Contribution of behavioural variability to representational drift

Sadra Sadeh[1*] and Claudia Clopath[1*]

[1]Bioengineering Department, Imperial College London, London SW7 2AZ, United Kingdom

*Correspondence: C.C. (c.clopath@imperial.ac.uk), or S.S. (s.sadeh@imperial.ac.uk)

Abstract

Neuronal responses to similar stimuli change dynamically over time, raising the question of how internal representations can provide a stable substrate for neural coding. Recent work has suggested a large degree of drift in neural representations even in sensory cortices, which are believed to store stable representations of the external world. While the drift of these representations is mostly characterized in relation to external stimuli, the behavioural state of the animal (for instance, the level of arousal) is also known to strongly modulate the neural activity. We therefore asked how the variability of such modulatory mechanisms can contribute to representational changes. We analysed large-scale recording of neural activity from the Allen Brain Observatory, which was used before to document representational drift in the mouse visual cortex. We found that, within these datasets, behavioural variability significantly contributes to representational changes. This effect was broadcasted across various cortical areas in the mouse, including the primary visual cortex, higher order visual areas, and even regions not primarily linked to vision like hippocampus. Our computational modelling suggests that these results are consistent with independent modulation of neural activity by behaviour over slower time scales. Importantly, our analysis suggests that reliable but variable modulation of neural representations by behaviour can be misinterpreted as representational drift, if neuronal representations are only characterized in the stimulus space and marginalised over behavioural parameters.

Introduction

Neuronal responses to stimuli, contexts or tasks change over time, creating a drift of representations from their original patterns (Deitch et al., 2021; Driscoll et al., 2017; Lütcke et al., 2013; Marks and Goard, 2021; Schoonover et al., 2021; Ziv et al., 2013). This representational drift can reflect the presence of intrinsic noise or plasticity in the circuitry and, depending on its origin, can be detrimental to or beneficial for the neural code (Clopath et al., 2017; Rule et al., 2019). Understanding the mechanisms contributing to the emergence of representational drift can therefore shed light on its relevance for neural computation (Lütcke et al., 2013; Rule et al., 2019).

Representational drift can arise from a variety of sources including bottom-up mechanisms, like changes in the feedforward input to neurons or from a dynamic reorganization of recurrent interactions in the network. Another important source of variability that can contribute to representational drift is changes in the behavioural state of the animal. Spontaneous behaviour has in fact been shown to heavily modulate responses in awake behaving animals (Musall et al., 2019; Niell and Stryker, 2010; Stringer et al., 2019). Drift of behavioural state – e.g. gradual changes in attention, arousal or running – can therefore change the way neural activity is modulated by top-down mechanisms (Niell and Stryker, 2010; Vinck et al., 2015) over different timescales.

The exact manner in which such top-down mechanisms modulate the neural activity (Cohen-Kashi Malina et al., 2021; Dipoppa et al., 2018; Fu et al., 2014; Garcia del Molino et al., 2017; Pakan et al., 2016) would in turn determine how behavioural variability affects the representational drift. One possibility is that stimulus-evoked responses are just scaled by arousal or running, as suggested by gain models (Ferguson and Cardin, 2020). Under this scenario, the behavioural state of the animal can modulate the similarity of sensory representations across multiple repeats of the same stimulus (representational similarity), by increasing or decreasing the signal-to-noise ratio. Another possibility is that the behaviour contributes independently to neuronal activity, and hence representational similarity is better described in a parameter space where internal and external parameters conjointly define the neural code. Under the latter scenario, variability in behavioural "signal" could be perceived as noise from the viewpoint of sensory representations, and could therefore be potentially mistaken as representational drift.

To delineate the contribution of behavioural variability to representational drift and to shed light on the involved mechanisms, we analysed publicly available datasets from the Allen Brain Observatory (de Vries et al., 2020; Siegle et al., 2021). These datasets provide a good opportunity to systematically address this question as standardized visual stimulation and large-scale recordings of population activity in response to multiple repeats of the same stimuli are combined with systematic measurements of behavioural parameters (like running speed and pupil size) across a

large number of animals. Our analysis suggested that changes in the behavioural state of the animal can strongly modulate the similarity of neuronal representations in response to multiple repeats of the same stimulus. In fact, within the datasets analysed, a significant fraction of what may be described as representational drift in a sensory cortex can be attributed to behavioural variability. Our results suggest that the contribution of behavioural variability to changes in neuronal activity should be carefully quantified and considered in the analysis of representational similarity and representational drift.

Results

Representational similarity depends on the behavioural state of the animal

We analysed publicly available, large-scale, standardized *in vivo* physiology datasets recently published by the Allen Brain Observatory (Siegle et al., 2021). The electrophysiology datasets obtained via Neuropixels probes (Jun et al., 2017) provide the possibility of studying the spiking activity of a large number of units to visual stimuli (see Methods). We studied similarity of neural activity in response to multiple repeats of the same natural movie (Figure 1a).

Previous studies have reported significant changes in stimulus-evoked representations even in sensory cortices, over the time scales of multiple hours to multiple days and weeks (Deitch et al., 2021; Marks and Goard, 2021; Schoonover et al., 2021). The Neuropixels electrophysiology datasets provide the opportunity of studying these representational changes while accounting for behavioural changes, although over a faster time scale (hours). Similar representational drift has in fact been reported for another dataset (obtained via two-photon calcium imaging) over the course of multiple days (Deitch et al., 2021). The insights obtained from this analysis may therefore help in understanding the mechanisms underlying representational drift over longer time scales.

To shed light on the involved mechanisms, we contrasted two potential scenarios (Figure 1b). Changes in population activity in response to the same stimulus can arise from a completely random and independently added noise. Alternatively, modulation of activity by other (stimulus-independent) factors like behavioural modulation can also contribute to these changes (Figure 1b). To delineate between these two scenarios, we characterized how neuronal representations change across repetitions of the same stimulus (Figure 1a,b). This was quantified by a measure of representational similarity (RS), which was characterized as the similarity of responses, at the population level, to multiple repeats of the stimulus (Methods and Figure 1—figure supplement 1). Representational drift was calculated as the difference of RS within and between multiple repeats of the natural movies (see Methods).

Metrics of representational drift (e.g., representational drift index, RDI (Marks and Goard, 2021); see Methods) are often calculated from RS. We therefore based our analysis primarily on comparing RS and its relationship with sensory and behavioural parameters, as this contained more information. RDI was calculated as a single metric to quantify representational drift between blocks of presentation in each session. The drop in representational similarity between individual pairs of movie repeats was used to analyse representational drift on a finer scale (this drop is mathematically related to representational drift; see Methods). The main behavioural parameters we analysed were the size of pupil and the running speed of animals. We refer to more slowly changing dynamics of these parameters (compared to sensory-evoked responses) as the behavioural state of the animal, and changes in these states are described as behavioural variability. Our analysis was performed in two datasets with different structure of stimulus presentations (Figure 1c,f; see Supplementary Table 1). In each dataset, the natural movie (30 second long) is presented multiple times in each block of presentation (10 and 30 repeats for dataset1 and dataset2, respectively). We analysed the data for two blocks of presentation separated by more than an hour (Figure 1c,f). For each presentation of the natural movie, we calculated a population vector of responses from the average activity of all the units recorded in the primary visual cortex (V1), in bin widths of 1 second starting from the onset of movie presentation (Methods). Representational similarity between two repeats of the natural movie was quantified by the correlation coefficient of the population vectors (Figure 1—figure supplement 1c and Methods).

Previous analysis has shown that representational similarity (as defined above) is higher within a block of presentation, and decreases significantly between different blocks, both in Neuropixels and calcium imaging datasets (Deitch et al., 2021). Our results confirmed this observation, but we also found that representational similarity is strongly modulated by the behavioural state of the animal. This was most visible in sessions where the behavioural state (as assayed by pupil diameter and the average running speed) changed clearly between the two repeats of the movie (Figure 1c,f). We observed that, firstly, change in the behavioural state strongly reduced the representational similarity between the two blocks (Figure 1d,g), reminiscent of the representational drift which has been reported over the scale of hours to days (Deitch et al., 2021; Marks and Goard, 2021; Schoonover et al., 2021). Secondly, increased pupil diameter and running during the second block of presentation in fact increased the similarity of responses to the same movie within that block (Figure 1d,g, left). Overall, there was a significant drop of representational similarity between the movie repeats in which the animal experienced the most changes in the average pupil size (Figure 1d,g, right). These results indicate that the behavioural state of the animal can bidirectionally modulate the representational similarity across repeats of the same stimulus.

We found similar dependence of representational similarity on the pupil change for other sessions (Figure 1--figure supplement 2a) and across all animals (Figure 1e,h). The effect was more prominent when focusing on movie repeats with significant changes in the average running (Figure 1e,h, left, black lines). Similar trend was also observed when considering units from all recorded regions, instead of only focusing on V1 units (Figure 1—figure supplement 2b). We also observed the same trend when repeating the analysis within blocks (Figure 1d-h, right, grey lines, and Figure 1—figure supplement 2b), although the drop of representational similarity across blocks was more prominent due to more drastic behavioural changes between the blocks, which is expected from the slow timescale of changes in behavioural states.

