numbers of equally-spaced bins were used for the five qualitative epochs of the trial: (1) pre-cue period; (2) cue period; (3) delay period; (4) a "turn" period up to the end of the trial; and (5) the ITI. For Fig. 1e the time-bins were ~200ms (72 bins) and for all other analyses ~1.1s (11 bins) in duration.

Because it is possible for cortical responses to have laterality preferences, we consistently expressed all variables relative to the brain hemisphere that was recorded from for a given mouse. That is, we defined choice, view angle, and treadmill velocity variables such that a positive sign corresponds to the mouse turning to the side ipsilateral to the recorded hemisphere.

Sequences

These analyses utilize only correct trials and followed previous work[53], but with cross-validation, i.e. the following were performed using half the trials in an imaging session. A neuron was defined as choice-specific if the distribution of activity in active periods (trial-average activity ≥ 25% of maximum for ~400ms) was significantly different in right- vs. left-choice trials (two-sample t-test, two tailed $p < 0.05$). A ridge-to-background excess was defined using the activity averaged over only preferred-choice (if relevant) trials, as the maximum minus the modal value across time. A neuron was determined to have significantly task-localized activity if no more than 5% of 1000 null hypothesis pseudo-datasets have a larger ridge-to-background excess. Each of these pseudo-datasets was generated by selecting a random time t_0 in the session, and then defining a pseudo activity time-series as $[F(t_0), F(t_0 + 1), \dots, F(t_n), F(1), \dots, F(t_0 - 1)]$ where $F(t)$is the original time-series. The preferred time of a neuron was defined as when its trial-average activity was maximal, and its activity field was defined as all contiguous time-points around this that have trial-average activity ≥ 50% of the maximum. Neurons were sorted by preferred time to determine order in a sequence.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The sequence display Fig. 1e utilized the above order of neurons, but the trial-average activity of a given neuron was computed using the left-out half of trials. Also using left-out trials, the reliability index was defined as the fraction of preferred-choice trials in which the activity averaged in a neuron's firing field is ≥ 3 times noise. Only significantly task-localized neurons with reliability ≥ 50% were included in Fig. 1e-g, and only significantly task-localized neurons in Fig. 1h. See Supplementary Fig. 1 for additional statistics for the above.

Distribution of projected neural states

Let $\underline{F}(t)$ be the trial-average neural state at time t in the trial. We defined the projection axis between two timepoints t_1 and t_2 as the unit vector $\underline{e}(t_1,t_2) \equiv [\underline{F}(t_1) - \underline{F}(t_2)]/||\underline{F}(t_1) - \underline{F}(t_2)||$. The neural state $F(t)$ projected onto this axis is defined as $proj_{t_1\to t_2}[F(t)] \equiv [F(t) - \underline{F}(t_1)] \cdot \underline{e}(t_1,t_2)$, i.e. the origin is at $\underline{F}(t_1)$. The actual distance along this projection axis depends on the number and activity scale of neurons, which we do not attempt to interpret. Thus for Fig. 2d we scale the projected distributions such that $proj_{t_1\to t_2}[F(t_1)] = 0$ and $proj_{t_1\to t_2}[F(t_2)] = 1$, in order to be able to pool data across sessions.

As a measure of overlap between the above projected distributions, we compute the Bhattacharyya coefficient[97] $\Sigma_i\sqrt{p_1(i)\,p_2(i)}$, where $p_1(i)$ is the probability density of $proj_{t_1\to t_2}[F(t_1)]$ in bin i, and analogously $p_2(i)$ is the probability density of $proj_{t_1\to t_2}[F(t_2)]$ in bin i. 101 bins were used with 0.1 spacing was used for evaluating the density histogram for this metric.

Decoding/encoding models

All these analyses used neural and behavioral data that were z-scored per timepoint in the trial, i.e. the time-dependent mean was subtracted and then the data divided by the time-dependent standard deviation. As we used linear models, this z-scoring can be easily "undone" because this corresponds to solving modified systems of equations (cf. Eq. 2 and Eq. 3):

$$\begin{aligned} &\underset{\widetilde{\mathbf{W}}_{\text{enc}}}{\text{argmin}} \left\| \mathbf{F}\,\mathbf{R}_{\mathbf{F}}^{-1} - \mathbf{X}\,\widetilde{\mathbf{W}}_{\text{enc}} \right\|_F^2 \quad \Rightarrow \quad \mathbf{W}_{\text{enc}} = \widetilde{\mathbf{W}}_{\text{enc}}\mathbf{R}_{\mathbf{F}} \\ &\underset{\widetilde{\mathbf{W}}_{\text{dec}}}{\text{argmin}} \left\| \mathbf{X} - \mathbf{F}\,\mathbf{R}_{\mathbf{F}}^{-1}\,\widetilde{\mathbf{W}}_{\text{dec}} \right\|_F^2 \quad \Rightarrow \quad \mathbf{W}_{\text{dec}} = \mathbf{R}_{\mathbf{F}}^{-1}\widetilde{\mathbf{W}}_{\text{dec}} \end{aligned} \qquad (1)$$

where $\mathbf{R}_{\mathbf{F}}$ is a diagonal matrix with elements being the standard deviations of each column of $\mathbf{F}$. The z-scored weights $\widetilde{\mathbf{W}}_{\text{enc}}$ were translated back to $\mathbf{W}_{\text{enc}}$ for Supplementary Fig. 10a and parts of Supplementary Fig. 11 (see caption) only. All models were constructed separately per timepoint, except for the time-independent decoders (Fig. 3e, Supplementary Fig. 3b,c) where all timepoints (or all timepoints within the stated phases of the trial) were included as if they were additional trials.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For decoding, we trained an L2-regularized Support Vector Machine classifier[98] (or L2-regularized Support Vector Regression for continuous variables) to predict a given task variable from the neural population state. For categorical variables, performance was defined as the accuracy (proportion correct) of classifying test trials, averaged across categories. For continuous variables, performance was defined as the Pearson's correlation coefficient between predicted and actual variable values. 3-fold cross-validation was used to evaluate all decoder performances. The p-value (significance) of this was defined as the fraction of shuffled datasets (activity of neurons permuted across trials) for which the decoding performance was greater or equal. We then used the Benjamini-Hochberg procedure[99] to control for false discovery rate as follows. For a given type of decoder, we sorted the p-values of all imaging sessions in ascending order, $[p_1, p_2, \dots, p_n]$, and found the first rank i_α such that $p_{i_\alpha} \leq i_\alpha \times 0.05/n$. The decoding performance was then considered to be significantly above chance for all $p \leq p_{i_\alpha}$.

For encoding we used unregularized linear regression so as not to impose additional task-variable-related structure, but results were qualitatively similar with L1 or L2 regularization (data not shown). Specifically, we computed the Singular Value Decomposition[100] (SVD) of the trial-by-variable matrix of task variables $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, then used the pseudo-inverse to solve for encoding weights $\vec{w} \equiv \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^\top\vec{f}$ where $\vec{f}$ is the activity over trials for a given neuron. Singular values $\leq 10^{-4}$ were set to 0 in $\mathbf{S}^{-1}$.

Leave-one-out angle w.r.t. other decoding directions

For a given decoding direction $\vec{d}$ and all other decoding directions (with statistically significant decoding performance) being the columns of $\mathbf{D}$, we first decomposed $\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}^{-1}$ (SVD) to obtain an orthonormal basis $\mathbf{U}$ for the span of $\mathbf{D}$, then compute the component within this subspace $\vec{d}_\mathbf{D} \equiv \mathbf{U}\mathbf{U}^\top\vec{d}$. Because $\vec{d}$ is a unit vector, the angle w.r.t. $\mathbf{D}$ is simply $\mathrm{acos}^{-1}(\|\vec{d}_\mathbf{D}\|)$. To evaluate statistical uncertainties, we constructed 100 bootstrap experiments where the data were sampled with replacement. Decoding models were trained and leave-one-out angles computed in the same way using these bootstrapped data, giving a distribution of angles as referred to in the text.

Multiplicative/additive time-modulation models

For the i^{th} neuron with activity $f_i(t)$ across trials, we fit two alternative encoding models with constrained changes in encoding weights vs. time t in the trial. Here $\vec{x}(t)$ are the values of nine task variables at time t, and have the time-dependent means across trials subtracted, but only scaled so that the standard deviation of each variable, computed across all timepoints as well as trials, are 1. This is in contrast to the per-timepoint encoding/decoding models, where the task variables were scaled differently per timepoint. We note that according to these models, the sensitivity of the neural-population encoding to the time-dependent scales of task variables (Supplementary Fig. 10a) should be ascribed to differences in task-variable encoding weights *across* individual neurons in the population.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The first, multiplicative time-modulation model (Fig. 7d-e) predicts that a neuron's activity has the form $f_i(t) = \mu_i(t) + g_i(t)\,\vec{u}_i \cdot \vec{x}(t)$, where $\mu_i(t)$ is a time-dependent shift in baseline that does not depend on task variables, $g_i(t)$ is a piecewise-constant function with 11 free parameters for the 11 time-bins, and $\vec{u}_i$ are 9 free parameters for linear dependencies on each of the task variables. Since $\vec{x}(t)$ has zero mean across trials for a fixed time t, without loss of generality $\mu_i(t)$ is just the trial-average mean activity level of the neuron. We estimated the 11+9 free parameters for the model by minimizing the L2-regularized least-squares cost function in a 3-fold cross-validation setting. The goodness-of-fit was defined as Pearson's correlation coefficient between the predicted and actual neural activity, and evaluated on test-sets of the cross-validation folds. The regularization hyperparameter was set to a small value $10^{-6} m/k$ where m is the number of data points (trials×timepoints) and k is the number of free parameters in the model.

The second, additive time-modulation model (Supplementary Fig. 10b-c) simply has $g_i(t) = 1$ for all timepoints, i.e. predicts neural activity of the form $f_i(t) = \mu_i(t) + \vec{u}_i \cdot \vec{x}(t)$. By comparing this to the multiplicative time-modulation model in a cross-validated setting, we explicitly test whether a non-constant $g_i(t)$ is important to explain the neural data.

Theoretical relationship between decoding and encoding

For a fixed time t in the trial and assuming neuronal responses are (locally) linear, we can derive a simple relationship between a hypothesized underlying brain structure and the estimated neural encoding/decoding weights, $\mathbf{W}_{\text{enc}}$ and $\mathbf{W}_{\text{dec}}$ respectively. For this purpose we treat the neural and behavioral data as matrices with rows being individual trials, i.e. $\mathbf{F}$ is a trial-by-neuron matrix of the neural state across trials, and $\mathbf{X}$ is a trial-by-variable matrix of behavioral factors across trials. Without loss of generality, we assume both $\mathbf{F}$ and $\mathbf{X}$ to be centered, i.e. means of each row subtracted. Then when we use linear regression to compute $\mathbf{W}_{\text{enc}}$ and $\mathbf{W}_{\text{dec}}$ from the data (neglecting regularization for simplicity), we solve the tightly related systems of equations:

$$\underset{\mathbf{W}_{\text{enc}}}{\text{argmin}} \ \|\mathbf{F} - \mathbf{X}\,\mathbf{W}_{\text{enc}}\|_F^2 \quad \Rightarrow \quad \widehat{\mathbf{W}}_{\text{enc}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{F} \qquad \mathbf{(2)}$$

$$\underset{\mathbf{W}_{\text{dec}}}{\text{argmin}} \ \|\mathbf{X} - \mathbf{F}\,\mathbf{W}_{\text{dec}}\|_F^2 \quad \Rightarrow \quad \widehat{\mathbf{W}}_{\text{dec}} = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{X} \qquad \mathbf{(3)}$$

In the above, $\|\cdot\|_F^2$ is the Frobenius norm (sum of squared elements of the enclosed matrix). If there are m task variables i.e. columns of $\mathbf{X}$, then the m rows of $\mathbf{W}_{\text{enc}}$ give the encoding directions and the m columns of $\mathbf{W}_{\text{dec}}$ give the decoding directions discussed in the text. For both encoding/decoding matrices $\widehat{\mathbf{W}}$ denotes the least-squares estimate of $\mathbf{W}$, and a derivation of the above solutions can be found in standard textbooks[100]. What we want to know is how Eq. 2 and Eq. 3 depend on the brain's "true" encoding scheme, i.e. we assume that the observed neural state $\mathbf{F}$ arises from an underlying stochastic relationship:

$$\mathbf{F} = \mathbf{X}\mathbf{W}_{\text{enc}} + \varepsilon \quad \text{where each row (trial) has} \quad \vec{\varepsilon} \sim \mathcal{N}(\vec{0}, \mathbf{S}) \qquad \mathbf{(4)}$$

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $\mathbf{S}$ is the neuronal noise covariance. Note that for conciseness in the main text, we omitted the overhat notation for all equations, but our derivation distinguishes between $\mathbf{W}_{\text{enc}}$ and related quantities, which involve the unknown (but assumed to be linear) encoding scheme of the brain, as opposed to $\widehat{\mathbf{W}}_{\text{enc}}$ and $\widehat{\mathbf{W}}_{\text{dec}}$, which are estimates computed from the neural and behavioral data. Our setup of the encoding/decoding system and solution for $\widehat{\mathbf{W}}_{\text{dec}}$ are highly similar to previous studies[28,29], but provided below for convenience.