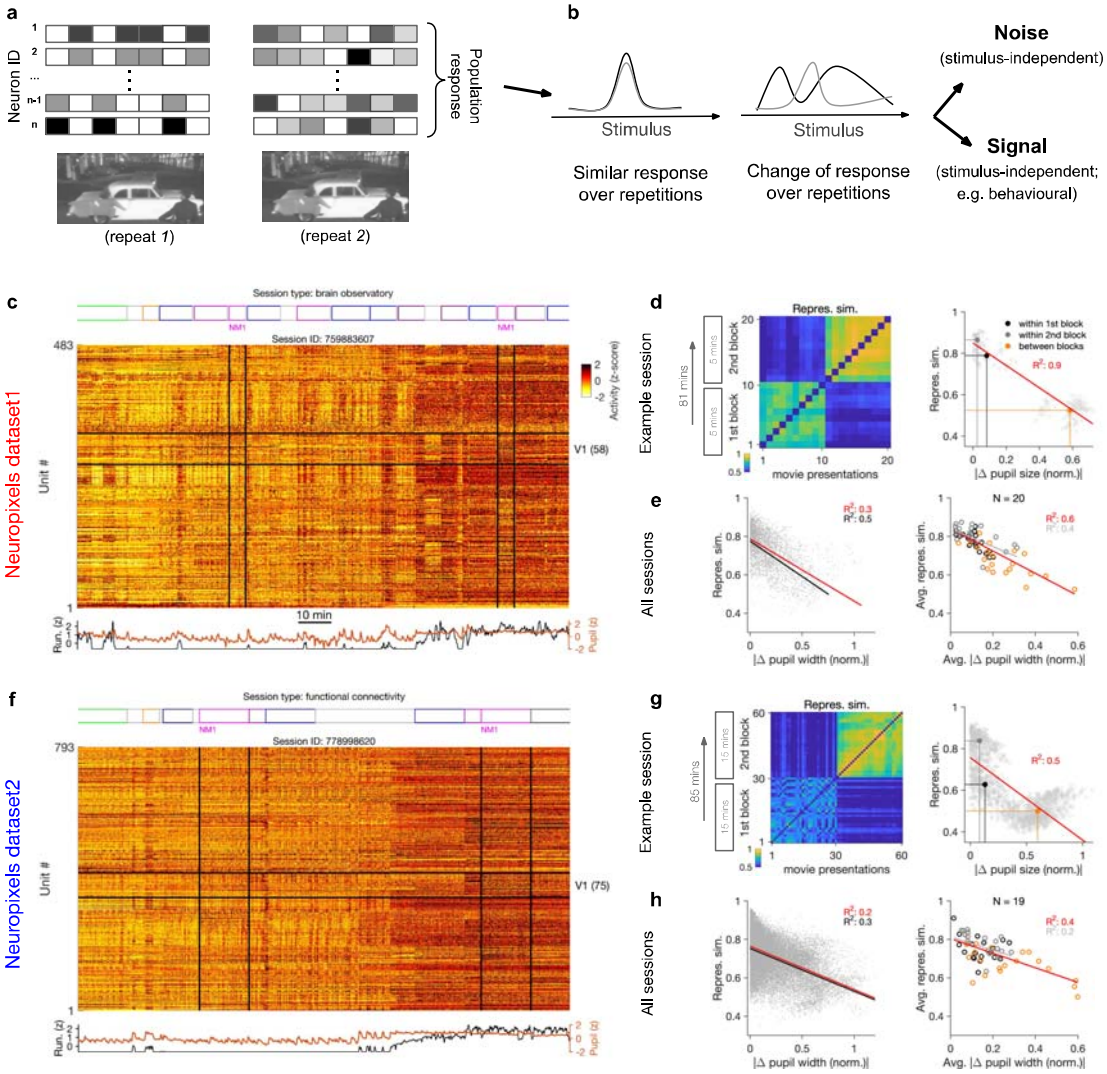


Figure 1: Representational similarity depends on the behavioural state of the animal.

a, Illustration of the response of a population of neurons to a stimulus (e.g., a natural movie) which is shown twice (left versus right). **b,** Neuronal response to different repetitions of the same stimulus can remain the

same (left) or change from the original pattern (right), leading to a change in neural representations. This change can arise from an added stimulus-independent noise, which is randomly and independently changing over each repetition. Alternatively, it can be a result of another signal, which is modulated according to a stimulus-independent parameter (e.g., behaviour), that is changing over each repetition. **c**, Population activity composed of activity of 483 units in an example recording session in response to different stimuli (timing of different stimulus types denoted on the top). Spiking activity of units is averaged in bins of 1 second and z-scored across the entire session for each unit. Units in primary visual cortex (V1; 58 units) and the two blocks of presentation of natural movie 1 (NM1) are highlighted by the black lines. Bottom: Pupil size and running speed of the animal (z-scored). **d**, Representational similarity between different presentations of natural movie 1. It is calculated as the correlation coefficient of vectors of population response of V1 units to movie repeats (see Methods). Left: The matrix of representational similarity for all pairs of movie repeats within and across the two blocks of presentation. Right: Representational similarity as a function of the pupil change, which is quantified as the normalized absolute difference of the average pupil size during presentations (see Methods). The best fitted regression line (using least squares method) and the R squared value (R^2) are shown. Filled circles show the average values within and between blocks. **e**, Same as (**d**, right) for all recording sessions. Left: Data similar to (**d**, grey dots) are concatenated across all mice and the best fitted regression line to the whole data is plotted. Black line shows the fit when movie repeats with significant change in the average running speed of the animal is considered (80th percentile). Right: The average values within and between blocks (filled circles in (**d**)) are plotted for all mice and the fitted regression line to these average values is plotted. Grey lines and R^2 values indicate the fit to within-block data only. N: number of mice. **f-h**, Same as (**c-e**) for a different dataset. Source data (for normalized changes in pupil width and representational similarity between pairs of movie repeats) are provided for individual sessions across the two datasets (Figure 1—source data 1).

In the above analysis, we considered the actual spiking activity of the units to build the population vectors. Calculating the representational similarity from these vectors can potentially bias the estimate by increasing the impact of highly active neurons. For instance, if the units which are firing higher remain consistently active, they may lead to some similarity of population vectors even independent of stimulus-evoked responses. To control for variance in the average activity of units, we repeated our analysis for population vectors composed of z-scored responses (as shown in Figure 1c,f; see Methods). Overall, representational similarity diminished when calculated from the z-scored activity (Figure 1—figure supplement 3a). However, we observed the same trend in terms of dependence on the behavioural state, whereby larger changes in pupil size were correlated with larger changes in representational similarity (Figure 1—figure supplement 3a).

We performed our original analysis with vectors of activity rendered in time bins of 1 second, corresponding to the interval of presentation of each frame of the natural movie (Methods). We tested whether our results hold for responses rendered in different time bins, by testing longer time

bins (2 seconds; Figure 1—figure supplement 3b). Overall, representational similarity was higher, as expected from averaging the activity. However, we observed similar drop of representational similarity with increases in pupil size (Figure 1—figure supplement 3b).

Our previous analyses were performed on wild-type mice as well as mice from three different transgenic lines (Pvalb-IRES-Cre × Ai32, n=8; Sst-IRES-Cre × Ai32, n=12; and Vip-IRES-Cre × Ai32, n=8; see Supplementary Table 1) (Siegle et al., 2021). To control for possible differences between different strains, we repeated our analysis for recording sessions in wild type mice only and observed similar results (Figure 1—figure supplement 4a). Our results also held when analysing female and male animals separately (female mice comprised a smaller fraction of the datasets; ~20%) (Figure 1—figure supplement 4b).

In our results above, we observed large changes in representational similarity between blocks of stimulus presentation with strong behavioural changes. But this effect can be confounded by the passage of time between the blocks, which may lead to other sources of variability such as changes in the excitability of neurons for instance. To account for this, we analysed the average block-wise representational similarity in individual animals (Figure1—figure supplement 5). For each animal, we quantified the representational drift index (RDI) between the two blocks as the normalized change of average correlations within and across blocks (Marks and Goard, 2021) (see Methods).

If the passage of time contributes strongly to the drop in representational similarity between the two blocks, we should see comparable levels of representational drift across animals with different levels of behavioural variability. Conversely, if the change in behaviour is the main factor contributing to the drop in representational similarity between the two blocks, we should see stronger levels of representational drifts for animals with larger behavioural variability. We indeed found evidence in favour of the latter: representational similarity remained rather stable for those animals which did not show large behavioural changes between the two blocks (Figure 1—figure supplement 5a). That is, passage of time per se did not contribute strongly to representational drift. Largest representational drifts were observed for animals with the largest changes in the average pupil width between the two blocks (Figure 1—figure supplement 5a). In fact, there was a good correlation between the two across animals (Figure 1—figure supplement 5b).

The relationship was weaker in Neuropixels dataset2, whereby more repeats of the natural movie are shown in each block (60 repeats for the total of 30 minutes, versus 20 repeats for 10 minutes in Neurpixels dataset1). Longer blocks of stimulus presentation increases the chance of behavioural changes within the blocks, which can in turn make the average block-wise representational similarity a less reliable metric. In line with this reasoning, further scrutiny into an outlier (with small average pupil changes but rather large representational drift between the two blocks; Figure 1—

figure supplement 5b, right) revealed that changes in the running speed of the animal within each block can be the factor contributing to changes in representational similarity (Figure 1—figure supplement 5c).

Taken together, these results suggest that, in awake behaving animals, variability of the behavioural state of the animal can be an important factor contributing to the modulation of representational similarity.

Evidence for independent modulation of responses by stimulus and behaviour

What is the mechanism by which behavioural state modulates representational similarity? Changes in pupil area are correlated with the level of arousal (Bradley et al., 2008), which can modulate the neuronal gain (Cohen-Kashi Malina et al., 2021). We therefore studied a possible gain model in which changes in pupil size modulate the neuronal responses to sensory inputs (Figure 2a; Methods). Alternatively, rather than scaling the stimulus-induced signal, behaviour can contribute independently to neuronal activity (Dipoppa et al., 2018; Saleem et al., 2013). We therefore compared the gain model to a model in which the neural tuning was obtained by an independent mixing of stimulus and behavioural signals (Figure 2b; Methods).

In each model, we calculated representational similarity from the neuronal responses to presentations of the same stimulus, and plotted that against the relative behavioural parameter (B) across the repeats (B_i/B_j, for the i-th and j-th repeats) (Figure 2c,d). Both models showed, on average, a similar dependence of representational similarity on relative behaviour (Figure 2c,d; the gain model only showed the same pattern if the signal was scaled by the behaviour; we observed different patterns, if behaviour scaled the noise, or both the signal and the noise; Figure 2—figure supplement 1a,b).

To compare different models with the experimental results, we took the relative pupil size as a proxy for relative behaviour and plotted the representational similarity of all units against it (Figure 2e). This revealed a similar average dependence as the signal-gain model and the independent-mixing model (Figure 2c-e). We observed a similar dependence for both datasets, and for most individual recording sessions within each dataset (Figure 2—figure supplement 1c-f). Similar results were observed when representational similarity was calculated from V1 units or all recorded units (Figure 2—figure supplement 1c-f).

We then asked how, at the level of individual units, the modulations of responses by stimulus and behaviour are related to each other (Figure 2f,g). To this end, instead of calculating representational similarity at the population level, we quantified the similarity of individual units' responses to multiple repeats of the stimulus (stimulus reliability; see Methods and Figure 1--figure supplement 1d). In the signal-gain model, stimulus reliability was highly correlated with behavioural modulation

of units (Figure 2f). This is a consequence of the scaling of the signal by the behaviour, which implies that neurons with higher signal component also show higher modulation with the behavioural parameter (see Methods). The signal-gain model therefore predicts that neurons which are strongly modulated by the stimulus also show strong modulation by the behaviour (Figure 2f). In contrast, the independent-mixing model predicted an independent relationship between stimulus and behavioural modulation of individual units (Figure 2g).

We tested these predictions in experimental data, by calculating behavioural modulation and stimulus reliability of individual units in all mice across both datasets. Behavioural modulation was calculated as the correlation of each unit's activity with pupil size, and stimulus reliability was obtained as the average correlation of each unit's activity vectors across multiple repeats of the natural movie (Methods and Figure 1—figure supplement 1d). As opposed to the signal-gain model, we did not observe a correlation between stimulus and behavioural modulation (Figure 2h). In fact, a regression analysis suggested that the two modulations are independent of each other in both datasets (Figure 2h), consistent with the independent-mixing model. The marginal distributions matched better with experimental distributions when we increased noise (x1.5 N) or decreased the behavioural signal (x0.5 B) (Figure 2—figure supplement 1g)

Overall, there was a wide distribution of stimulus reliability (Figure 2—figure supplement 2a) and behavioural modulation (Figure 2—figure supplement 2c) across recorded units, with patterns highly consistent across the two datasets. Most V1 units showed variable responses to repeats of the natural movie, as indicated by the peak of the distribution at low stimulus reliability (Figure 2—figure supplement 2a). However, the distribution had a long tail composed of units with high stimulus reliability, which showed highly reliable responses across repeats of the movie (Figure 2—figure supplement 2a,b). There was a wide spectrum of behavioural modulation too, with most units showing positive correlations with pupil size (Figure 2—figure supplement 2c,d), and a smaller population of negatively modulated units (Figure 2—figure supplement 2c).

The units that showed significant modulation with the stimulus were not necessarily modulated strongly by the behaviour, and vice versa; in fact, it was possible to find example units from all four combinations of weak/strong x stimulus/behavioural modulations (Figure 2—figure supplement 2e,f). A clear example of the segregation of stimulus and behavioural modulation was observed in CA1, where the units showed, on average, very weak stimulus reliability across movie repeats, consistently across different mice and datasets (Figure 2—figure supplement 3a). However, they were largely modulated by behaviour, to an extent comparable to V1 units (Figure 2—figure supplement 3a-c). Taken together, these results suggest that, rather than scaling the stimulus-

evoked responses, behaviour modulates the activity in a more independent and heterogeneous manner.

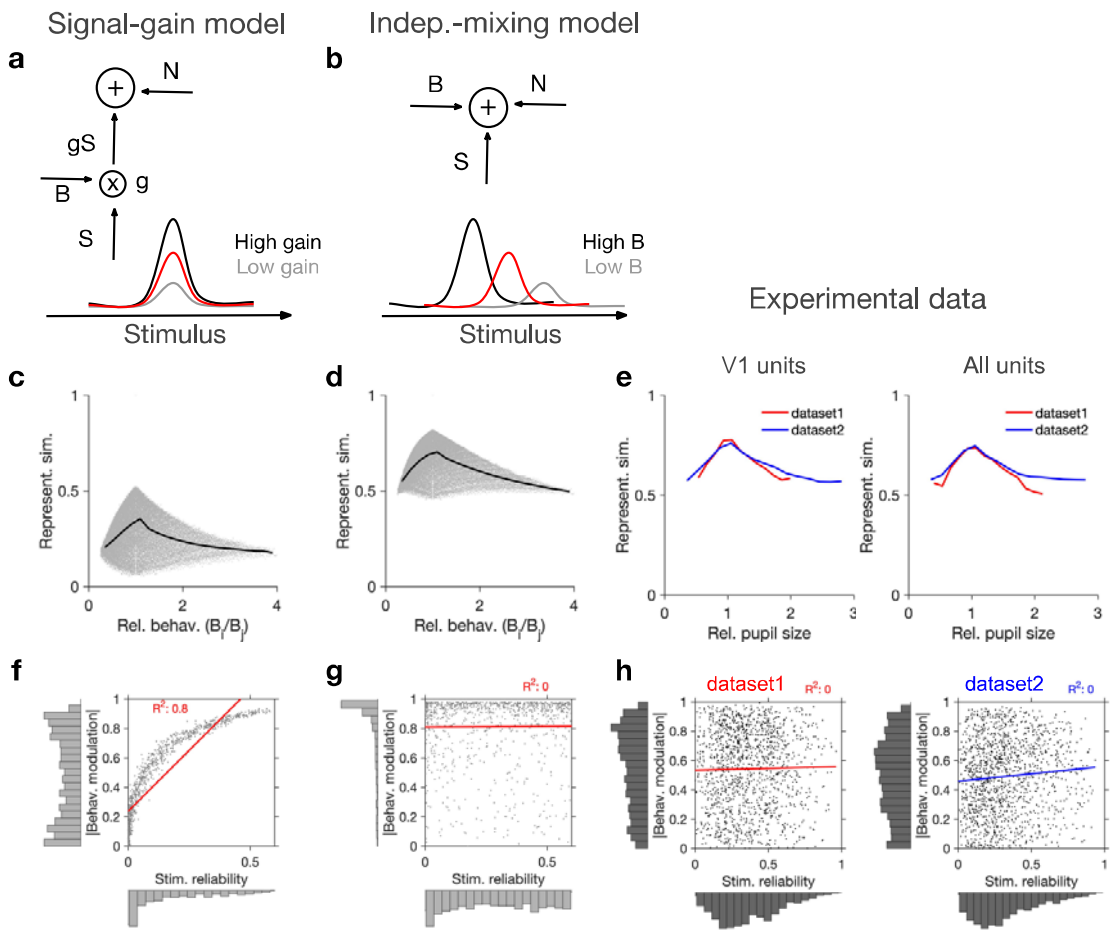


Figure 2: Independent modulation of activity by stimulus and behaviour.