Given the above assumption that the brain's response is stochastic, we derive the expected value of $\widehat{\mathbf{W}}_{\text{enc}}$ and $\widehat{\mathbf{W}}_{\text{dec}}$ across experiments (many possible neural responses $\mathbf{F}$) under the same behavioral conditions ($\mathbf{X}$ is fixed and known). Indicating expectation values by $\langle \cdot \rangle$, we plug Eq. 4 into Eq. 2 to get:

$$\langle \widehat{\mathbf{W}}_{\text{enc}} \rangle = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \langle \mathbf{F} \rangle = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\mathbf{W}_{\text{enc}} + \langle \varepsilon \rangle) = \mathbf{W}_{\text{enc}} \qquad (5)$$

The same is a little more complicated for Eq. 3:

$$\begin{aligned} \langle \widehat{\mathbf{W}}_{\text{dec}} \rangle &= \langle (\mathbf{F}^\top \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{X} \rangle \\ &= \langle [(\mathbf{W}_{\text{enc}}^\top \mathbf{X}^\top + \varepsilon^\top)(\mathbf{X}\mathbf{W}_{\text{enc}} + \varepsilon)]^{-1} (\mathbf{W}_{\text{enc}}^\top \mathbf{X}^\top + \varepsilon^\top) \mathbf{X} \rangle \end{aligned}$$

If the neural responses are not trivial ($\mathbf{W}_{\text{enc}} \neq \mathbf{0}$) and there are enough trials per experiment (see next section for potential concerns when otherwise), then the random coincidences of the neuronal noise ε with the task variables $\mathbf{X}$ should be negligible compared to the signal-related pieces. That is, we can approximately drop the terms $\mathbf{X}^\top \varepsilon$ even within expectation values. Then defining the behavioral covariance matrix to be $\mathbf{C} = \mathbf{C}^\top \equiv \mathbf{X}^\top \mathbf{X}$, we obtain:

$$\langle \widehat{\mathbf{W}}_{\text{dec}} \rangle \approx \langle (\mathbf{W}_{\text{enc}}^\top \mathbf{X}^\top \mathbf{X} \mathbf{W}_{\text{enc}} + \varepsilon^\top \varepsilon)^{-1} \mathbf{W}_{\text{enc}}^\top \mathbf{X}^\top \mathbf{X} \rangle \approx (\mathbf{W}_{\text{enc}}^\top \mathbf{C} \mathbf{W}_{\text{enc}} + \mathbf{S})^{-1} \mathbf{W}_{\text{enc}}^\top \mathbf{C} \qquad (6)$$

For the last approximate equality, we also assumed that the number of trials is large, so that $\varepsilon^\top \varepsilon \approx \mathbf{S}$ within an experiment. Finally, it will be convenient to rewrite the above using the push-through matrix identity[101] $(\mathbf{A} + \mathbf{M}\mathbf{B}\mathbf{N})^{-1}\mathbf{M} = \mathbf{A}^{-1}\mathbf{M}(\mathbf{B}^{-1} + \mathbf{N}\mathbf{A}^{-1}\mathbf{M})^{-1}\mathbf{B}^{-1}$, where $\mathbf{A} \to \mathbf{S}$, $\mathbf{B} \to \mathbf{I}$ (the identity matrix), $\mathbf{M} \to \mathbf{W}_{\text{enc}}^\top \mathbf{C}$, and $\mathbf{N} \to \mathbf{W}_{\text{enc}}$, giving:

$$\langle \widehat{\mathbf{W}}_{\text{dec}} \rangle \approx \mathbf{S}^{-1} \mathbf{W}_{\text{enc}}^\top \mathbf{C} \, (\mathbf{I} + \mathbf{W}_{\text{enc}} \mathbf{S}^{-1} \mathbf{W}_{\text{enc}}^\top \mathbf{C})^{-1} \qquad (7)$$

Although Eq. 7 indicates that the decoding weights should more correctly be understood relative to the structure of noise correlations between neurons[45,102–106], it empirically turned out that we were able to neglect $\mathbf{S}$ when modeling the relationships between decoding directions. In particular, we approximated $\mathbf{S} \approx \sigma^2 \mathbf{I}$ and introduced $\phi_{\text{enc}} = \phi_{\text{enc}}^\top \equiv \sigma^{-2} \mathbf{W}_{\text{enc}} \mathbf{W}_{\text{enc}}^\top$ to obtain:

$$\langle \widehat{\mathbf{W}}_{\text{dec}} \rangle \approx \sigma^{-2} \mathbf{W}_{\text{enc}}^\top \mathbf{C} \, (\mathbf{I} + \phi_{\text{enc}} \mathbf{C})^{-1} \qquad (8)$$

This approximation also holds if neurons have noise correlations in $\mathbf{S}$ that are proportional to their signal correlations $\mathbf{W}_{\text{enc}}^\top \mathbf{C} \mathbf{W}_{\text{enc}}$. We see this by plugging $\mathbf{S} = \gamma \mathbf{W}_{\text{enc}}^\top \mathbf{C} \mathbf{W}_{\text{enc}} + \rho^2 \mathbf{I}$

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

into Eq. 6, which gives $\langle\widehat{\mathbf{W}}_{\text{dec}}\rangle \approx ([1+\gamma]\mathbf{W}_{\text{enc}}^{\top}\mathbf{C}\mathbf{W}_{\text{enc}} + \rho^2\mathbf{I})^{-1}\mathbf{W}_{\text{enc}}^{\top}\mathbf{C}$ $= (1+\gamma)^{-1}(\mathbf{W}_{\text{enc}}^{\top}\mathbf{C}\mathbf{W}_{\text{enc}} + \nu^2\mathbf{I})^{-1}\mathbf{W}_{\text{enc}}^{\top}\mathbf{C}$. This differs from Eq. 6 only by an overall scale $(1+\gamma)^{-1}$ that does not change geometrical properties discussed below and in the text, with $\nu^2 \equiv \rho^2/(1+\gamma)$ playing the role of σ^2. Furthermore, if we consider an L2-regularized linear decoding model (as we use in practice), this means that instead of Eq. 3 we use the estimator[100] $\widehat{\mathbf{W}}_{\text{dec}} = (\mathbf{F}^{\top}\mathbf{F} + \lambda\mathbf{I})^{-1}\mathbf{F}^{\top}\mathbf{X}$ for some small regularization hyperparameter λ. This is equivalent to replacing $\mathbf{S} \to \mathbf{S} + \lambda\mathbf{I}$ in the derivation of Eq. 7, which if anything improves the $\mathbf{S} \approx \sigma^2\mathbf{I}$ approximation.

Assuming $\mathbf{S} \approx \sigma^2\mathbf{I}$ (and again neglecting terms like $\mathbf{X}^{\top}\varepsilon$, but see next section for caveats), the geometry of the estimated decoding and encoding directions are given by:

$$\begin{aligned}\widehat{\phi}_{\text{dec}} &\equiv \langle\widehat{\mathbf{W}}_{\text{dec}}^{\top}\widehat{\mathbf{W}}_{\text{dec}}\rangle \approx \sigma^{-2}(\mathbf{I}+\mathbf{C}\phi_{\text{enc}})^{-1}\mathbf{C}\,\phi_{\text{enc}}\,\mathbf{C}\,(\mathbf{I}+\phi_{\text{enc}}\mathbf{C})^{-1}\\ \widehat{\phi}_{\text{enc}} &\equiv \langle\widehat{\mathbf{W}}_{\text{enc}}\widehat{\mathbf{W}}_{\text{enc}}^{\top}\rangle \approx \mathbf{W}_{\text{enc}}\mathbf{W}_{\text{enc}}^{\top} = \sigma^2\phi_{\text{enc}}\end{aligned} \qquad \textbf{(9)}$$

$\widehat{\phi}_{\text{dec}}$ is the expected value of the matrix of dot products between pairs of columns of $\widehat{\mathbf{W}}_{\text{dec}}$, which gives a prediction for the cosine angles between observed decoding directions ($\widehat{\theta}_{\text{dec}}$) after dividing by the vector norms of the two columns in question (i.e. the square-root of the diagonal entries of $\widehat{\phi}_{\text{dec}}$). Because of this normalization, any overall scaling of $\widehat{\phi}_{\text{dec}}$ cannot change the predicted $\widehat{\theta}_{\text{dec}}$; or in other words, scaling two vectors by the same factor does not change the angle between them. In particular, this means that the factor of σ^{-2} that multiplies all entries of $\widehat{\phi}_{\text{dec}}$ has no effect on the predicted $\widehat{\theta}_{\text{dec}}$. An identical argument can be made for the predicted angles between encoding directions. Together, this means that we can ignore σ in Eq. 9 entirely and discuss the observed encoding/decoding angles as related through (only) the measured behavioral covariance matrix, $\mathbf{C}$, and the unknown “true” encoding structure of the brain, ϕ_{enc}.

As a scientific note that does not affect the above formulae, the neuronal data in our hands seemed to be better explained by z-scored behavioral factors $\mathbf{X}$, i.e. mean subtracted and divided by the standard deviation of each row, for each timepoint t (see Supplementary Fig. 10a and text). This means we can best model the data by taking $\mathbf{C}$ to be the correlation instead of covariance matrix, as we have done in the main text.

When there is no signal

In the above derivations, we assumed that the neural population has nonzero response to task variables, $\mathbf{W}_{\text{enc}} \neq \mathbf{0}$. Here we consider the case where there is little to no signal response, $\mathbf{X}\mathbf{W}_{\text{enc}} \ll \varepsilon$, where ε is a trial-by-neuron matrix of random noise fluctuations in one experiment. This means that we can no longer neglect terms like $\mathbf{X}^{\top}\varepsilon$ as we did to obtain Eq. 5 and Eq. 6. In particular, the angles between pairs of encoding/decoding directions can have nonzero expectation values because they depend on second-order statistics of the neural noise[69]:

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\begin{aligned}\widehat{\phi}_{\text{enc}}^{\text{noise}} &\equiv \langle \widehat{\mathbf{W}}_{\text{enc}} \widehat{\mathbf{W}}_{\text{enc}}^{\top} \rangle_{\text{noise}} = \mathbf{C}^{-1}\mathbf{X}^{\top}\langle \varepsilon\varepsilon^{\top}\rangle \mathbf{X}\mathbf{C}^{-1} \\ \widehat{\phi}_{\text{dec}}^{\text{noise}} &\equiv \langle \widehat{\mathbf{W}}_{\text{dec}}^{\top} \widehat{\mathbf{W}}_{\text{dec}} \rangle_{\text{noise}} \approx \mathbf{X}^{\top}\langle \varepsilon \left(\varepsilon^{\top}\varepsilon\right)^{-2}\varepsilon^{\top}\rangle \mathbf{X} \qquad (10)\end{aligned}$$

Since ε is different for every experiment and has no structure related to the behavior, we might expect to see high variability of encoding/decoding angle structures across experiments in this no-signal neural scenario. The worst case regarding false discoveries is when the terms involving ε actually turn out to be highly similar across experiments, leading to a false impression of clear, behavior-related structure. We consider such a case, which happens under two assumptions. One, $\varepsilon\varepsilon^{\top}$ is a matrix of dot products (a.k.a. similarities) between neural-noise states of various pairs of trials. If there is no across-trial structure of the neural noise, and there are many neurons, then each element of $\varepsilon\varepsilon^{\top}$ is a sum of many random numbers. Except for the diagonal elements, this converges to zero according to the central limit theorem, i.e. $\varepsilon\varepsilon^{\top} \propto \mathbf{I}$. Two, when there are many trials then $\varepsilon^{\top}\varepsilon \approx \mathbf{S}$ even within an experiment, and we again assume $\mathbf{S} \approx \sigma^2\mathbf{I}$. With these two assumptions, we obtain:

$$\begin{aligned}\widehat{\phi}_{\text{enc}}^{\text{noise}} &\propto \mathbf{C}^{-1}\mathbf{X}^{\top}\mathbf{X}\mathbf{C}^{-1} = \mathbf{C}^{-1} \\ \widehat{\phi}_{\text{dec}}^{\text{noise}} &\propto \mathbf{X}^{\top}\langle\varepsilon\varepsilon^{\top}\rangle\mathbf{X} \propto \mathbf{X}^{\top}\mathbf{X} = \mathbf{C} \qquad (11)\end{aligned}$$

Eq. 11 pose potential confounds for the interpretation of our experimental observations. In particular, we observed encoding angles $\propto \mathbf{C}^{-1}$ and decoding angles $\propto \mathbf{C}$, which we wished to interpret as a biological statement about how the brain encoded task variables. Unfortunately, Eq. 11 shows that even if the neural population does not respond to task variables, we can still obtain these kind of encoding/decoding observations. Fortunately, there are several predictions of this no-signal scenario that do not match the data in our hands. First, the encoding and decoding predictions should be at chance levels compared to permutation tests where we randomized the task variables $\mathbf{X}$ across trials, creating a no-signal scenario. However, Fig. 3 shows that all variables could be decoded to significantly above-chance levels, Supplementary Fig. 8b-right shows that for many neurons, the variance explained by encoding models is substantial for at least one timepoint in the trial, and Supplementary Fig. 10a shows that the distribution of encoding weights across neurons had a tail that extended to weights of large magnitudes (on the order of neural activity scales).

The data did not match a no-signal scenario, but it could still be that the noise-only terms in Eq. 11 should have been included in Eq. 9 for interpretation of encoding/decoding angles. For example, retaining all ε-related terms when computing $\widehat{\phi}_{\text{enc}} \equiv \langle \widehat{\mathbf{W}}_{\text{enc}} \widehat{\mathbf{W}}_{\text{enc}}^{\top} \rangle$ yields:

$$\widehat{\phi}_{\text{enc}} = \mathbf{W}_{\text{enc}}\mathbf{W}_{\text{enc}}^{\top} + \mathbf{W}_{\text{enc}}\langle\varepsilon^{\top}\rangle\mathbf{X}\mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{X}^{\top}\langle\varepsilon\rangle + \mathbf{C}^{-1}\mathbf{X}^{\top}\langle\varepsilon\varepsilon^{\top}\rangle\mathbf{X}\mathbf{C}^{-1}$$

Assuming $\langle\varepsilon\rangle = \mathbf{0}$, we find that $\widehat{\phi}_{\text{enc}} = \mathbf{W}_{\text{enc}}\mathbf{W}_{\text{enc}}^{\top} + \mathbf{C}^{-1}\mathbf{X}^{\top}\langle\varepsilon\varepsilon^{\top}\rangle\mathbf{X}\mathbf{C}^{-1}$ is a mixture of the signal-related piece in Eq. 9 and the noise-only term in Eq. 10. The data could thus include signal responses, yet the structure of encoding/decoding angles shown in Fig. 5 and Fig. 6 could be driven mostly by pure-noise contributions. However, if this was the case we should find that the $\widehat{\phi}_{\text{enc}} \propto \mathbf{C}^{-1}$ and $\widehat{\phi}_{\text{dec}} \propto \mathbf{C}$ correspondences are better at low signal-to-

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

noise (SNR), i.e. exhibit a decreasing trend w.r.t. SNR. Using the performance of task-variable decoders as a measure of SNR at the neural-population level, we instead see in Supplementary Fig. 6a that the $\widehat{\phi}_{\text{dec}} \propto \mathbf{C}$ agreement *increased* with SNR for all posterior cortical areas. Albeit the generally more noisy nature of encoding model comparisons, Supplementary Fig. 6b shows a similar increase of the $\widehat{\phi}_{\text{enc}} \propto \mathbf{C}^{-1}$ agreement with SNR for all areas. These increasing trends vs. SNR argue against a noise-induced origin.