a, Schematic of a signal-gain model in which behaviour controls the gain with which the stimulus-driven signal is scaled. Individual units are driven differently with the stimulus, leading to different tuning curves which determines their stimulus signal, S. Behavioural parameter, B, sets the gain, g, with which the stimulus signal is scaled, before being combined with the noise term, N, to give rise to the final response. S is the same across repetitions of the stimulus, while N is changing on every repeat (see Figure 2—figure supplement 1a and Methods). **b**, An alternative model (independent-mixing) in which the response of a unit is determined by the summation of its independent tuning to stimulus, S (red) and behaviour, B (black: high B, grey: low B), combined with noise, N (see Methods for details). **c**, Representational similarity of population responses to different repeats of the stimulus as a function of the relative behavioural parameter (B_i/B_j) in the signal-gain model. Black line shows the average (in 20 bins). **d**, Same as c, for the independent-mixing model. **e**, Same as (c,d) for the experimental data from Neuropixels dataset1 (red) or dataset2 (blue). For each pair of movie repeats, the average representational similarity of population responses (left: V1 units; right: all units) are

plotted against the relative pupil size (P_i/P_j). (**f,g**) Relation between behavioural modulation and stimulus reliability of units in different models. The stimulus signal-gain model predicts a strong dependence between behavioural modulation and stimulus reliability of units (f), whereas the independent-mixing model predicts no correlation between the two (g). Best fitted regression lines and R^2 values are shown for each case. Marginal distributions are shown on each axis. **h,** Data from V1 units in the two datasets show no relationship between the stimulus reliability of units and their absolute behavioural modulation, as quantified by the best fitted regression lines and R^2 values. Stimulus reliability is computed as the average correlation coefficient of each unit's activity vector across repetitions of the natural movie, and behavioural modulation is calculated as the correlation coefficient of each unit's activity with the pupil size (see Methods). Marginal distributions are shown. Source data (for stimulus reliability of V1 units and their modulation by pupil size) are provided for individual sessions across the two datasets (Figure 2—source data 1).

Behavioural variability modulates the low-dimensional components of population activity independent of stimulus reliability

If the behavioural state of the animal modulates the neuronal responses independently of the stimulus, it should be possible to see its signature in the low-dimensional space of neural activity. To test this, we analysed the principal components (PCs) of population responses in individual sessions (Figure 3; see Methods). For the two example sessions we analysed previously (shown in Figure 1c,f), the first two PCs explained a significant fraction of variance (Figure 3a,g). Low-dimensional population activity showed a distinct transition between two behavioural states, which were corresponding to low versus high arousal, as quantified by different pupil sizes (Figure 3b,h). The first PC, which explained most of the variance was strongly correlated with both pupil size and running speed (Figure 3—figure supplement 1a). These results suggest that behavioural modulation contributes significantly to the low-dimensional variability of neural activity.

To link the low-dimensional modulation of activity by behaviour to single neurons, we next analysed the projection of units' activity over the PCs by looking at PC loadings. Individual units were projected in the PC space according to their respective PC loadings (PC projections). Visualizing the average activity of units in the space of PC projections suggested a spectrum of weakly- to highly-active units (Figure 3—figure supplement 1b). In fact, neural projections over the first two PCs were correlated with the average activity of neurons (Figure 3—figure supplement 1c). In contrast to the average activity, the PC projections did not reveal any relationship with the stimulus reliability of units (Figure 3c,i), suggesting that the low-dimensional neural activity is modulated independently of stimulus-evoked modulation of responses.

These results were remarkably consistent across different datasets and across difference mice. The first two PCs explained similar levels of variance across more than 20 mice in each dataset (Figure

3d,j). The distance over PC1 was highly correlated with the difference of pupil size on the trajectory (Figure 3e,k). Importantly, this relationship also held when controlled for the passage of time (Figure 3e,k and Figure 3—figure supplement 2). In both datasets, the regression analysis revealed no relationship between the two PC projections and the stimulus reliability of units (Figure 3f,l; see Figure 3—figure supplement 1d for individual sessions). We therefore conclude that behaviour significantly modulates the low-dimensional components of neural activity, but this modulation does not specifically enhance the activity of neurons which are more reliably representing the stimulus.

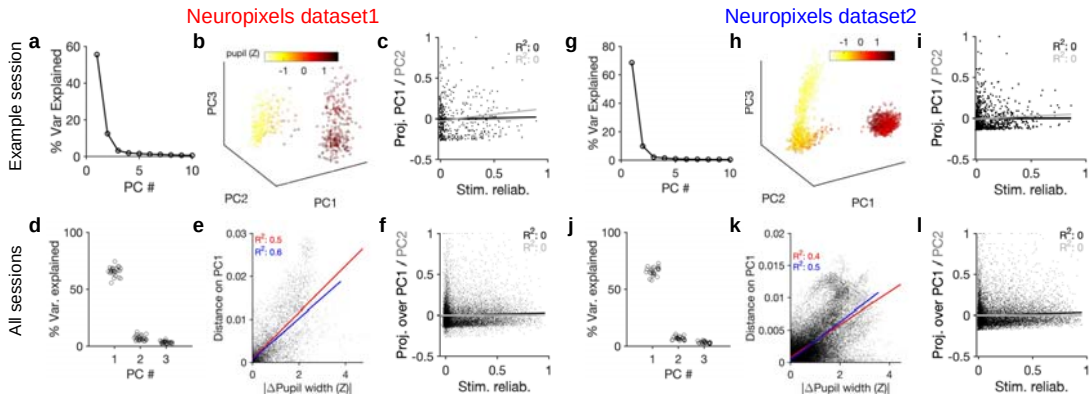


Figure 3. Behavioural variability modulates the low-dimensional components of population activity independent of stimulus reliability.

a, Relative contribution of the first 10 principal components (PCs) of population responses to the variability of activity (quantified by the fraction of explained variance by each PC) for an example session (same as shown in Figure 1). **b**, Population activity in the low-dimensional space of the first three PCs (see Methods for details). Pseudo colour code shows the pupil size at different times, indicating that the sudden transition in the low-dimensional space of activity is correlated with changes in the behavioural state. **c**, Projection of units' activity over PC1 (black) or PC2 (grey) (respective PC loadings) versus stimulus reliability of the units. It reveals no correlation between the two, as quantified by best fitted regression lines in each case (the best fitted regression lines and R^2 values shown by respective colours). PC loadings are normalized to the maximum value in each case. **d**, Fraction of variance explained by PCs 1-3 for all sessions. **e**, Distance on PC1 versus the difference of the average pupil size between different time points on the trajectory, for all sessions. For each pair of non-identical time points on the PC trajectory, the absolute difference of PC1 components is calculated as the distance on PC, and plotted against the absolute difference of the average pupil width (normalized by z-scoring across the entire session). Best fitted regression line and the corresponding R^2 value are shown in red for all data points, and in blue for data points with small time difference between the pairs (see Figure 3—figure supplement 2). **f**, Same as (**c**) for all units from all sessions in dataset1. **g-l**, Same as (**a-f**) for Neuropixels dataset2.

Behaviour modulates the setpoint of responses

To gain further insight into how the behaviour modulates the low-dimensional pattern of population activity, we analysed the relation between behavioural parameters and the average activity of units. In the two example sessions analysed previously (shown in Figure 1), there was a transition in the average pupil size and running speed in the second block, which was correlated with an overall increase in the average population activity (Figure 4a,e and Figure 4—figure supplement 1a). In general, change in the pupil size explained a significant fraction of changes in population activity of V1 units in all sessions (Figure 4—figure supplement 1b).

We also looked at the average activity of individual units across all movie frames and repetitions (their setpoints). Units had a wide range of setpoints, which were relatively stable within each block (small variance across repetitions relative to the mean) (Figure 4—figure supplement 1c). However, the setpoints changed upon transition to the next block, with most units increasing their setpoints, without an apparent link to their previous setpoint levels (Figure 4—figure supplement 1d). The population vectors composed of setpoints in each repeat can be further used to quantify setpoint similarity (Figure 4b,f). Within-block correlations were high, indicating the stability of setpoints when behavioural changes were minimum – although occasional, minor changes of pupil size still modulated these correlations (Figure 4b,f). Most changes in setpoint similarity, however, appeared between the blocks, when the animal experienced the largest change in its behavioural state.

Quantifying the dependence of setpoint similarity on changes in pupil size revealed a strong relationship, both for V1 units and for all recorded units (Figure 4c,g). The relationship was rather stable when calculated from responses to single frames of movie presentation, instead of the average activity across the entire movie (Figure 4—figure supplement 1e). We obtained similar results when the dependence was calculated from the average block activity across all mice, from both datasets (Figure 4d,h).

The relationship of setpoint similarity with changes in pupil size was not always monotonic. For the example session in Figure 4g, despite an overall drop of setpoint similarity with increased changes in pupil width, another trend was observed for large pupil changes. To understand what underlies this nonmonotonic trend we plotted setpoint similarity as a function of changes in both pupil size and running speed (Figure 4—figure supplement 2). The multi-dimensional space of changes in behaviour suggests that different trends exist at different regimes of behavioural changes (Figure 4—figure supplement 2b). Notably, limiting our analysis to the pairs of movie repeats with small changes in running speed revealed a monotonic relationship between setpoint similarity and changes in pupil size, with similar levels of regression between the two parameters (Figure 4—figure supplement 2a-c).

The opposite trend emerged at higher levels of changes in pupil size and running speed, and seemed to be related to changes in pupil size in the first block of presentation (Figure 4—figure supplement 2d,e). Indeed, movie presentations with higher (/lower) average pupil size in the first block showed smaller (/larger) setpoint similarity with the movie repeats in the second block, on average (Figure 4—figure supplement 2d,e). This trend was accompanied, and might be explained, by an average movement in the centre of the pupil between the two blocks (Figure 4—figure supplement 2f). Such a movement can shift the receptive fields, and therefore change the setpoint activity of units, leading to changes in setpoint similarity between the two blocks. This effect is expected to be higher for higher pupil sizes, and therefore the setpoint similarity is specifically reduced for movie presentations with larger pupil sizes in the first block.

The resulting change in the average visual input may also explain why the nonmonotonic trend is specifically present in V1 units and tend to disappear when all units are considered (Figure 4g). Consistent with this reasoning, shifts in the average centre of pupil also showed correlation with changes in setpoint similarity across many sessions (Figure 4—figure supplement 3,4). This was reminiscent of the dependence of setpoint similarity on pupil width and tended to be higher for sessions with strong correlation between changes in pupil size and position (Figure 4—figure supplement 3,4).

Our results therefore suggest that the changes in setpoint similarity can arise from complex interaction between multiple behavioural parameters and their modulation of neural activity. On a case-by-case basis, it remains to be seen how behavioural parameters and their interactions specifically modulate neural activity on finer scales. On average, however, there was a gradual increase of the average pupil width during the recording session, which was paralleled by a gradual increase of the average activity of units (~50% increase for the average activity of all units across all mice in both datasets; Figure 4—figure supplement 5). The results of our analysis therefore suggest that the behavioural signal can modulate the setpoint of neural activity independent of and in addition to stimulus, and, in doing so, induce a similarity (/dissimilarity) of population responses when behaviour is stable (/changing).

Note that an unintuitive connotation of this finding is that quantifying response similarity from population vectors may reveal representational drift upon behavioural changes, even independent of stimulus-evoked modulation of activity. This is because the constancy of setpoint activity of units would lead to some degree of similarity between population vectors, even if the stimulus-evoked component is different (Figure 2—figure supplement 3d). The behaviourally induced component of similarity changes more slowly, leading to a drop in representational similarity on a longer timescale (e.g., between blocks of stimulus presentation, rather than within them). In line with this

reasoning, we observed a similar drop of representational similarity in CA1 (Figure 2—figure supplement 3e), although individual units in this region had, on average, no reliable visual representations (Figure 2—figure supplement 3a). Modulation of the average setpoint activity – and hence setpoint similarity – by the behaviour can, therefore, contribute to representational similarity, in addition to specific stimulus-induced tuning of responses.

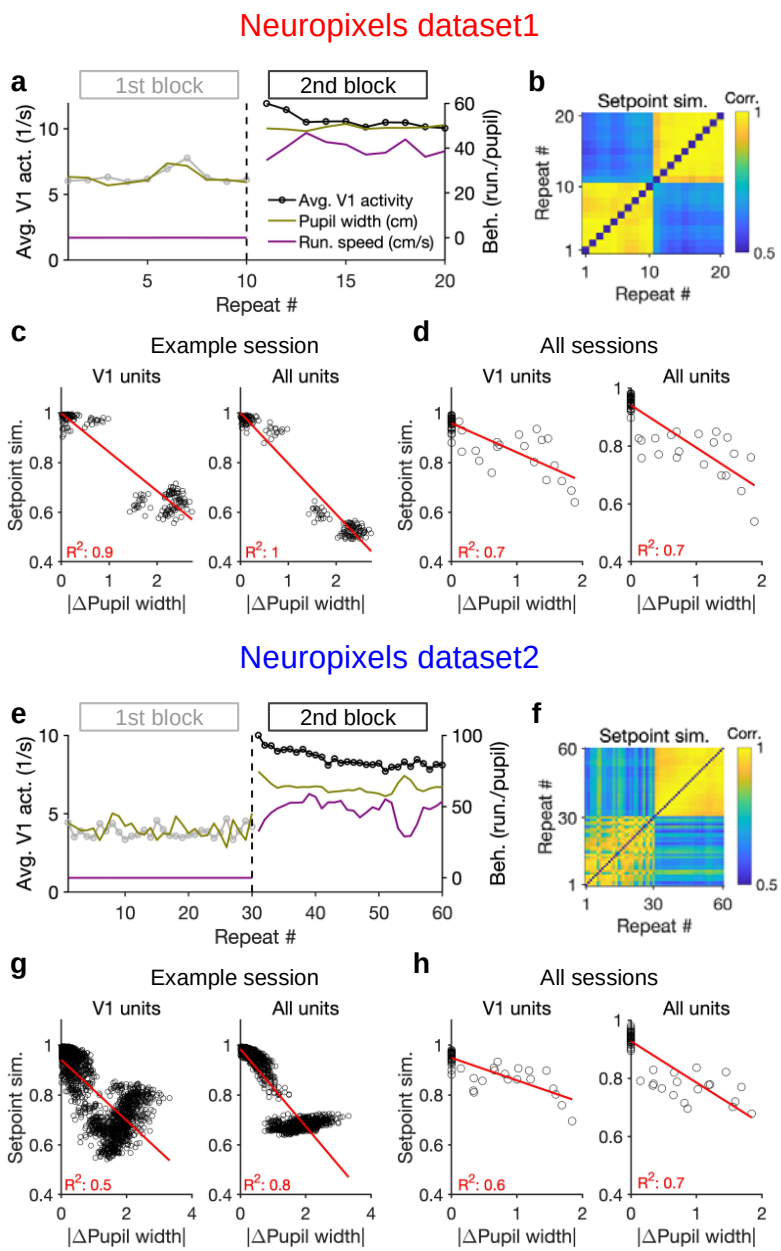


Figure 4: Behaviour modulates the setpoint of responses.

a, Average population activity and behavioural parameters (pupil size and running speed) during the first and second blocks of presentation of natural movie 1 (same examples sessions as Figure 1). Grey, first block; black, second block; each point corresponding to the average in one repeat of the movie. **b**, Setpoint similarity is

calculated as the correlation coefficient of population vectors composed of average activity of units during each repeat of movie presentation. Change in the behavioural state (as quantified by the pupil size) between the two blocks is correlated with a drastic decrease in the average between-block setpoint similarity. Note that transient changes of pupil within each block also modulate the setpoint similarity. **c**, Setpoint similarity (as in **b**) as a function of change in pupil size (z-scored pupil width) between the movie repeats, when the population vectors of setpoints are composed of V1 units (left) or all recorded units (right). **d**, Dependence of setpoint similarity on pupil change for all sessions, calculated from within-block and across-block averages in each session. **e-h**, Same as (**a-d**) for dataset2. Source data (for the change in pupil width and setpoint similarity between pairs of movie repeats) are provided for individual sessions across the two datasets (Figure 4—source data 1).

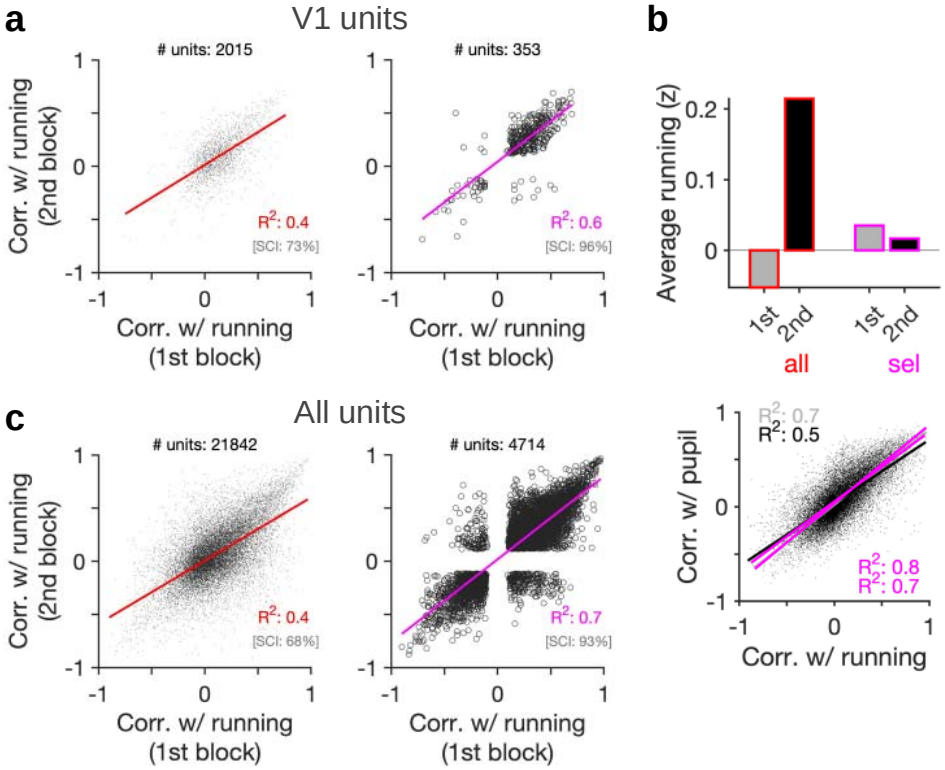
Behaviour reliably modulates responses during active states

What distinguishes an independent behavioural signal from a modulatory component or from noise is that it brings about reliable responses for different states of behaviour. That is, there should exist a reliable and independent tuning of units with behavioural parameters (like pupil size or running speed). We therefore investigated more precisely how the neural activity is modulated by behaviour (Figure 5). We used the correlation of units' activity with running as a metric for behavioural tuning. To obtain a measure of significance of correlations, we calculated bootstrapped correlations (see Methods). More than half of the units showed significant modulation by running, and the fraction and distribution of significant correlations were similar between the two blocks and across the two datasets (Figure 5—figure supplement 1).