An even more stringent test uses a conservative cross-validation paradigm[107–109], where two sets of decoding directions were computed separately using two disjoint subsets of the data. In this case, random neural fluctuations selected by (say) decoding directions computed using the first half of trials should have no relationship to those computed using the second half of trials, and we should thus find no noise-induced structure in angles between decoding directions computed in two different halves of the data. Mathematically, this means that we compute $\langle\widehat{\mathbf{W}}_{\text{dec}}\rangle\langle\widehat{\mathbf{W}}_{\text{dec}}^{\top}\rangle$ instead of $\langle\widehat{\mathbf{W}}_{\text{dec}}\widehat{\mathbf{W}}_{\text{dec}}^{\top}\rangle$ (and actually Eq. 9 should have more correctly be written as $\widehat{\phi}_{\text{dec}} \equiv \langle\widehat{\mathbf{W}}_{\text{dec}}\rangle\langle\widehat{\mathbf{W}}_{\text{dec}}^{\top}\rangle$). Supplementary Fig. 7a shows that the cross-validated $\langle\widehat{\mathbf{W}}_{\text{dec}}\rangle\langle\widehat{\mathbf{W}}_{\text{dec}}^{\top}\rangle$ matches $\mathbf{C}$ nearly as well as the $\langle\widehat{\mathbf{W}}_{\text{dec}}\widehat{\mathbf{W}}_{\text{dec}}^{\top}\rangle$ estimations used throughout the rest of the paper. Supplementary Fig. 7b shows that when using such a cross-validation scheme, noise does not contribute any relationship between $\langle\widehat{\mathbf{W}}_{\text{dec}}\rangle\langle\widehat{\mathbf{W}}_{\text{dec}}^{\top}\rangle$ and $\langle\widehat{\mathbf{W}}_{\text{dec}}\widehat{\mathbf{W}}_{\text{dec}}^{\top}\rangle$ (a permutation test where neural activities were randomly shuffled across trials, breaking neuron-behavior relationships but preserving correlations between task variables). A similar line of reasoning applies to encoding angles, and Supplementary Fig. 7c shows that there is a significant trend in how well $\langle\widehat{\mathbf{W}}_{\text{enc}}\rangle\langle\widehat{\mathbf{W}}_{\text{enc}}^{\top}\rangle$ vs. $\langle\widehat{\mathbf{W}}_{\text{enc}}\widehat{\mathbf{W}}_{\text{enc}}^{\top}\rangle$ match the $\widehat{\phi}_{\text{enc}} \propto \mathbf{C}^{-1}$ hypothesis, which is not present in the permutation test (Supplementary Fig. 7d). We note an overall reduction of $\widehat{\phi}_{\text{enc}} \propto \mathbf{C}^{-1}$ agreement in the cross-validated scenario, due to the higher variability of encoding directions across different halves of the dataset compared to decoding directions. Nevertheless, these cross-validated results for decoding and encoding angles support that our encoding/decoding structure observations were driven by signal-related neural responses, and not artifacts of finding spurious structure in noise.

Three possible encoding scenarios, and what they say about decoding

From Eq. 8, the decoding weights depend on the neural noise level σ through the term $\phi_{\text{enc}} \equiv \sigma^{-2}\mathbf{W}_{\text{enc}}\mathbf{W}_{\text{enc}}^{\top}$. As $\mathbf{W}_{\text{enc}}$ specifies how strongly neurons respond to the task variables a.k.a. the signal strength, ϕ_{enc} can be thought of as the brain's encoding structure in units of signal-to-noise (SNR). We consider here three interesting cases for what ϕ_{enc} might be.

In the first, high SNR scenario ϕ_{enc} is large, and in the limiting case of $\phi_{\text{enc}}\mathbf{C} \gg \mathbf{I}$ we find:

$$\langle\widehat{\mathbf{W}}_{\text{dec}}\rangle \approx \sigma^{-2}\mathbf{W}_{\text{enc}}^{\top}\mathbf{C}(\phi_{\text{enc}}\mathbf{C})^{-1} = \sigma^{-2}\mathbf{W}_{\text{enc}}^{\top}\mathbf{C}\mathbf{C}^{-1}\phi_{\text{enc}}^{-1} = \mathbf{W}_{\text{enc}}^{\top}\left(\mathbf{W}_{\text{enc}}\mathbf{W}_{\text{enc}}^{\top}\right)^{-1} \qquad \mathbf{(12)}$$

where we have assumed that both $\mathbf{C}$ and ϕ_{enc} are invertible. The rightmost formula is recognizable as the Moore-Penrose pseudoinverse[110] of the brain's presumed encoding

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

matrix $\mathbf{W}_{\text{enc}}$. It follows that the angles between decoding directions are given by (up to an ignorable factor of σ^{-2}:

$$\begin{aligned} &\langle \widehat{\mathbf{W}}_{\text{dec}} \rangle \propto \mathbf{W}_{\text{enc}}^{\top} \phi_{\text{enc}}^{-1} \\ &\quad \widehat{\phi}_{\text{dec}} \equiv \langle \widehat{\mathbf{W}}_{\text{dec}}^{\top} \widehat{\mathbf{W}}_{\text{dec}} \rangle \propto \phi_{\text{enc}}^{-1} \mathbf{W}_{\text{enc}} \mathbf{W}_{\text{enc}}^{\top} \phi_{\text{enc}}^{-1} \propto \phi_{\text{enc}}^{-1} \, \phi_{\text{enc}} \, \phi_{\text{enc}}^{-1} = \phi_{\text{enc}}^{-1} \end{aligned} \qquad \textbf{(13)}$$

In the opposite extreme where $\phi_{\text{enc}}\mathbf{C} \ll \mathbf{I}$, then $\langle \widehat{\mathbf{W}}_{\text{dec}} \rangle \approx \sigma^{-2}\mathbf{W}_{\text{enc}}^{\top}\mathbf{C}$, which means that the optimal decoding directions (columns of $\langle \widehat{\mathbf{W}}_{\text{dec}} \rangle$) are simply the correlation-weighted sum of encoding directions (rows of $\mathbf{W}_{\text{enc}}$). More generally we define the second, low (but nonzero) SNR scenario as one where $\lim\limits_{n \to \infty} (\phi_{\text{enc}}\mathbf{C})^n = 0$, as then $(\mathbf{I} + \phi_{\text{enc}}\mathbf{C})^{-1}$ can be expanded in a Neumann series[111] to obtain:

$$\widehat{\phi}_{\text{dec}} \propto \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} (-\mathbf{C}\phi_{\text{enc}})^i \, \mathbf{C} \, \phi_{\text{enc}} \, \mathbf{C} \, (-\phi_{\text{enc}}\mathbf{C})^j = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} (-1)^{i+j} (\mathbf{C} \, \phi_{\text{enc}})^{i+1} \, \mathbf{C} \, (\phi_{\text{enc}}\mathbf{C})^j$$

If ϕ_{enc} has no special structure related to $\mathbf{C}$, then the above indicates that roughly speaking, the lowest power to which $\widehat{\phi}_{\text{dec}}$ can depend on $\mathbf{C}$ is $\mathbf{C}^2$, with higher powers contributing successively smaller corrections to the series expansion. This is explicitly so for orthogonal encoding $\phi_{\text{enc}} = \zeta\mathbf{I}$, where ζ is a scalar that specifies the SNR:

$$\widehat{\phi}_{\text{dec}} \propto \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} (-\zeta)^{i+j} \mathbf{C}^{i+j+2} = \sum_{k=0}^{\infty} (k+1)(-\zeta)^k \mathbf{C}^{k+2} \qquad \textbf{(14)}$$

The third and last scenario that we consider is where $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$, which has a simple form that is qualitatively different from Eq. 14:

$$\widehat{\phi}_{\text{dec}} \propto [(\mathbf{I}+\mathbf{I})^{-1}]^{\top} \mathbf{C} \, \mathbf{I} \, (\mathbf{I}+\mathbf{I})^{-1} \propto \mathbf{C} \qquad \textbf{(15)}$$

We note that this is in some sense an intermediate SNR scenario, since neither the $\phi_{\text{enc}}\mathbf{C} \gg \mathbf{I}$ (high SNR) nor $\phi_{\text{enc}}\mathbf{C} \ll \mathbf{I}$ (low SNR) criteria are fulfilled when $\phi_{\text{enc}}\mathbf{C} = \mathbf{I}$. However the particular choice of $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ makes the decoding observations in this third scenario indistinguishable from the first, high-SNR scenario (Eq. 13), although there are of course other possibilities for ϕ_{enc} in the latter.

Why we call $\phi_{\mathbf{U}} \propto \mathbf{C}^{-1}$ an (effectively) whitening/decorrelation operation

Given a trial-by-variable data matrix $\mathbf{Y}$ with arbitrary covariance matrix, and the transformed data $\mathbf{Z} = \mathbf{YT}$, we call $\mathbf{T}$ a whitening transformation if the transformed covariance is $\mathbf{Z}^{\top}\mathbf{Z} = \mathbf{I}$. Below, we show that constraining $\phi_{\text{enc}} \propto \mathbf{C}^{-1}$ corresponds to such a whitening transformation for $\mathbf{Z} = \mathbf{FD}$, where $\mathbf{D}$ is an orthonormal basis ($\mathbf{D}^{\top}\mathbf{D} = \mathbf{I}$) for the brain's encoding matrix $\mathbf{W}_{\text{enc}}^{\top}$ (Eq. 4). $\mathbf{Z}$ can be understood as the projection of the neural state $\mathbf{F}$ onto the information-coding subspace (spanned by basis vectors $\mathbf{D}$) as discussed in the text. We first show that optimal linear decoding depends only on this projected neural state $\mathbf{Z} = \mathbf{FD}$, and then show that the covariance of the projected neural state is $\mathbf{Z}^{\top}\mathbf{Z} = \mathbf{I}$.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For the following derivations, we will need the fact that applying the projection operator $\mathbf{D}\mathbf{D}^\top$ to $\mathbf{W}_{\mathrm{enc}}^\top$ does nothing: $\mathbf{D}\mathbf{D}^\top\mathbf{W}_{\mathrm{enc}}^\top = \mathbf{W}_{\mathrm{enc}}^\top$, because the columns of $\mathbf{W}_{\mathrm{enc}}^\top$ are by construction vectors that live in the subspace spanned by $\mathbf{D}$. No other properties of $\mathbf{D}$ are required, i.e. $\mathbf{D}$ can be any orthonormal basis for the information-coding subspace.

Recalling from Eq. 4 that the neural representation of the task variables is $\mathbf{F} \approx \mathbf{X}\mathbf{W}_{\mathrm{enc}}$, the covariance of the neural data is $\mathbf{F}^\top\mathbf{F} \approx \mathbf{W}_{\mathrm{enc}}^\top\mathbf{C}\mathbf{W}_{\mathrm{enc}} + \mathbf{S}$ (see derivation of Eq. 6). In the presumably typical case where there are more neurons than encoded variables, specifying the variable-by-variable matrix $\phi_{\mathrm{enc}} \propto \mathbf{W}_{\mathrm{enc}}\mathbf{W}_{\mathrm{enc}}^\top$ will not fully constrain the variable-by-neuron matrix $\mathbf{W}_{\mathrm{enc}}$, and is therefore insufficient to completely whiten the neural data in the sense of achieving $\mathbf{F}^\top\mathbf{F} \propto \mathbf{I}$ even with neural noise covariance $\mathbf{S} \propto \mathbf{I}$. However, we can show that the decoded information does not depend on the full high-dimensional $\mathbf{F}$, but only the neural-state values in a subspace spanned by $\mathbf{D}$ (assuming $\mathbf{S} \propto \mathbf{I}$). The decoded values are given by projecting the neural state onto the decoding vectors $\mathbf{W}_{\mathrm{dec}}$ (Eq. 3). Plugging $\mathbf{D}\mathbf{D}^\top\mathbf{W}_{\mathrm{enc}}^\top = \mathbf{W}_{\mathrm{enc}}^\top$ into Eq. 8, we see that the decoded values are $\mathbf{F}\langle\widehat{\mathbf{W}}_{\mathrm{dec}}\rangle \propto (\mathbf{F}\mathbf{D})\,\mathbf{E}^\top\mathbf{C}\,(\mathbf{I} + \phi_{\mathrm{enc}}\mathbf{C})^{-1}$, which depends only on the projected neural state $\mathbf{F}\mathbf{D}$ and not the full $\mathbf{F}$.

To understand how the encoding structure affects the covariance of $\mathbf{F}\mathbf{D}$, we first note that although the encoding directions $\mathbf{W}_{\mathrm{enc}}^\top$ are vectors in the high-dimensional neural state space, the dot products between encoding directions (ϕ_{enc}) depend only on the projected coordinates of these encoding directions in the information-coding subspace, $\mathbf{E}^\top \equiv \mathbf{D}^\top\mathbf{W}_{\mathrm{enc}}^\top$ (a variable-by-variable matrix). This is because plugging in $\mathbf{D}\mathbf{D}^\top\mathbf{W}_{\mathrm{enc}}^\top = \mathbf{W}_{\mathrm{enc}}^\top$ to $\phi_{\mathrm{enc}} \propto \mathbf{W}_{\mathrm{enc}}\mathbf{W}_{\mathrm{enc}}^\top$ gives $\phi_{\mathrm{enc}} \propto \mathbf{W}_{\mathrm{enc}}\mathbf{D}\mathbf{D}^\top\mathbf{W}_{\mathrm{enc}}^\top = \mathbf{E}\mathbf{E}^\top$. Our claim is that *any* invertible $\mathbf{E}$ that satisfies $\phi_{\mathrm{enc}} \propto \mathbf{C}^{-1}$ will whiten $\mathbf{F}\mathbf{D}$ (up to an overall scale, which we ignore). To see this, start from $\phi_{\mathrm{enc}}^{-1} = (\mathbf{E}^{-1})^\top\mathbf{E}^{-1} = \mathbf{C}$, left-multiply by $\mathbf{E}^\top$ and right-multiply by $\mathbf{E}$ to get $\mathbf{I} = \mathbf{E}^\top\mathbf{C}\mathbf{E}$. The covariance of the projected neural state is $(\mathbf{F}\mathbf{D})^\top(\mathbf{F}\mathbf{D}) = \mathbf{D}^\top(\mathbf{W}_{\mathrm{enc}}^\top\mathbf{C}\mathbf{W}_{\mathrm{enc}} + \mathbf{S})\mathbf{D}$. Again assuming $\mathbf{S} = \sigma^2\mathbf{I}$, we get $(\mathbf{F}\mathbf{D})^\top(\mathbf{F}\mathbf{D}) = \mathbf{E}^\top\mathbf{C}\mathbf{E} + \sigma^2\mathbf{D}^\top\mathbf{D} \propto \mathbf{I}$, as claimed. In sum, we call $\phi_{\mathrm{enc}} \propto \mathbf{C}^{-1}$ a whitening operation because it is the constraint that if exactly satisfied, will whiten $\mathbf{F}\mathbf{D}$.

Effect of per-timepoint z-scoring

As mentioned above, the per-timepoint encoding models were fit to neural data that had been z-scored per timepoint. If we "undo" this time-dependent scaling of neural data as previously explained for Eq. 1, Supplementary Fig. 11a-e shows that there is some increase time variation in encoding directions and angles. This in fact suggests a way for us to more directly gauge how much of an effect time-varying encoding directions can have on how well the encoding angles tracked $\mathbf{C}^{-1}(t)$. Note that according to the multiplicative model, changes in encoding directions are due to the time-modulations $g_i(t)$ of each neuron. We can *reduce* these time-modulations by z-scoring the neural data per timepoint, which as illustrated in Supplementary Fig. 11g stabilizes task-variable responses around the peak activity period of each neuron. Encoding models trained with z-scored data produced better agreement of encoding angles with $\mathbf{C}^{-1}(t)$ than models without z-scoring

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Supplementary Fig. 11h). The time-modulations $g_i(t)$ of neurons a.k.a. time-variations in encoding directions can thus indeed add substantial variability to the encoding angles, in the sense of causing them to deviate from $\mathbf{C}^{-1}(t)$. However and most intriguingly, the effect of z-scoring significantly diminished for sessions with larger numbers of recorded neurons (Supplementary Fig. 11h-right). This trend hints at how large, sequentially active populations could effectively behave as if neurons had near-stable behavioral responses analogous to the z-scored data, for reasons hypothesized in the text.