Another way to assay the reliability of behavioural tuning is to test whether the correlation of units with behaviour remains stable between the two blocks of presentation (Figure 5a,d). Random correlations with running should be uncorrelated across the two repeats. In contrast, regression analysis revealed a good correlation between the two blocks (Figure 5a,d, left). The distributions of correlations with behaviour were also similar between the two blocks (Figure 5—figure supplement 1). Notably, focusing on sessions with similar levels of running between the two blocks (Figure 5b,e), and on units with significant behavioural modulation, improved the similarity of tuning between the two repeats (Figure 5a,d, right). Specifically, most units which were positively (/negatively) modulated during the first block remained positively (/negatively) modulated in the second block (Figure 5a,d, right). These results therefore suggest that a significant fraction of the population shows reliable modulation by running – similar result is expected for pupil, as we observed a high correlation between modulation of units with running and pupil in both datasets (Figure 5b,e, lower).

Our results held when repeating the analysis for all units instead of V1 units only (Figure 5c,f and Figure 5—figure supplement 1). We also observed similar results when quantifying the reliability of tuning between two blocks of presentation of another stimuli (drifting grating; Figure 5—figure supplement 2). Notably, the tuning of units remained stable from one stimulus type to another: modulation of units during presentation of drifting gratings had a good correlation with their modulation during natural movie presentations for both blocks (Figure 5—figure supplement 2d,h). The tuning with running was even reliable between the first (30-90 mins) and second (90-150 mins) parts of the entire session, with each part containing different stimuli (Figure 5—figure supplement 3). We did a region-specific analysis of this reliability and found that reliable tuning exists in various regions (Figure 5—figure supplement 3). Overall, these analyses suggest that behaviour reliably and independently modulates neuronal responses.

Neuropixels dataset1



Neuropixels dataset2

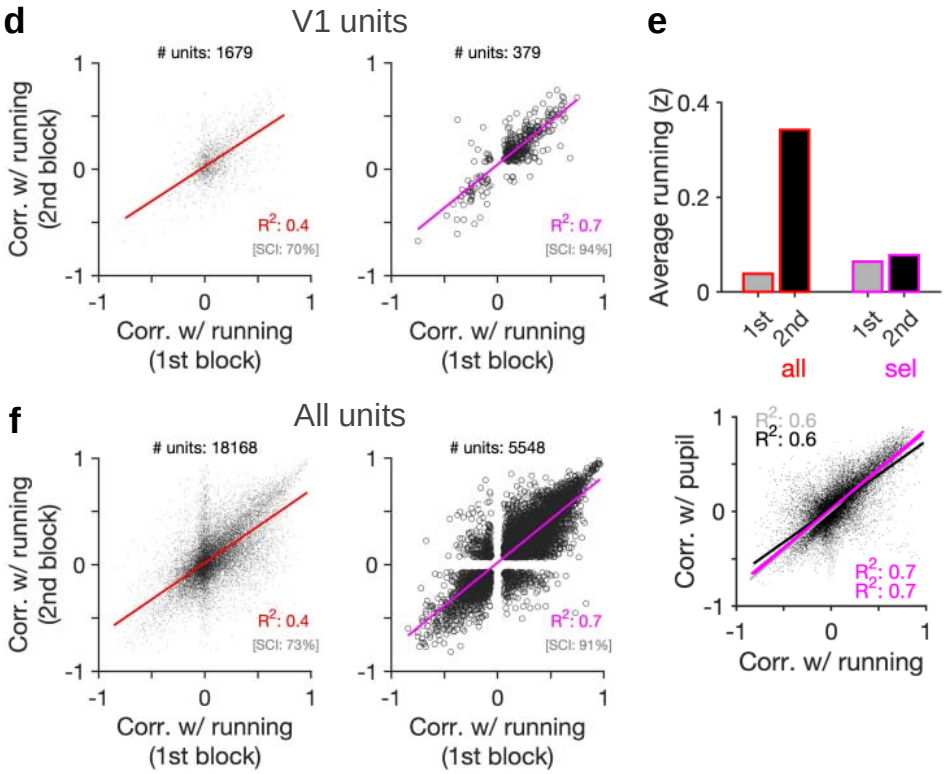


Figure 5: Behaviour reliably modulates responses during active states.

a, Correlation of activity with running during second block against the first block for all V1 units (left), and for selected sessions and units (right). In the latter case, only sessions with similar average running between the two blocks, and units with significant correlation with running, are selected (see Figure 5—figure supplement 1 and Methods for details). In addition to the regression analysis (quantified by the value of R^2), a second metric (sign constancy index, SCI) is also used to quantify the reliability of modulations by behaviour. For each unit, the sign of correlation with running in different blocks are compared and the fraction of units with the same sign is reported as SCI. **b**, Upper: Average z-scored value of running in the first (1st) and the second (2nd) block across all units/sessions (all; red) and for selected ones (sel; magenta). Lower: Correlation of all V1 units with pupil size against their correlation with running in first (grey) and second (black) blocks. Magenta: regression fits for selected units/sessions only. **c**, Same as (**a**) for recorded units from all regions. **d-f**, Same as (**a-c**) for dataset2.

Stimulus-dependence of behavioural variability and setpoint similarity

External stimulus directly modulates the responses by activating selective receptive fields of neurons, which can be measured under anaesthesia (Niell and Stryker, 2008; Yoshida and Ohki, 2020). In awake behaving animals, however, it is possible that different stimulus types indirectly modulate the responses by inducing different patterns of behavioural variability. We indeed found that this was the case when comparing natural movies with an unnatural stimulus (drifting gratings) (Figure 6). Natural movies induced more variability of pupil size and running in the animals across the two blocks of stimulus presentations: both measures significantly increased during the second block for natural movies, whereas changes were not significant for drifting gratings (Figure 6a,d). The result was consistent across the two datasets with different length and sequence of stimulus presentations (cf. Figure 5—figure supplement 3a,b).

To see if and how this difference affects response similarity of units, we calculated average setpoint similarity (cf. Figure 4) between the two blocks of presentations from the shuffled activity of units in response to different stimuli (see Methods). Average setpoint similarity was high for both stimuli, but it was significantly larger for drifting gratings for most sessions (Figure 6b,e). Plotting setpoint similarity as a function of behavioural changes for the entire distribution revealed its composition across the two stimulus types. Responses to drifting gratings showed, on average, a higher setpoint similarity for similar behavioural states (small behavioural changes) (Figure 6c,f), arguing for more stability of average responses even independent of behavioural variability. Larger behavioural changes were more prevalent for the natural movie presentations, and units' responses showed a large drop of setpoint similarity at these deviations (Figure 6c,f), leading to a significant drop of average setpoint similarity compared to drifting gratings. Taken together, these results suggest

that stability of population responses to different stimulus types might be determined by the combined effect of stimulus-evoked reliability of responses and its indirect modulation by behavioural variability.