Acknowledgements

We thank J. Pillow for some of the math in this paper, M. S. Goldman and A. M. Essin for help with the presentation, as well as S. Lewallen, R. Low, and M. Schottdorf for impetus to dive into the strange new world of high-dimensional geometry. B. Engelhard and L. Pinto built rigs for high-throughput training of mice, and S. Stein helped in the training of mice in this study. We additionally thank all members of the BRAIN COGS team, and present and past members of the Tank and Brody labs. This work was supported by the NIH grants 5U01NS090541 and 1U19NS104648, and the Simons Collaboration on the Global Brain (SCGB).

Author Contributions

SAK performed the experiments, data analysis and conceptualization/modeling. SYT and DWT designed the experimental setups. SAK wrote the manuscript with input from DWT and CDB. SAK, DWT, and CDB conceived the project.

Declaration of Interests

The authors declare no competing interests.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

Custom code (Matlab and C++) used to perform all analyses are available from the corresponding author upon reasonable request.

References

1. Keller, G. B. & Mrsic-Flogel, T. D. Predictive Processing: A Canonical Cortical Computation.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Neuron **100**, 424–435 (2018).

2. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
3. Bastos, A. M. *et al.* Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).
4. Helmholtz, H. von. Concerning the perceptions in general, 1867. in *Readings in the history of psychology.* 214–230
5. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–1438 (2006).
6. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).
7. Steinmetz, N., Zatka-Haas, P., Carandini, M. & Harris, K. Distributed correlates of visually-guided behavior across the mouse brain. *bioRxiv* (2018).
8. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, 255 (2019).
9. Minderer, M., Brown, K. D. & Harvey, C. D. The Spatial Structure of Neural Encoding in Mouse Posterior Cortex during Navigation. *Neuron* **102**, 232–248.e11 (2019).
10. Musall, S., Kaufman, M. T., Gluf, S. & Churchland, A. K. Movement-related activity dominates cortex during sensory-guided decision making. *BioRxiv* (2018).
11. Pinto, L. *et al.* An Accumulation-of-Evidence Task Using Visual Pulses for Mice Navigating in Virtual Reality. *Front. Behav. Neurosci.* **12**, 36 (2018).
12. BRAIN CoGS Collaboration. BRAIN Circuits of coGnitive Systems. Available at: https://www.braincogs.org/.
13. Pinto, L. *et al.* Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions. *Neuron* (2019). doi:10.1016/j.neuron.2019.08.025
14. Jazayeri, M. & Afraz, A. Navigating the Neural Space in Search of the Neural Code. *Neuron* **93**,

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1003–1014 (2017).

15. Ganguli, S. & Sompolinsky, H. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu. Rev. Neurosci.* **35**, 485–508 (2012).
16. Gallego, J. A., Perich, M. G., Miller, L. E. & Solla, S. A. Neural Manifolds for the Control of Movement. *Neuron* **94**, 978–984 (2017).
17. Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423–426 (2014).
18. Tsodyks, M., Kenet, T., Grinvald, A. & Arieli, A. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science* **286**, 1943–1946 (1999).
19. Okun, M. *et al.* Diverse coupling of neurons to populations in sensory cortex. *Nature* **521**, 511–515 (2015).
20. Stopfer, M., Jayaraman, V. & Laurent, G. Intensity versus identity coding in an olfactory system. *Neuron* **39**, 991–1004 (2003).
21. Luczak, A., Barthó, P. & Harris, K. D. Spontaneous events outline the realm of possible sensory responses in neocortical populations. *Neuron* **62**, 413–425 (2009).
22. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. *Annu. Rev. Neurosci.* **36**, 337–359 (2013).
23. Yu, B. M. *et al.* Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *Journal of Neurophysiology* **102**, 614–635 (2009).
24. Pang, R., Lansdell, B. J. & Fairhall, A. L. Dimensionality reduction in neuroscience. *Curr. Biol.* **26**, R656–60 (2016).
25. Williamson, R. C., Doiron, B., Smith, M. A. & Yu, B. M. Bridging large-scale neuronal recordings and large-scale network models using dimensionality reduction. *Curr. Opin. Neurobiol.* **55**, 40–47 (2019).
26. Barlow, H. B. Unsupervised Learning. *Neural Comput.* **1**, 295–311 (1989).
27. Bialek, W. & Zee, A. Coding and computation with neural spike trains. *J. Stat. Phys.* **59**, 103–115

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(1990).

28. Diamantaras, K. I., Hornik, K. & Strintzis, M. G. Optimal linear compression under unreliable representation and robust PCA neural models. *IEEE Trans. Neural Netw.* **10**, 1186–1195 (1999).

29. Doi, E., Balcan, D. C. & Lewicki, M. S. A Theoretical Analysis of Robust Coding over Noisy Overcomplete Channels. in *Advances in Neural Information Processing Systems 18* (eds. Weiss, Y., Schölkopf, B. & Platt, J. C.) 307–314 (MIT Press, 2006).

30. Chalk, M., Marre, O. & Tkačik, G. Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 186–191 (2018).

31. Młynarski, W. F. & Hermundstad, A. M. Adaptive coding for dynamic sensory inference. *Elife* **7**, (2018).

32. Ganguli, D. & Simoncelli, E. P. Neural and perceptual signatures of efficient sensory coding. *arXiv [q-bio.NC]* (2016).

33. Pouget, A., Zhang, K., Deneve, S. & Latham, P. E. Statistically efficient estimation using population coding. *Neural Comput.* **10**, 373–401 (1998).

34. Attneave, F. Some informational aspects of visual perception. *Psychol. Rev.* **61**, 183–193 (1954).

35. Barlow, H. B. & Others. Possible principles underlying the transformation of sensory messages. *Sensory communication* **1**, 217–234 (1961).

36. Rieke, F., Bodnar, D. A. & Bialek, W. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. Biol. Sci.* **262**, 259–265 (1995).

37. Laughlin, S. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch. C* **36**, 910–912 (1981).

38. Dan, Y., Atick, J. J. & Reid, R. C. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J. Neurosci.* **16**, 3351–3362 (1996).

39. Baddeley, R. *et al.* Responses of neurons in primary and inferior temporal visual cortices to

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

natural scenes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **264**, 1775–1783 (1997).

40. Vinje, W. E. & Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).

41. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).

42. Marsat, G. & Maler, L. Neural heterogeneity and efficient population codes for communication signals. *J. Neurophysiol.* **104**, 2543–2555 (2010).

43. Onken, A., P P Chamanthi, Kayser, C. & Panzeri, S. Understanding Neural Population Coding: Information Theoretic Insights from the Auditory System. *Advances in Neuroscience* **2014**, 1–14 (2014).

44. Weliky, M., Fiser, J., Hunt, R. H. & Wagner, D. N. Coding of natural scenes in primary visual cortex. *Neuron* **37**, 703–718 (2003).

45. Ganguli, D. & Simoncelli, E. P. Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Comput.* **26**, 2103–2134 (2014).

46. Beyeler, M., Rounds, E. L., Carlson, K. D., Dutt, N. & Krichmar, J. L. Neural correlates of sparse coding and dimensionality reduction. *PLoS Comput. Biol.* **15**, e1006908 (2019).

47. Brinkman, B. A. W., Weber, A. I., Rieke, F. & Shea-Brown, E. How Do Efficient Coding Strategies Depend on Origins of Noise in Neural Circuits? *PLoS Comput. Biol.* **12**, e1005150 (2016).

48. Averbeck, B. B. & Lee, D. Effects of noise correlations on information encoding and decoding. *J. Neurophysiol.* **95**, 3633–3644 (2006).

49. Doi, E. & Lewicki, M. S. A simple model of optimal population coding for sensory systems. *PLoS Comput. Biol.* **10**, e1003761 (2014).

50. Dombeck, D. A., Harvey, C. D., Tian, L. & Looger, L. L. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nature* (2010).

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

51. Zhuang, J. *et al.* An extended retinotopic map of mouse cortex. *Elife* **6**, (2017).

52. Pnevmatikakis, E. A. *et al.* Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data. *Neuron* **89**, 285–299 (2016).

53. Harvey, C. D., Coen, P. & Tank, D. W. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* **484**, 62–68 (2012).

54. Morcos, A. S. & Harvey, C. D. History-dependent variability in population dynamics during evidence accumulation in cortex. *Nat. Neurosci.* **19**, 1672–1681 (2016).

55. Saleem, A. B., Diamanti, E. M., Fournier, J., Harris, K. D. & Carandini, M. Coherent encoding of subjective spatial position in visual cortex and hippocampus. *Nature* **562**, 124–127 (2018).

56. Krumin, M., Lee, J. J., Harris, K. D. & Carandini, M. Decision and navigation in mouse parietal cortex. *Elife* **7**, (2018).

57. Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).

58. Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N. & Harvey, C. D. Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell* **170**, 986–999.e16 (2017).

59. Tiganj, Z., Jung, M. W., Kim, J. & Howard, M. W. Sequential Firing Codes for Time in Rodent Medial Prefrontal Cortex. *Cereb. Cortex* **27**, 5663–5671 (2017).

60. Singh, I., Tiganj, Z. & Howard, M. W. Is working memory stored along a logarithmic timeline? Converging evidence from neuroscience, behavior and models. *Neurobiol. Learn. Mem.* **153**, 104–110 (2018).

61. Hattori, R., Danskin, B., Babic, Z., Mlynaryk, N. & Komiyama, T. Area-Specificity and Plasticity of History-Dependent Value Coding During Learning. *Cell* **177**, 1858–1872.e15 (2019).

62. Osborne, L. C., Palmer, S. E., Lisberger, S. G. & Bialek, W. The neural basis for combinatorial coding in a cortical population response. *Journal of Neuroscience* **28**, 13522–13531 (2008).

63. Stevens, C. F. A statistical property of fly odor responses is conserved across odors. *Proc. Natl.*

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acad. Sci. U. S. A. **113**, 6737–6742 (2016).

64. Malnic, B., Hirono, J., Sato, T. & Buck, L. B. Combinatorial receptor codes for odors. *Cell* **96**, 713–723 (1999).
65. Curto, C., Itskov, V., Morrison, K., Roth, Z. & Walker, J. L. Combinatorial neural codes from a mathematical coding theory perspective. *Neural Comput.* **25**, 1891–1925 (2013).
66. Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory*. (Psychology Press, 2005).
67. Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R. & Warland, D. Reading a neural code. *Science* **252**, 1854–1857 (1991).
68. Atick, J. J. & Redlich, A. N. What Does the Retina Know about Natural Scenes? *Neural Comput.* **4**, 196–210 (1992).
69. Cai, M. B., Schuck, N. W., Pillow, J. W. & Niv, Y. Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. doi:10.1101/347260
70. Wark, B., Lundstrom, B. N. & Fairhall, A. Sensory adaptation. *Curr. Opin. Neurobiol.* **17**, 423–429 (2007).
71. Dasgupta, S. & Gupta, A. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct. Algorithms* **22**, 60–65 (2003).
72. Morcos, A. S. & Harvey, C. D. History-dependent variability in population dynamics during evidence accumulation in cortex. *Nat. Neurosci.* **19**, 1672–1681 (2016).
73. Wei, Z., Inagaki, H., Li, N., Svoboda, K. & Druckmann, S. An orderly single-trial organization of population dynamics in premotor cortex predicts behavioral variability. *Nature Communications* **10**, (2019).
74. Lebedev, M. A. *et al.* What, if anything, is the true neurophysiological significance of 'rotational dynamics'? doi:10.1101/597419

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

75. Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).

76. O'keefe, J. & Nadel, L. *The hippocampus as a cognitive map*. (Oxford: Clarendon Press, 1978).

77. Pastalkova, E., Itskov, V., Amarasingham, A. & Buzsáki, G. Internally generated cell assembly sequences in the rat hippocampus. *Science* **321**, 1322–1327 (2008).

78. MacDonald, C. J., Lepage, K. Q., Eden, U. T. & Eichenbaum, H. Hippocampal 'Time Cells' Bridge the Gap in Memory for Discontiguous Events. *Neuron* **71**, 737–749 (2011).

79. MacDonald, C. J., Carrow, S., Place, R. & Eichenbaum, H. Distinct hippocampal time cell sequences represent odor memories in immobilized rats. *J. Neurosci.* **33**, 14607–14616 (2013).

80. Eichenbaum, H. On the Integration of Space, Time, and Memory. *Neuron* **95**, 1007–1018 (2017).

81. Lisman, J. E. Relating Hippocampal Circuitry to Function: Recall of Memory Sequences by Reciprocal Dentate–CA3 Interactions. *Neuron* **22**, 233–242 (1999).

82. Howard, M. W. *et al.* A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *J. Neurosci.* **34**, 4692–4707 (2014).

83. Williams, A. H. *et al.* Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron* **98**, 1099–1115.e8 (2018).

84. Bartal, Y., Recht, B. & Schulman, L. Dimensionality reduction: beyond the Johnson-Lindenstrauss bound. in *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms* 868–887 (Society for Industrial and Applied Mathematics, 2011).

85. National Research Council, Division on Earth and Life Studies, Institute for Laboratory Animal Research & Committee for the Update of the Guide for the Care and Use of Laboratory Animals. *Guide for the Care and Use of Laboratory Animals: Eighth Edition*. (National Academies Press, 2011).

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

86. Chen, T.-W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).

87. Dana, H. *et al.* Thy1-GCaMP6 transgenic mice for neuronal population imaging in vivo. *PLoS One* **9**, e108697 (2014).

88. Madisen, L. *et al.* Transgenic Mice for Intersectional Targeting of Neural Sensors and Effectors with High Specificity and Performance. *Neuron* **85**, 942–958 (2015).

89. Gorski, J. A. *et al.* Cortical excitatory neurons and glia, but not GABAergic neurons, are produced in the Emx1-expressing lineage. *J. Neurosci.* **22**, 6309–6314 (2002).

90. Aronov, D. & Tank, D. W. Engagement of neural circuits underlying 2D spatial navigation in a rodent virtual reality system. *Neuron* **84**, 442–456 (2014).

91. Garrett, M. E., Nauhaus, I., Marshel, J. H. & Callaway, E. M. Topography and areal organization of mouse visual cortex. *J. Neurosci.* **34**, 12587–12600 (2014).

92. Kalatsky, V. A. & Stryker, M. P. New paradigm for optical imaging: temporally encoded maps of intrinsic signal. *Neuron* **38**, 529–545 (2003).

93. Ratzlaff, E. H. & Grinvald, A. A tandem-lens epifluorescence macroscope: hundred-fold brightness advantage for wide-field imaging. *J. Neurosci. Methods* **36**, 127–137 (1991).

94. Juavinett, A. L., Nauhaus, I., Garrett, M. E., Zhuang, J. & Callaway, E. M. Automated identification of mouse visual areas with intrinsic signal imaging. *Nat. Protoc.* **12**, 32–43 (2017).

95. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).

96. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2**, 13 (2003).

97. Comaniciu, D., Ramesh, V. & Meer, P. Real-time tracking of non-rigid objects using mean shift. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)* doi:10.1109/cvpr.2000.854761

98. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: A Library for Large

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Linear Classification. *J. Mach. Learn. Res.* **9**, 1871–1874 (2008).

99. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

100. William H. Press, Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. (Cambridge University Press, 2007).

101. Henderson, H. V. & Searle, S. R. On deriving the inverse of a sum of matrices. *SIAM Rev.* **23**, 53–60 (1981).

102. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).

103. Eyherabide, H. G. & Samengo, I. When and why noise correlations are important in neural decoding. *J. Neurosci.* **33**, 17921–17936 (2013).

104. Panzeri, S., Macke, J. H., Gross, J. & Kayser, C. Neural population coding: combining insights from microscopic and mass signals. *Trends Cogn. Sci.* **19**, 162–172 (2015).

105. Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).

106. Chicharro, D. A Causal Perspective on the Analysis of Signal and Noise Correlations and Their Role in Population Coding. *Neural Computation* **26**, 999–1054 (2014).

107. Alink, A., Walther, A., Krugliak, A., van den Bosch, J. J. F. & Kriegeskorte, N. Mind the drift - improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. *bioRxiv* 032391 (2015). doi:10.1101/032391

108. Ramírez, F. M., Cichy, R. M., Allefeld, C. & Haynes, J.-D. The neural code for face orientation in the human fusiform face area. *J. Neurosci.* **34**, 12155–12167 (2014).

109. Henriksson, L., Khaligh-Razavi, S.-M., Kay, K. & Kriegeskorte, N. Visual representations are dominated by intrinsic fluctuations correlated between areas. *Neuroimage* **114**, 275–286 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

110. Penrose, R. A generalized inverse for matrices. *Math. Proc. Cambridge Philos. Soc.* **51**, 406–413 (1955).

111. Stewart, G. W. Matrix Algorithms: Basic Decompositions, vol. 1. *Society for Industrial and Applied Mathematics, Philadelphia, PA* (1998).

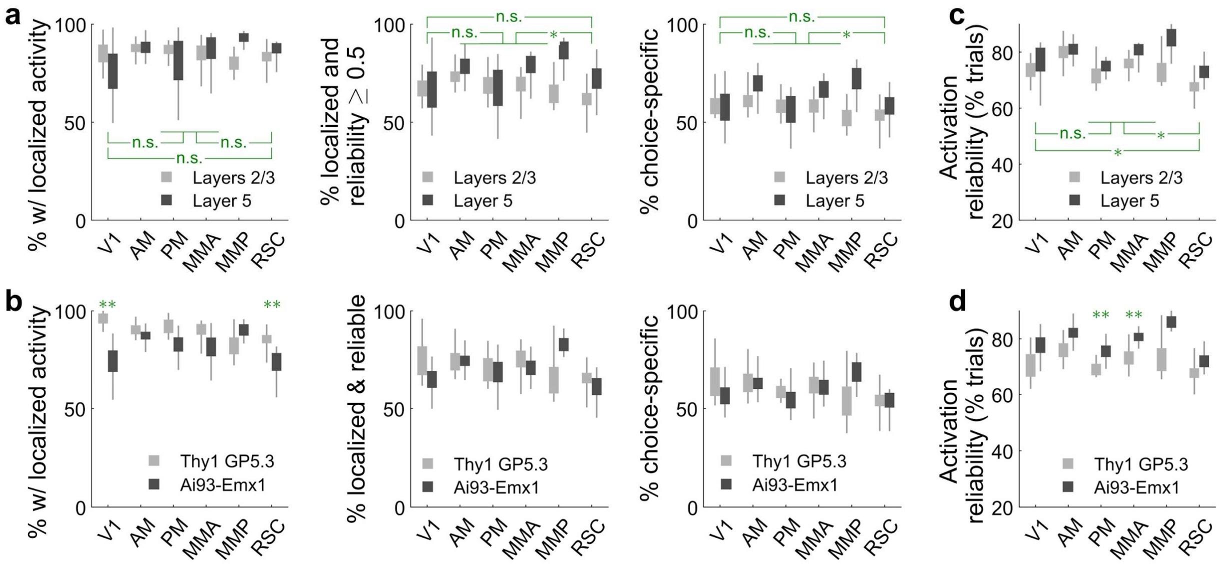
112. Higham, N. J. *Matrix nearness problems and applications*. (Citeseer, 1988).

Supplemental Information

	Layers 2/3						Layer 5					
	V1	**AM**	**PM**	**MMA**	**MMP**	**RSC**	**V1**	**AM**	**PM**	**MMA**	**MMP**	**RSC**
# sessions	9	18	11	15	8	30	8	12	6	9	7	12
# mice imaged	4	8	6	9	4	8	4	7	3	8	4	6

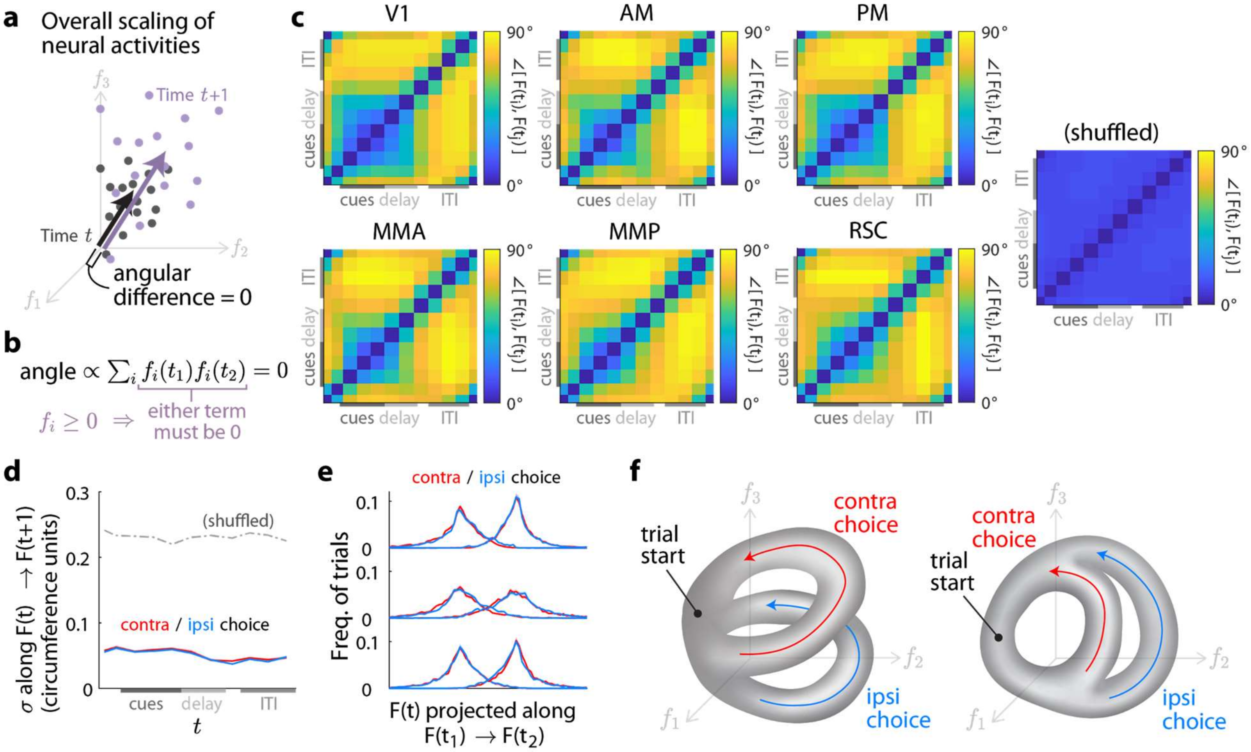
Supplementary Table 1. Number of imaging sessions and mice for various areas and layers, for the main experiment.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



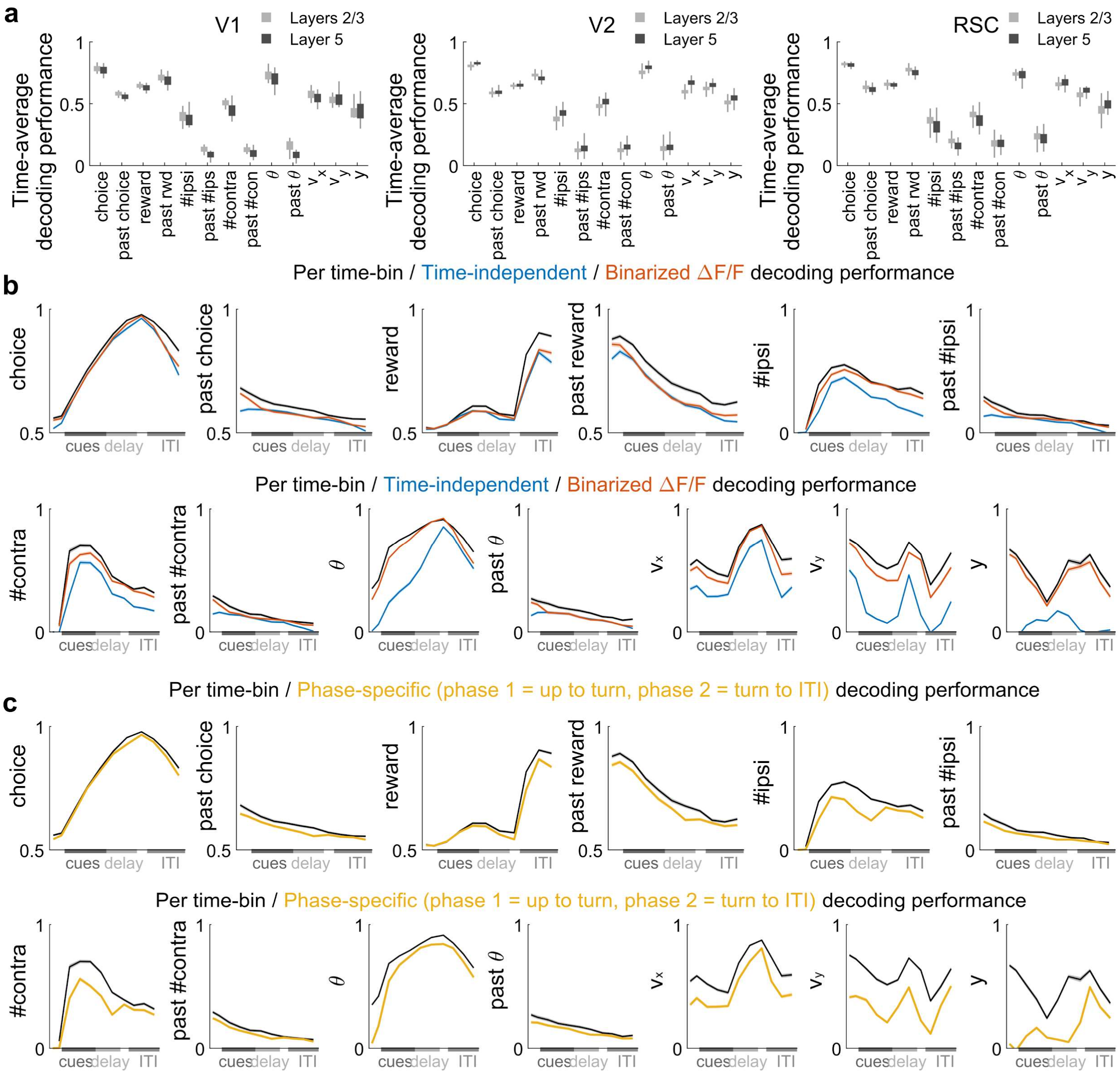
Supplementary Figure 1. ***Statistics for choice-specific sequences, and cross-strain comparison.*** **(a)** Percents of neurons that had significant ridge-to-background excess vs. a permutation test (left plot), and additionally were active within their firing fields in ≥ 50% of their (preferred-choice, if any) trials (middle plot), and additionally had different activity levels in right- vs. left-choice trials (right plot). Error bars: std. dev. across imaging sessions. Rectangles: Median and S.E.M. Stars: significant differences in means (Wilcoxon rank-sum test). **(b)** Like (a), but comparing two strains of mice. Data were pooled across layers. Double-stars indicate areas for which there was a significant difference in means (Wilcoxon rank-sum test). **(c)** Average reliability of choice-specific neurons in a given area/layer, defined as the fraction of trials in which the neuron was significantly active within its putative firing field . Error bars as in (a). **(d)** Like (c), but comparing two strains of mice. Data were pooled across layers. Double-stars indicate areas for which there was a significant difference in means (Wilcoxon rank-sum test).

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



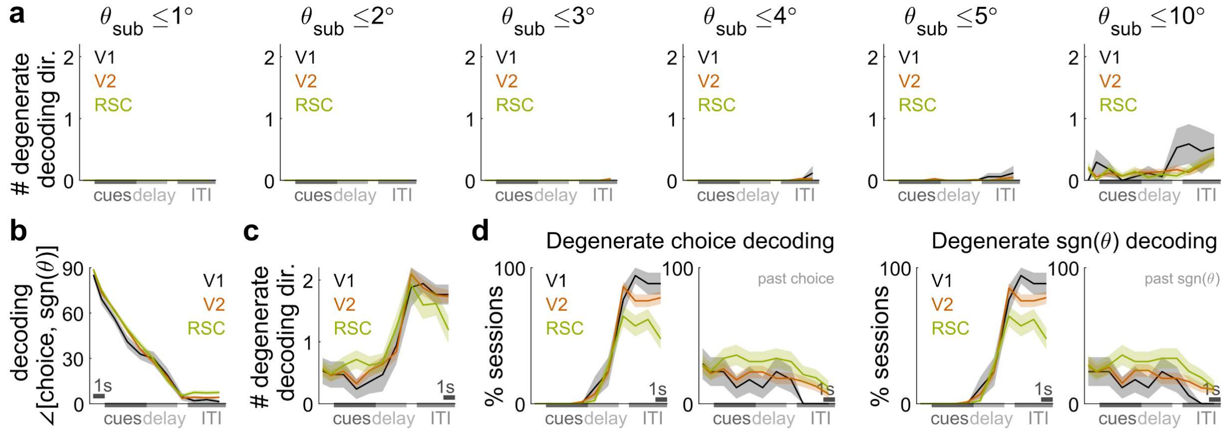
Supplementary Figure 2. ***Manifold geometry metrics for all pairs of timepoints.*** **(a)** Illustration of a case where the neural states across trials at time $t + 1$ are an overall scaling of the neural states at time t (i.e. by the same scale factor for the activity of each neuron). This results in zero angular difference between the centers of the two point clouds (black vs. purple arrows), because the vector to the center of the $t + 1$ point cloud is also just a scaling of the vector to the center of the t point cloud. **(b)** The angle between two vectors $\vec{f}(t_1)$ and $\vec{f}(t_2)$ is proportional to the dot product of the two vectors. If all components $f_i(t)$ of these vectors are nonnegative, then the only way for the angle to be zero is for all terms in the sum to be zero (since there are no negative terms, so they cannot cancel). This means that either $f_i(t_1)$ or $f_i(t_2)$ must be zero in the sum, which can be interpreted as a neuron activity f_i switching from active at t_1 to inactive at t_2, or vice versa (or this is a silent neuron). Across all terms in the sum a.k.a. the neural population, this means that the identities of active neurons must completely change between t_1 and t_2. **(c)** Angles between time-average neural state vectors $F(t_1)$ and $F(t_2)$, for all possible pairs of timepoints t_1 and t_2. imaging sessions for the stated areas were averaged for each plot. Shuffled: Pseudo-data with activity randomly shuffled per trial, per neuron. **(d-e)** Same as Fig. 2c-d, but plotted separately for trials with different eventual choices. The same projection axes were used regardless of trial type. **(f)** Illustration of how manifolds with global time parameters (arrowed curves) can have strong sub-structure in that (for example) trials of different choices bifurcate mid-trial and follow well-separated trajectories until the end of the trial, then gradually losing this distinction in the ITI.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



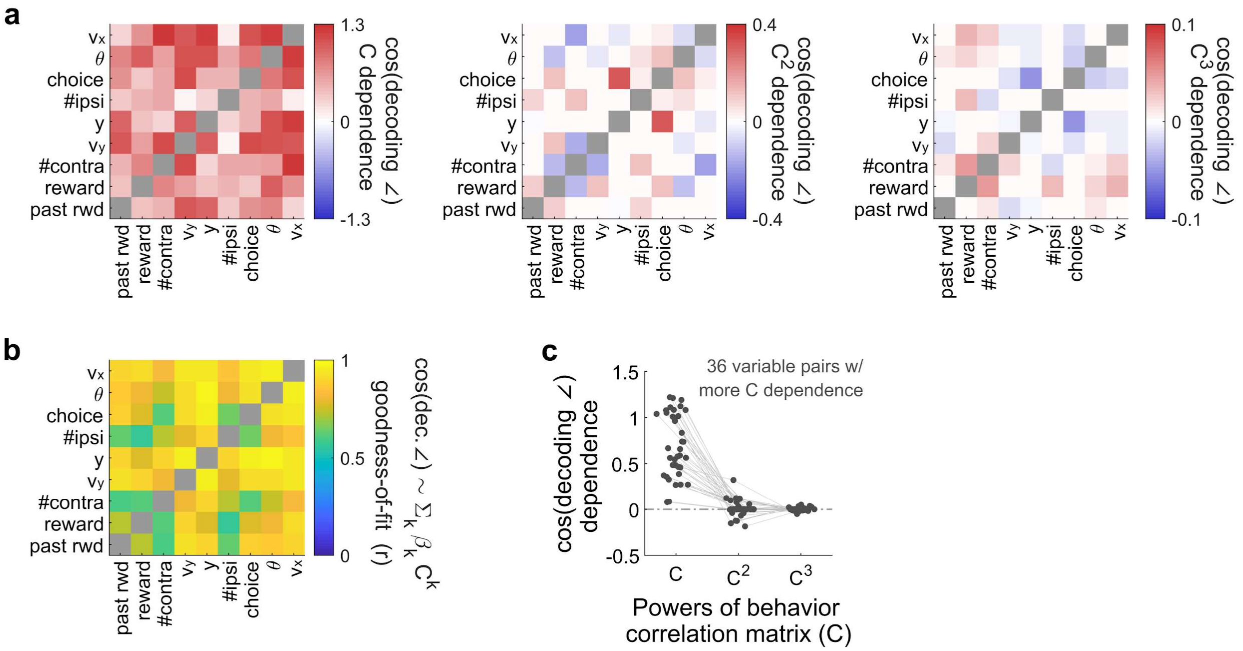
Supplementary Figure 3. ***Details of decoding accuracies for different types of decoding methods.*** **(a)** Time-average performance of various decoders (x-axis), shown separately for datasets in layer 2/3 vs. 5. Error bars: std. dev. across sessions. Rectangles: Median and S.E.M. **(b)** Decoding performance for alternative decoding methods as discussed in the text. Lines: mean across imaging sessions. Band: S.E.M. **(c)** Decoding performance for per-timepoint decoders used throughout the text and in (b), vs. phase-specific decoders (yellow lines). The phase-specific decoders were trained using data in 2 separate phases of the trial. The first phase included all timepoints from the start of the trial to the end of the delay region, and these timepoints were treated like additional trials when training the decoders. The second phase included the remaining timepoints from the start of the turn region to the end of the ITI. Decoding performances were evaluated at each timepoint, as usual. Lines: mean across imaging sessions. Band: S.E.M.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



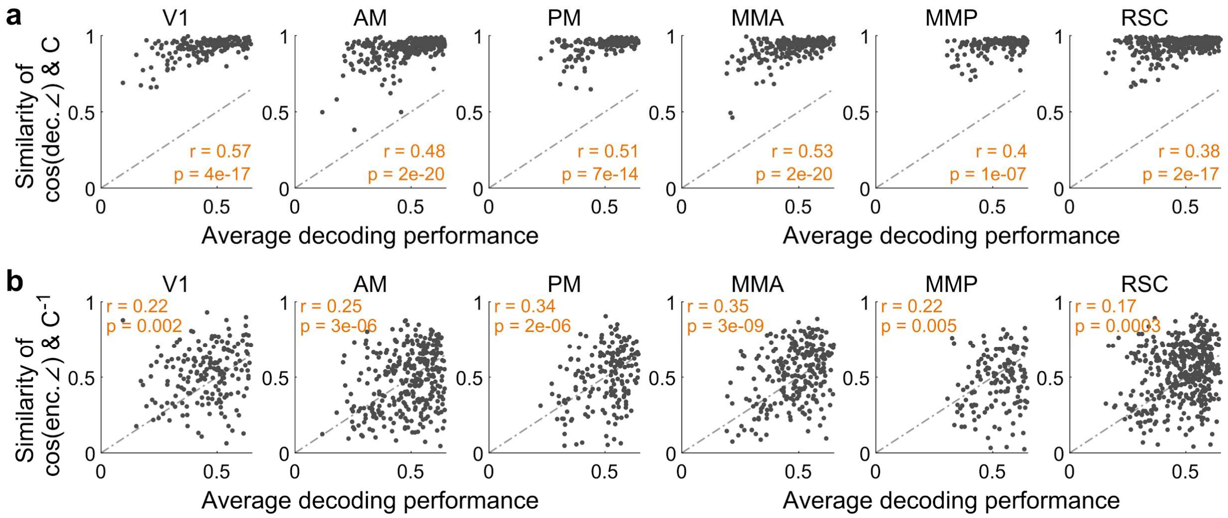
Supplementary Figure 4. ***Diagnoses for degenerate decoding directions.*** **(a)** Number of decoding directions vs. time that were identified as degenerate, i.e. having a near-zero angle w.r.t. the subspace spanned by the other decoding directions. These plots use six different angular thresholds (θ_{sub} as indicated at the top of the plots) for deciding whether the angle is "close enough" to zero. Lines: Mean across sessions. Bands: S.E.M. **(b)** Angle between choice and view-angle-sign decoders, vs. time in the trial. Lines: Mean across sessions. Bands: S.E.M. **(c)** Same as (b) but for a set of thirteen variables that includes the sign of the view angle in lieu of the continuously-valued view angle. **(d)** For the modified set of thirteen variables in (c), the percent of imaging sessions in which the choice or view-angle-sign decoding directions were identified as being degenerate.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



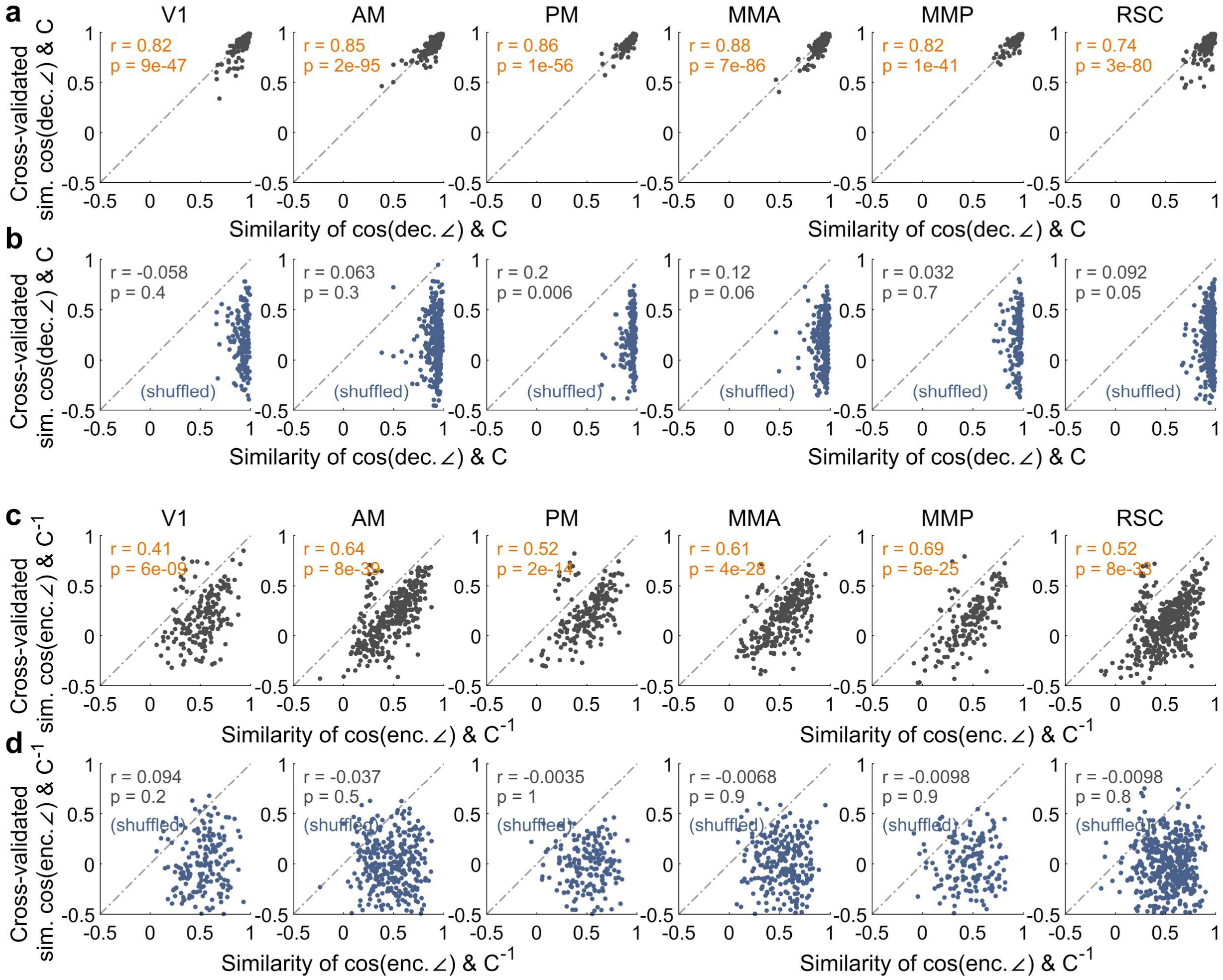
Supplementary Figure 5. ***Dependence of decoding angles on the behavioral correlation matrix.*** **(a)** Coefficients from a linear regression model of the cosine decoding angles as a function of powers of the behavioral correlation matrix C. This was computed separately for all pairs of task variables. Note the different color scales for each plot. **(b)** Goodness-of-fit (Pearson's correlation) for the regression model in (a). The more poorly fit variable pairs tend to involve past-trial quantities, which also had lower decoding performances. **(c)** Same information as (a) except that each upper-triangular matrix entry is plotted as a point for the indicated C^k dependence (i.e. as for insets of Fig. 5a). Lines link the three C^kdependence coefficients for the same variable pair. There were no variable pairs with a smaller magnitude of C than C^2 dependence.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



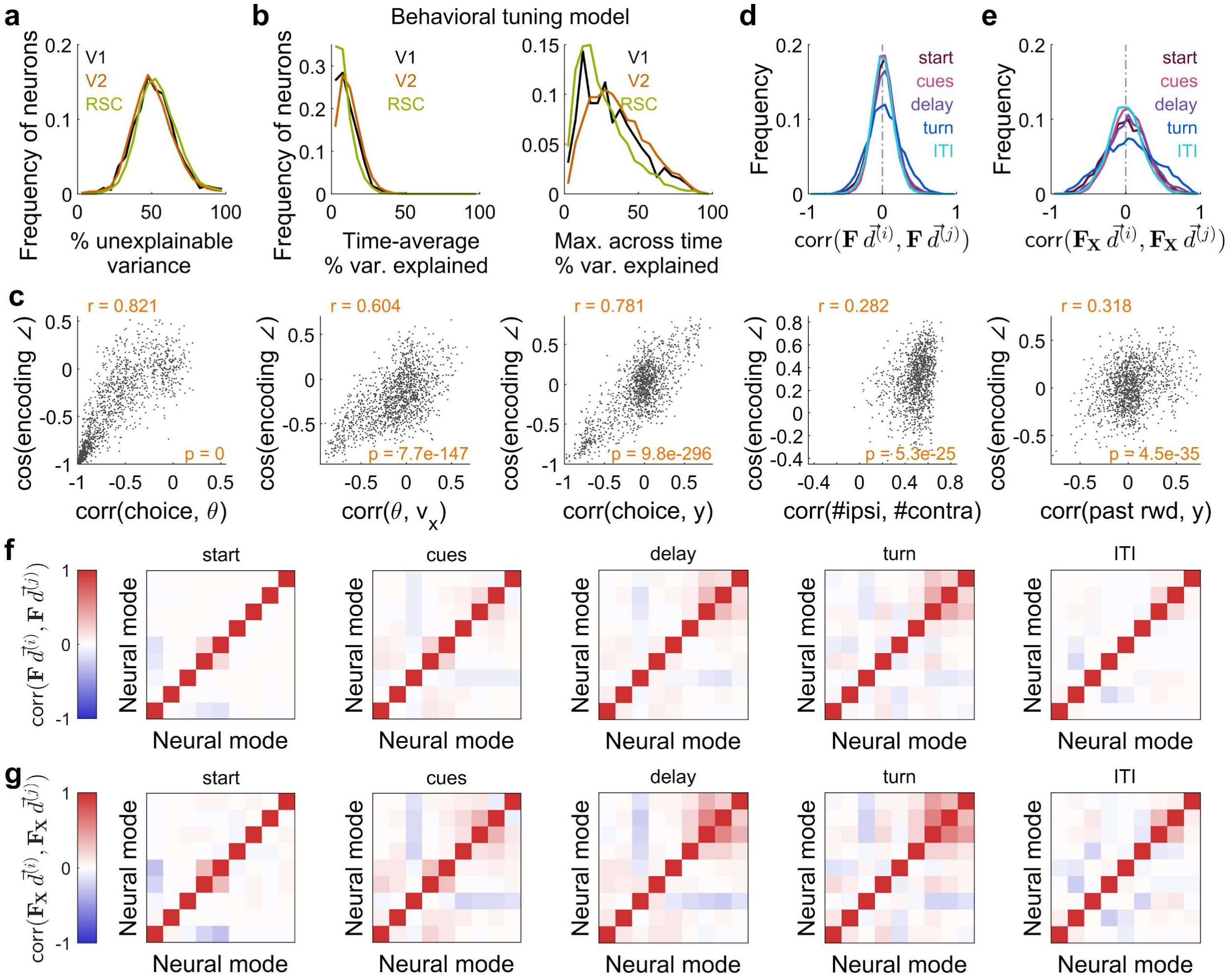
Supplementary Figure 6. ***Behavior-related structure of encoding/decoding geometry improves (or does not degrade) with signal-to-noise (SNR).*** **(a)** Similarity of the matrix of decoding angles to the matrix of task-variable correlations (y coordinate), vs. the average performance of task-variable decoders (x coordinate). Each imaging session contributes 11 data points in each plot, i.e. one per timepoint in the trial, and the various plots (columns) are for recordings in the stated cortical areas. Similarity scores (y coordinates) were computed as Pearson's correlation with data points being the upper-triangular elements of the two matrices, i.e. as in Fig. 5b. Decoding performances (x coordinates) were calculated separately for timepoint in the trial, as the average performance (Pearson's correlation between predicted and actual values) across task variables. To avoid averaging together random chance-level accuracies, before computing the average the decoding performance of a given variable was set to zero if it was not significant vs. a permutation test (cross-validated and corrected for multiple comparisons, as in Fig. 3). **(b)** Same as (a), except that the similarity scores (y coordinates) are for between the matrix of encoding angles and the inverse task-variable correlation matrix.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