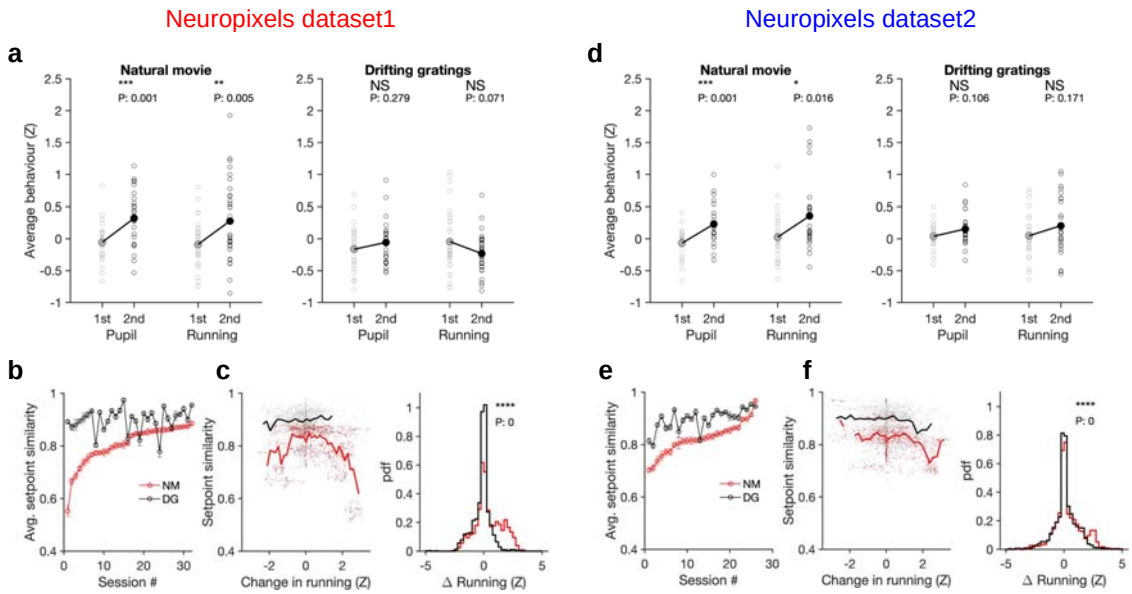


Figure 6: Stimulus-dependence of behavioural variability and setpoint similarity.

a, Average pupil size and running speed during the 1st (grey) and 2nd (black) blocks of presentation of natural movies (left) and drifting gratings (right) for different sessions (empty circles). Filled circles: the mean across sessions. Pupil size and running speed are z-scored across each session, respectively. P-values on top show the result of two-sample t-tests between the two blocks. NS: P > 0.05. *: P ≤ 0.05; **: P ≤ 0.01; ***: P ≤ 0.001; ****: P ≤ 0.0001. **b,** Average setpoint similarity between the two blocks of presentation of natural movie 1 (NM) and drifting gratings (DG) for different sessions. Sessions are sorted according to their average setpoint similarity for NM. Population vectors are built out of the average responses of all units to 30 randomly chosen frames (1 second long). The correlation coefficient between a pair of population vectors from different blocks (within the same stimulus type) is taken as setpoint similarity. The procedure is repeated for 100 pairs in each session and the average value is plotted. Error bars show the std across the repeats. **c,** Left: Setpoint similarity as a function of the difference in average running, $\Delta Z = Z_2 - Z_1$, where Z_1 and Z_2 are the average running during randomly chosen frames in the 1st and 2nd block, respectively. The lines show the average of the points in 40 bins from the minimum ΔZ to the maximum. Right: Distribution of changes in running for different stimuli. The probability density function (pdf) is normalized to the interval chosen (0.25). **d-f,** Same as (**a-c**) for dataset2. Source data (for the average z-scored pupil width and running speed in each block) are provided for individual sessions across the two datasets (Figure 6—source data 1).

Decoding generalizability improves by focusing on reliable units

How does behavioural variability affect the decoding of stimulus-related information, and how can decoding strategies be optimized to circumvent the drift of representations? Our analyses so far suggested that behaviour modulates the responses in addition to and independently of stimulus-evoked modulations (independent model in Figure 2). This independent behavioural modulation would be perceived as noise, if a downstream decoder is decoding stimulus-related signals, and can compromise the generalizability of decoding. For instance, the activity of a subpopulation of units (A) might be assigned to stimulus A by the decoder, in the absence of significant behavioural modulation. If the decoder is now tested in a new condition where behaviour modulates the responses independently of the presented stimulus, the activity of subpopulation A can be interpreted as presence of stimulus A, independent of the presented stimulus. This is in contrast to the gain model (signal-gain model in Figure 2b) in which behavioural state scales the stimulus-evoked signal, and can therefore not compromise the generalizability of decoding (subpopulation A only responds to stimulus A, but with different gains). In the signal-gain model, focusing on units which are strongly modulated by behaviour should in fact enhance the decoding generalizability under behavioural variability, whereas in the independent model the focus should be on units with more stimulus reliability.

We tested these two alternatives directly by training a linear decoder to discriminate between different frames of the natural movie (Figure 7a and Methods). The decoder was trained on the activity of units in the first block to detect a target frame; it was then tested on the second block of presentation to discriminate between the target frame and other frames, in order to evaluate the generalizability of decoding (i.e., out-of-distribution transfer) (Figure 7a). When the decoder was trained on the activity of all units in the first block, discriminability (d') was very low in the second block (Figure 7b,c,e,f). However, focusing on the reliable units (units with high stimulus reliability) shifted the distribution of d' to larger values and increased the average discriminability (Figure 7c,f). Focusing on units with strong behavioural modulation, on the other hand, did not yield higher discriminability in the second block (Figure 7—figure supplement 1). These results suggest that behavioural modulation is detrimental to generalizability of stimulus decoding, and that this problem can be circumvented by focusing on units with more stimulus information.

This effect was consistent across mice in both datasets (Figure 7d,g). In dataset2, we observed higher average d' in the second block, when the decoder was trained and tested on all units. This could be due to more presentations of the natural movie in dataset2 (30 repetitions in each block compared to 10 in dataset1). Larger training samples can help the decoder in learning the signal from the noise, suggesting that the effect of behavioural "noise" on corrupting the stimulus signal is

more significant for small sample sizes. On the other hand, longer presentations can lead to sampling from responses under more behavioural variability, which can in turn inform the decoder on how to ignore the stimulus-irrelevant modulation by behaviour. Altogether, these results corroborate our previous analysis that the contribution of behavioural variability to neural activity is orthogonal to stimulus modulations, and suggest that such behavioural noise limits the decoding capacity especially with limited data.

We also analysed how the behaviour can be decoded from the population activity (Figure 7—figure supplement 2). A decoder was trained on half of movie presentations (randomly chosen from the two blocks of presentation) to predict the pupil width and running speed, and was then tested on the other half of movie presentations (see Methods). For the example sessions (shown in Figure 1), both behavioural parameters were predicted with high accuracy (Figure 7—figure supplement 2a). The accuracy was higher for dataset2 (correlation of more than 90% for both parameters; compared to 60% and 80% for pupil width and running speed predictions, respectively, in dataset1; Figure 7—figure supplement 2a,b). Similar results were obtained when the decoding of behavioural parameters was performed for other mice in both datasets (Figure 7—figure supplement 2b).

Higher prediction accuracy in dataset2 is consistent with our reasoning above that longer episodes of stimulus presentation provide a higher chance for behavioural variability, which can in turn enable a decoder to learn the modulations arising from the behaviour better. An extreme case for this was when the decoder was trained on the first block in the example sessions and tested on the second block. For both datasets, the accuracy was very low (results not shown). This was a consequence of drastic changes of behaviour between the two blocks and specifically very little running in the first block (cf. Figure 1 and Figure 5); in such cases, the decoder cannot learn how behaviour modulates the neural activity from the training data due to paucity of relevant samples.

We also asked whether the decoding of behaviour may create any systematic bias in processing of sensory stimulus. If the behavioural modulation has a systematic relation with the modulation of neuronal activity by stimulus, a decoder which is predicting the animals' behaviour might be biased to infer illusionary stimuli due to the correlations between behaviour and stimulus. To test this, we weighted the responses of neurons to each movie presentation with the weights optimized for the behavioural decoder. We then calculated the correlation of the overall response of the decoder across multiple repeats of the natural movie. Systematic sensory bias in the read-out of the behavioural decoder should lead to significant correlations between the repeats of the same stimulus. Contrary to this, we observed correlations close to zero between different repeats, for the example sessions in both datasets, and across all mice (Figure 7—figure supplement 2c). These results suggest that behavioural decoding can be performed independent of sensory inference.

Taken together, the results of our behavioural decoding reveal two main insights: first, the behavioural state of the animal can reliably be predicted from the neural activity. This is consistent with our results on reliable modulation of neuronal activity by behaviour (Figure 5).

Second, behavioural decoding did not create a significant bias in sensory inference, which supports our previous results on independent modulation of neuronal responses by behaviour and sensory stimulus (Figure 2 and Figure 3).