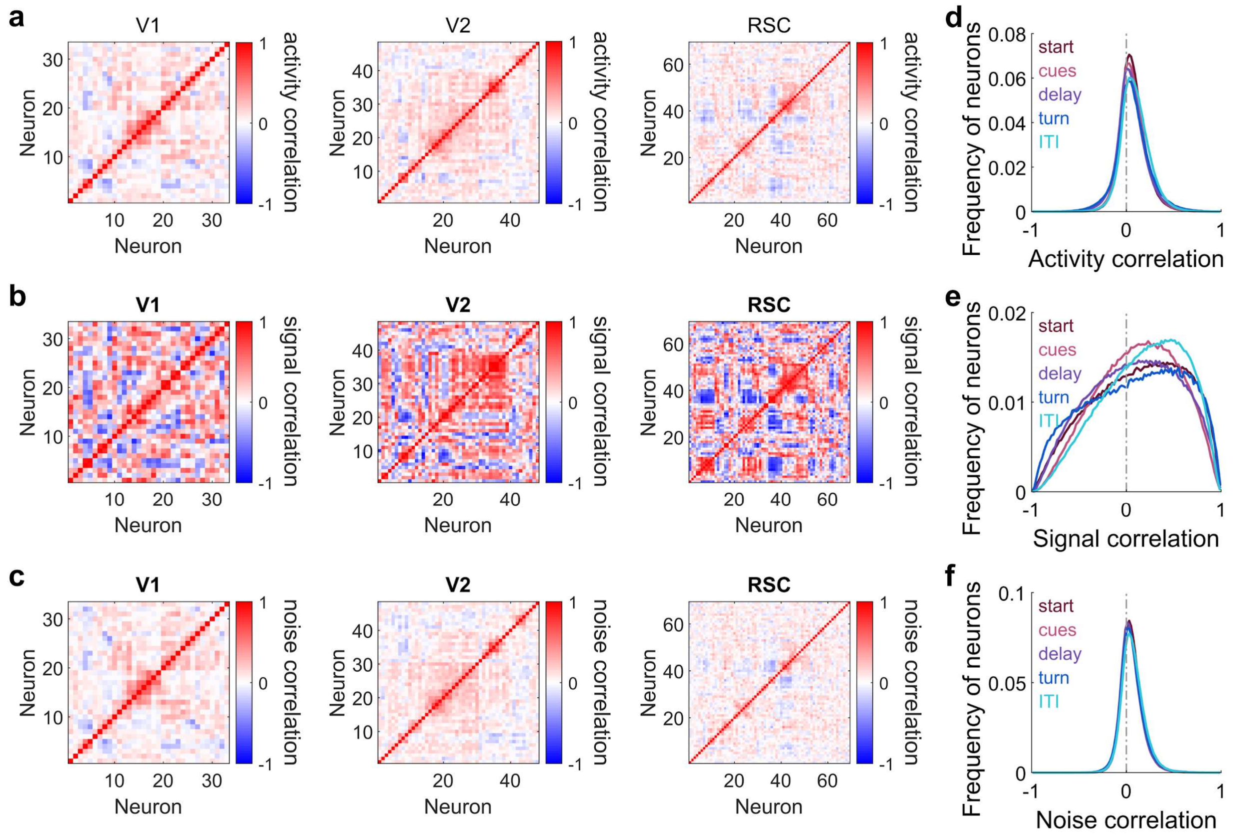
Supplementary Figure 7. ***Observations of decoding/encoding geometry are preserved under cross-validation.*** **(a)** Similarity (Pearson's correlation) score for how well the matrix of decoding angles matched the task-variable correlation matrix $\mathbf{C}$ (as in Fig. 5b), but comparing this similarity score in a cross-validated scenario (y-coordinate) vs. the nominal method used everywhere else (x-coordinate). For cross-validation, two sets of nine decoding directions were separately computed using two disjoint subsets of trials, so for task variable i we obtained two independent estimates of its decoding direction, $\vec{w}_1^{(i)}$ and $\vec{w}_2^{(i)}$. Angles between decoding directions for variables i and j were then computed as the average of $\cos\angle(\vec{w}_1^{(i)}, \vec{w}_2^{(j)})$ and $\cos\angle(\vec{w}_1^{(j)}, \vec{w}_2^{(i)})$. Each plot (columns) corresponds to data from imaging sessions for the stated cortical region. **(b)** Same as (a), but the y-coordinate of each data point was computed using data where neural responses were permuted across trials, breaking the correspondence between neural activity and task variables while retaining correlations between task variables. **(c-d)** Same as (a-b), but for encoding directions.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



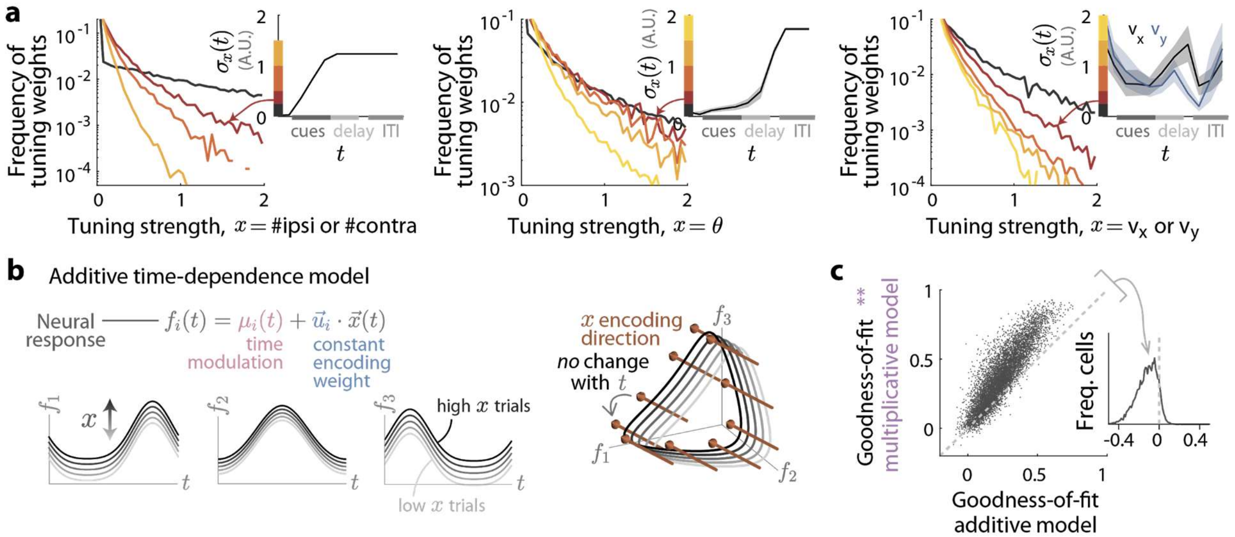
Supplementary Figure 8. ***Additional comparisons of encoding angles to inverse behavioral correlation matrix.*** **(a)** Distribution across neurons of the percent unexplainable variance. This was estimated per neuron as the variance of activity across trials with identical stimuli and behavioral outcome, relative to the variance across all trials. **(b)** Distribution across neurons of the variance explained by a per-timepoint linear behavioral response model. The left plot shows the time-average variance explained, whereas the right plot shows the maximum possible variance explained for that neuron, i.e. taking the maximum over time. **(c)** Scatter plot of the cosine angle between pairs of encoding directions vs. the corresponding entry in the inverse task-variable correlation matrix. Each data point corresponds to a timepoint within a recording session, all sessions included. **(d)** Distribution of correlation coefficients for pairs of neural modes in the information-coding subspace, for various timepoints in the trial. The neural modes are defined as the projection of the high-dimensional neural state $\mathbf{F}$ onto the various orthogonal basis vectors $\vec{d}^{(i)}$ for the subspace spanned by the encoding directions. **(e)** Same format as (d), but for the predicted neural signal $\mathbf{F_X} \equiv \mathbf{X}\widehat{\mathbf{W}}_{\text{enc}}$ instead of the full neural state $\mathbf{F}$, which includes noise. **(f)** Color scale: neural mode correlations as in (d). For each pair of neural modes (matrix entries), this correlation was averaged across imaging sessions and timepoints within the indicated period within the trial. For comparability across datasets, the basis vectors $\vec{d}^{(i)}$ were computed using polar decomposition of the encoding weight matrix, so that each $\vec{d}^{(i)}$ was as close as possible to one encoding direction in the least-squares sense[112]. Neural modes (rows and columns) were ordered using the order of their nearest encoding direction in Fig. 6f. **(g)** As in (f), but for the predicted neural signal as in (e).

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



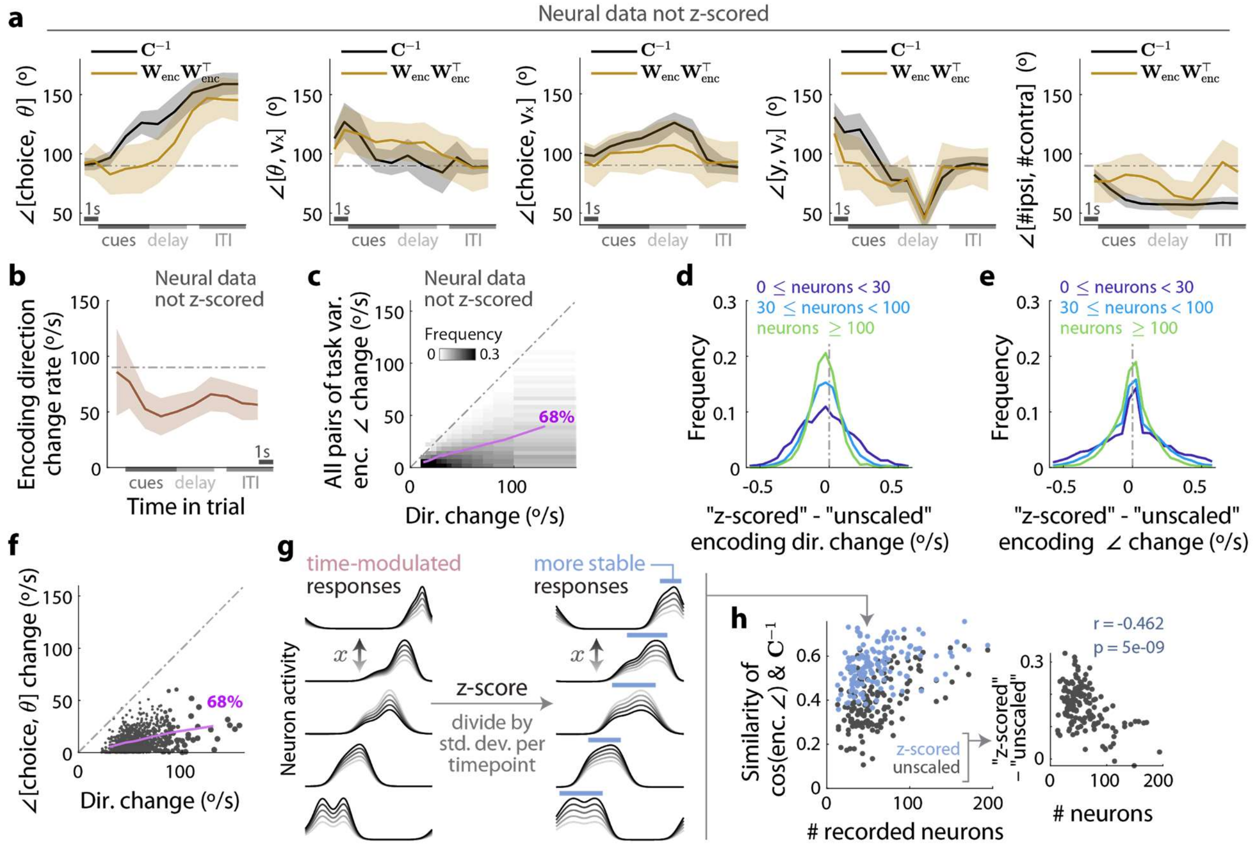
Supplementary Figure 9. ***Correlations between pairs of neurons.*** **(a)** Pairwise correlation coefficients for the activities of neurons across trials, evaluated at a fixed timepoint at the end of the cue region. The various plots are for three example imaging sessions in the stated brain areas, selected to have the median number of neurons across all imaging sessions for that region. Neurons (i.e. the displayed order of rows and columns) were sorted using hierarchical clustering of this matrix. **(b)** Same format and order of neurons as (a), but for signal correlations between neurons, defined as correlations between the predicted activities of neurons according to the per-timepoint behavioral encoding models. **(c)** Same format and order of neurons as (a), but for estimated noise correlations between neurons, where "noise" was defined as the residual activity of neurons after subtracting the behavior-based prediction used in (b). **(d-f)** Distributions of correlation coefficients as in **(a-c)**, for neurons pooled across all sessions but restricted to the stated time periods in the trial.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



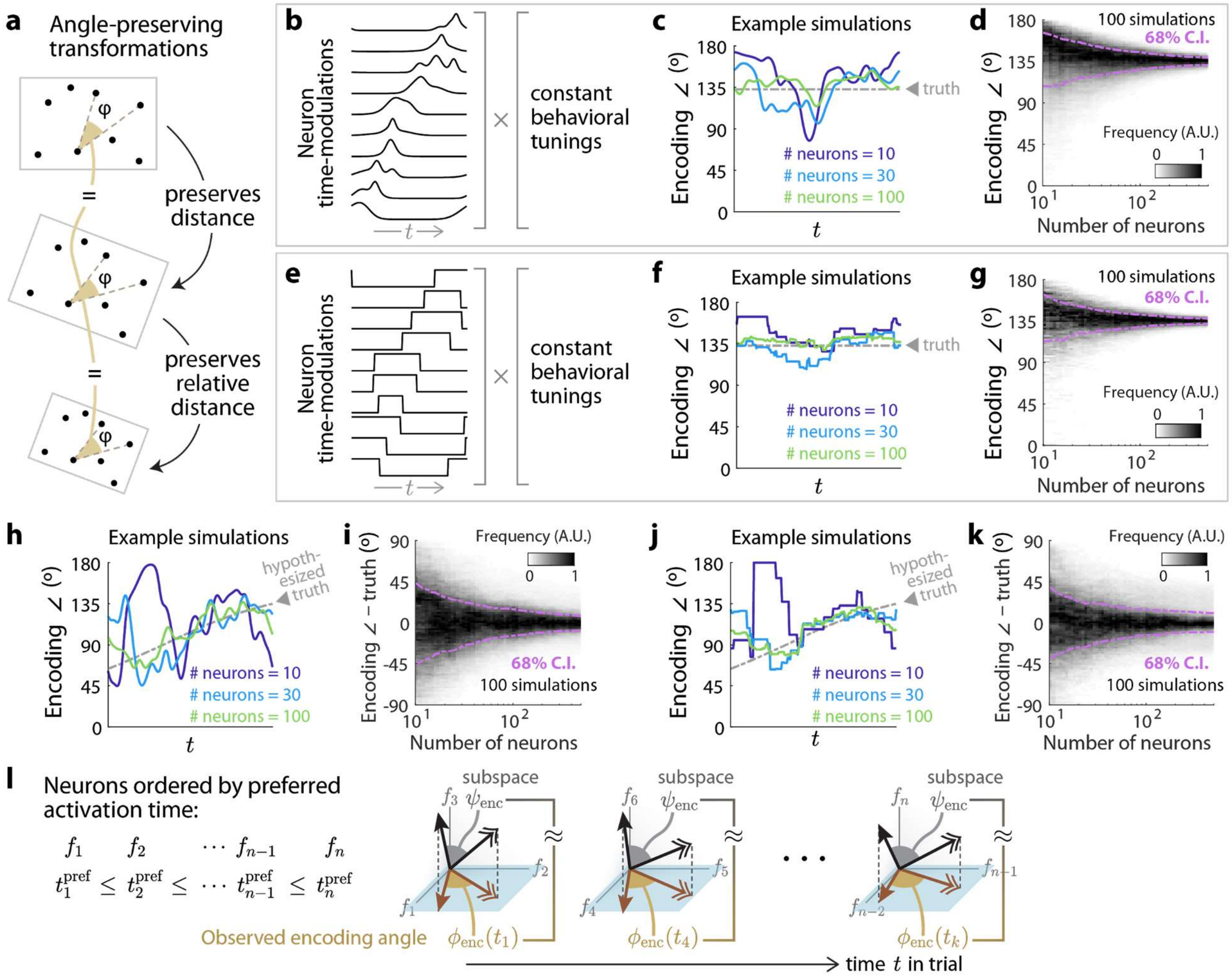
Supplementary Figure 10. ***Time-dependence of task-variable encoding weights and encoding geometry.*** **(a)** Dependence of encoding weights on the time-dependent scale of the respective task variables (insets). 11 encoding models were fit separately per timepoint for each neuron, which yields 11 encoding weights per neuron, for a given variable. The colored lines are distributions of these encoding weights restricted to timepoints in the trial where the time-dependent scale (standard deviation) of the behavioral variable fell within the indicated bins (vertical colored bars in the inset plot). For comparability across neurons and task variables, encoding weights were expressed in units of $\underline{\sigma_F}/\underline{\sigma_x}$, where $\underline{\sigma_F}$ is the standard deviation of the activity level of a given neuron across all timepoints in the imaging session, and $\underline{\sigma_x}$ is the standard deviation of the task variable again across all timepoints. **(b)** Simulation of three neurons with constant linear dependence on a task variable x and an additive, time-dependent baseline. Even though the neural activity was sequential and formed a ring-shaped manifold (right plot), there was no change in the encoding directions. **(c)** Same as Fig. 7e, but comparing the multiplicative model in Fig. 7e vs. the additive model in (b). The multiplicative model performed significantly better ($p\approx 0$ given statistics of our dataset, Wilcoxon rank-sum test).

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 11. ***Effect of per-timepoint z-scoring of neural data on encoding directions and angles.*** **(a-c)** Same as Fig. 7a-c, but using "unscaled" encoding models where the neural data had *not* been z-scored per timepoint. **(d)** Distribution of differences between the rate of angular change of encoding directions, for encoding models using z-scored neural data vs. unscaled neural data. Each imaging session contributes $9 \times 11 = 99$ points for encoding directions of 9 variables for each 11 timepoints. **(e)** As in (d) but for the rate of change of encoding angles. **(f)** Same as Fig. 7c, but for an example pair of task variables, i.e. the angle between choice and θ encoding directions (absolute value), vs. the rate of change of the encoding directions (each x-coordinate is the average of the change in the choice direction and the change in the θ direction). Each data point corresponds to one timepoint in one imaging session. Lines: 68% C.I. of encoding-angle change (y-coordinate), calculated in bins of the direction change rate. **(g)** Illustration of how z-scoring neural responses per timepoint corrects for some part of the time-modulation of behavioral responses, resulting in more stable encoding directions and angles around the peak activity time of each neuron. For this illustration, the standard deviation was computed as the spread across simulated responses to different x levels (grayscale curves per neuron), plus 0.1 to avoid division by zero and to show the effect of non-task-related variability. **(h)** Similarity scores (Pearson's correlation as in Fig. 6e) for how well cosine encoding angles matched cosine angles between columns of the inverse task-variable correlation matrix, vs. number of recorded neurons in the imaging session. Blue points are for encoding models constructed using neural data that was z-scored per timepoint as in the rest of the article, whereas black points are for alternative encoding models constructed using neural data that did *not* have this per-timepoint scaling. Right plot: the difference between z-scored (blue) and unscaled (black) points in the left plot, as a function of number of neurons.

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Supplementary Figure 12. ***Convergence of multiplicative sequential time-modulations to a stable per-timepoint encoding geometry, with large numbers of neurons.*** **(a)** Illustration of two transformations that preserve distances (rotation) and relative distances (uniform scaling) between points. Also shown is an angle between two vectors (dashed lines), which remains the same if distances are preserved, and also if relative distances are preserved, as angles do not depend on lengths of vectors. **(b)** Illustration of time-modulation functions for the simulations in (c) and (d). Each simulated neuron (rows) had a uniformly random time preference within $t = [0,1]$, and time-modulation $g_i(t)$ being the sum of 5 gaussian bumps randomly distributed around this time preference ($\sim N(\mu = 0, \sigma = 0.08)$). The width of each bump was drawn randomly $\sim N(\mu = 0.06, \sigma = 0.03)$ with a minimum of 0.02, and for simplicity we selected a scale such that the maximum over time of $g_i(t)$ is 1 for each neuron. As in Fig. 7b, these time-modulations multiply random, time-independent behavioral responses, so that each simulated (i^{th}) neuron's activity has the form $f_i(t) = g_i(t)\,[1 + \vec{u}_i \cdot \vec{x}]$ where $\vec{u}_i$ are constant weights for encoding "task variables" $\vec{x}$. The time-dependent contribution of the neuron to the population-level encoding direction for variable x_k is $\partial f_i(t)/\partial x_k = g_i(t)\,u_{ik}$. **(c)** Encoding angle vs. time for three simulated experiments described in (b). In each simulation with n neurons, two underlying (constant) encoding directions $\vec{U}^{(1)} \equiv [u_{11}\, u_{21}\, \cdots u_{n1}]^T$ and $\vec{U}^{(2)} \equiv [u_{12}\, \cdots\, u_{n2}]^T$ were generated randomly (entries $u_{ik} \sim N(\mu = 0, \sigma = 1)$), but constrained to have a 135° angle between $\vec{U}^{(1)}$ and $\vec{U}^{(2)}$. The per-timepoint encoding directions computed using the simulated neural activities $\{f_i(t)\}$ are $\vec{w}^{(1)}(t) = [g_1(t)\, u_{11}\, \cdots g_n(t)\, u_{n1}]$ and $\vec{w}^{(2)}(t) = [g_1(t)\, u_{12}\, \cdots g_n(t)\, u_{n2}]$, and have time-varying encoding angles as shown in this plot. **(d)** Distribution of encoding angles over 100 simulated experiments

bioRxiv preprint doi: https://doi.org/10.1101/801654; this version posted October 13, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as in (b), as a function of the number of simulated neurons. Each simulation contributes 201 timepoints to the distribution. **(e-g)** Same as (b-d), except that $g_i(t)$ of each neuron was set to 0 if $< 0.1 \times$ the maximum, and to 1 otherwise. **(h)** Same as (c), except that the underlying encoding directions were generated as $\vec{U}^{(1)} = [u_{11} \; \cdots \; u_{n1}]^T$ being random as before, but $\vec{U}^{(2)} = [v_1 \; \cdots \; v_n]^T + [(0.5 - 1.5\, t_1^{pref})\, u_{11} \; \cdots \; (0.5 - 1.5\, t_n^{pref})\, u_{n1}]^T$ where $v_i \sim N(\mu = 0, \sigma = 1)$ are random and $t_i^{pref} \equiv argmax_t\; g_i(t)$ is the peak activity time of the i^{th} neuron. The "hypothesized truth" for the time-dependence of encoding angles is based on the assumption that at time t, the neurons that contribute most to the encoding directions $\vec{w}^{(1)}(t)$ and $\vec{w}^{(2)}(t)$ are those with time preferences $t_i^{pref} = t$. Recalling that $g_i(t_i^{pref}) = 1$ by construction, then under this assumption $\vec{w}^{(1)}(t) \approx P_t\, \vec{U}^{(1)}$ where P_t is a projection matrix that zeroes out the rows of $\vec{U}^{(1)}$ for which $t_i^{pref} \neq t$ and leaves the same rows for which $t_i^{pref} = t$, and also $\vec{w}^{(2)}(t) \approx P_t\vec{v} + (0.5 - 1.5\, t)\, \vec{w}^{(1)}(t)$. Because both $P_t\vec{v}$ and $\vec{w}^{(1)}$ are random vectors, when there are sufficiently many neurons in the population (that are active at time t), then these two random high-dimensional vectors are likely to be orthogonal, $(P_t\vec{v}\,) \cdot \; \vec{w}^{(1)} \approx 0$. Since the entries of $\vec{v}$ and $\vec{U}^{(1)}$ were drawn from a normal distribution $N(\mu = 0, \sigma = 1)$, their norms are likely to be $|P_t\vec{v}| \approx \sqrt{n_t}$ and $|\vec{w}^{(1)}| \approx |P_t\vec{U}^{(1)}| \approx \sqrt{n_t}$, where n_t is the number of neurons with $t_i^{pref} = t$. Given all this, $|\vec{w}^{(2)}|^2 \approx |P_t\vec{v}|^2 + (0.5 - 1.5\, t)^2\, |\vec{w}^{(1)}|^2 \approx [1 + \; (0.5 - 1.5t)^2]\, n_t$. Lastly, the cosine angle between the encoding directions is $cos\, \angle(\vec{w}^{(1)}, \vec{w}^{(2)}) \approx (0.5 - 1.5\, t)\; (\vec{w}^{(1)} \cdot \vec{w}^{(1)}) \,/\, |\vec{w}^{(1)}|\, |\vec{w}^{(2)}| = (0.5 - 1.5\, t) \,/\, \sqrt{1 + (0.5 - 1.5t)^2}$. This formula for $\angle(\vec{w}^{(1)}, \vec{w}^{(2)})$ is shown as the "hypothesized truth" in the plot, for comparison to the empirically calculated angle vs. time between $\vec{w}^{(1)}(t) = [g_1(t)\, u_{11} \; \cdots g_n(t)\, u_{n1}]$ and $\vec{w}^{(2)}(t) = [g_1(t)\, u_{12} \; \cdots g_n(t)\, u_{n2}]$ for three simulated experiments (colored lines) with the stated number of neurons. The empirical calculation better resembles the hypothesized truth at large number of neurons. **(i)** Distribution of the difference between the empirically calculated encoding angle and the hypothesized truth as explained in (h), as a function of the number of simulated neurons. **(j-k)** Same as (h-i), but $g_i(t)$ of each neuron was set to 0 if $< 0.1 \times$ the maximum, and to 1 otherwise. **(l)** Conceptualization of how structure in the underlying constant encoding weights $\mathbf{U}_{\text{enc}}^{\top} \equiv [\vec{U}^{(1)} \;\; \vec{U}^{(2)} \;\; \cdots]$ that systematically varies e.g. with the time preference of neurons as in (h), can result in observed encoding angles that vary with time in the trial. We assume that there are n neurons denoted $f_1, f_2\,, \ldots,\; f_n$, which are ordered by time preference so that $t_1^{pref} \leq t_2^{pref} \leq \cdots \leq t_n^{pref}$. Each subplot illustrates a subset of $\mathbf{U}_{\text{enc}}$ in a 3-dimensional subspace, with coordinates being encoding weights for neurons 1-3, neurons 4-6, and so forth. The systematic differences in relationships between encoding weights u_{i1} and u_{i2} as a function of neuron time preference and therefore neuron index i means that the angle between the subsets of $\mathbf{U}_{\text{enc}}$ ("subspace ψ_{enc}") differs systematically across the 3 subplots. A scenario where the observed encoding directions can vary rapidly in time yet exhibit slow changes in encoding angles, is when $g_i(t)$ are sharply peaked around t_i^{pref} whereas the abovementioned systematic differences in $\mathbf{U}_{\text{enc}}$ vary slowly as a function of neuron index. In the illustration, this corresponds to $\mathbf{U}_{\text{enc}}$ being mostly randomly oriented in each subplot, despite gradual changes in the projected angle ψ_{enc} across subplots. The fast changes in encoding directions corresponds to $g_i(t)$ acting effectively as projections of $\mathbf{U}_{\text{enc}}$ onto the (f_1, f_2) plane, the (f_2, f_3) plane, and so forth as a function of time in the trial. With many neurons i.e. high dimensions, the projected $\mathbf{U}_{\text{enc}}$ can be highly similar to the randomly oriented $\mathbf{U}_{\text{enc}}$ per subplot according to the Johnson-Lindenstrauss theorem, but these projections will also reflect the gradual changes in $\mathbf{U}_{\text{enc}}$ across subplots.