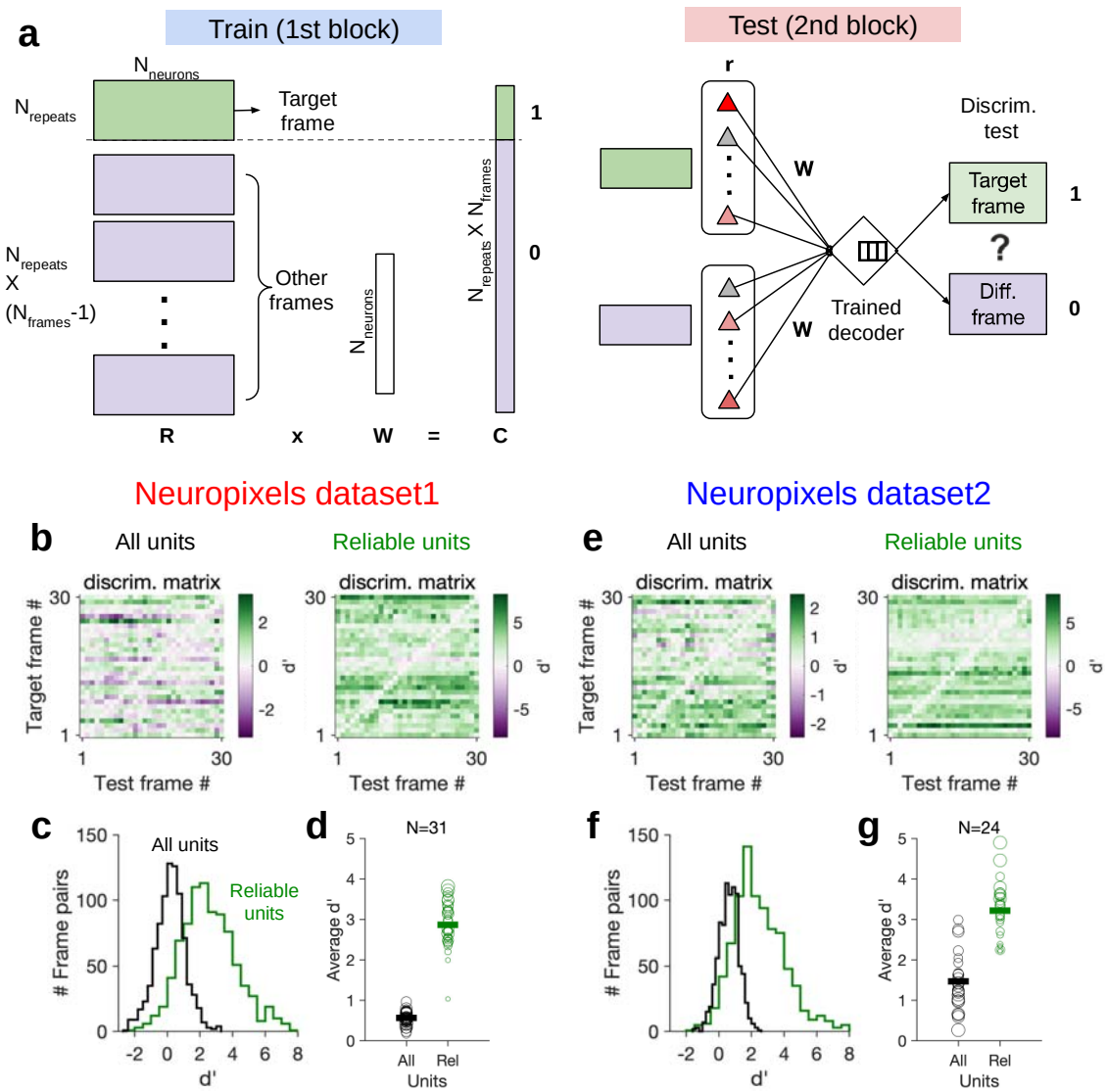


Figure 7: Decoding generalizability of natural images improves by focusing on reliable units.

a, Schematic of a decoder which is trained on the population activity during the first block of presentation of the natural movie (upper) and tested on the second block of presentation to discriminate different frames of natural images from each other (out-of-distribution transfer, see Methods for details). **b**, Matrix of discriminability index (Methods), d', between all combination of movie frames as target and test, when all

units (left) or only units with high stimulus reliability (right) are included in training and testing of the decoder. **c**, Distribution of d' from the example discriminability matrices shown in (**b**) for decoders based on all units (black) and reliable units (green). Reliable units are chosen as units with stimulus reliability (Methods) of more than 0.5. **d**, Average d' for all mice, when all units or only reliable units are included. Size of each circle is proportionate to the number of units available in each session (sessions with >10 reliable units are included). Filled markers: average across mice. **e-g**, Same as (**b-d**) for dataset2. Data in (**b,c**) and (**e,f**) are from the same example sessions shown in Figure 1c and Figure 1f, respectively.

Discussion

The results of our analysis here suggest that variability of the behavioural state of animal can contribute significantly to changes in representational similarity. We found that population responses to different repeats of the same natural movie was the most similar when behavioural parameters like pupil size and running speed changed the least. This was a result of an independent modulation of neural activity by behaviour, which was mixed with stimulus-evoked responses to represent a multidimensional code. Our results are consistent with a view in which behaviour modulates the low-dimensional, slowly changing setpoints of neurons, upon which faster operations like sensory processing are performed.

Small modulation of ongoing neural dynamics by sensory stimuli was reported before in awake, freely viewing animals (Fiser et al., 2004), in line with other reports on the significance of internal signals even in sensory cortices (Arieli et al., 1996; Kenet et al., 2003; Tsodyks et al., 1999). Our results here are consistent with these reports, and our analysis provides a mechanism by which variability of the internal state can contribute to ongoing signal correlations. It suggests that two distinct sources of response similarity exist in neuronal networks, with one set by baseline responses modulated on a slower timescale via internal parameters (setpoint similarity), and the other arising from finer and faster modulations invoked by sensory stimuli. Importantly, changes in representational similarity (which can lead to representational drift) can arise from changes in both sources, and hence attributing it purely to the drift of the sensory component might be inaccurate.

Internal, behavioural states of the animal can contribute independently to neural processing, or can act as a modulator for external stimuli, for instance by increasing the input gain and enhancing the saliency of the sensory signal. Notably, our results could not be explained by a model in which behaviour acted as a gain controller for sensory inputs. Such a model would predict a direct relationship between the stimulus modulation and behavioural modulation of neurons. One would therefore expect that the most reliable neurons in representing sensory information to be modulated the most by arousal or running states. However, we found that the reliability of

stimulus-evoked responses to different repeats of the same natural movie was independent of behavioural modulation, in line with a previous report (Dipoppa et al., 2018).

A gain-model account of behavioural modulation would only change the signal-to-noise ratio of sensory representations by behaviour. Therefore, if the level of arousal or attention of the animal drifts away over time, the signal component of the representations becomes weaker compared to the noise, leading to some drop in representational similarity. In contrast, independent modulation of neuronal responses by behaviour affects representational similarity in more complex ways. First, similarity of population vectors across repeats of the same stimuli can be due, at least partly, to the behavioural signal rather than stimulus-evoked responses. Second, changes in behavioural signal might be perceived as sensory-irrelevant noise, if the parameter space of representations (composed of internal and external parameters) is only analysed over the external stimulus dimension (Montijn et al., 2016; Stringer et al., 2019). Reliable changes in behavioural signals might, therefore, be misinterpreted as the drift of stimulus-induced representations in the latter scenario.

Note that our results here do not rule out the contribution of other parameters, like slow latent learning and plasticity within the circuitry, leading to the drift of sensory representations, especially over longer time scales (days to weeks). Our analysis here revealed changes in neural representations in two blocks of multiple presentations of natural movies within the same day. To determine whether neuronal representations are gradually changing, there must be multiple (more than two) different compared time points (Clopath et al., 2017). Gradual change of neural representations over multiple time points can distinguish between representational drift and random neuronal variability which can arise independently in each repetition. In fact, we observed a gradual *increase* of both pupil width and average activity of units during the entire session (Figure 4—figure supplement 5). Changes in the setpoint similarity arising from such gradual changes can therefore lead to representational drift over multiple time points. It would be interesting to repeat our analysis in future studies on other datasets which contain multiple blocks of stimulus presentations within and across days.

A recent analysis of similar datasets from the Allen Brain Observatory reported similar levels of representational drift within a day and over several days (Deitch et al., 2021). The study showed that tuning curve correlations between different repeats of the natural movies were much lower than population vector and ensemble rate correlations (Deitch et al., 2021); it would be interesting to see if, and to which extent, similarity of population vectors due to behavioural signal that we observed here (cf. Figure 4) may contribute to this difference. In fact, previous studies showed gradual changes in the cells' activity rates during periods of spontaneous activity, suggesting that

these changes can occur independently of the presented stimulus (Deitch et al., 2021; Schoonover et al., 2021).

The fact that there are changes that are not purely related to the tuning of cells is demonstrated in several previous studies. For instance, place cells in the mouse hippocampal CA1 can drop in and out of the ensemble during different visits to the same familiar environment over weeks, leading to a gradual change in the population of place cells that encode the location of the animal during the task (Gonzalez et al., 2019; Ziv et al., 2013). These changes, which reflect changes in activity rates, were independent of the position the neurons encode and were found in place cells and non-place cells alike. A similar turnover of the active cells in the cortex was also shown by (Aschauer et al., 2022; Driscoll et al., 2017). Notably, (Rubin et al., 2015) showed that hippocampal representations of two distinct environments (which had different place-cell representations) co-evolve over timescales of days/weeks, with the shared component of the drift stemming from gradual context-independent changes in activity rates.

The stimulus-independent component of representational drift due to behavioural variability is a global phenomenon that can affect all regions, even independent of their involvement in the processing of natural images. In fact, we found similar representational drift in many areas, including regions like CA1 (see also (Deitch et al., 2021)), although units in this region had no reliable representation of natural images (Figure 2—figure supplement 3a). Global, low-dimensional modulation of activity by behavioural state, like arousal and running, or other aspects of behaviour that we did not have access to their quantification here (e.g., whisking, posture, or body movements other than running), might underlie such changes in representational similarity -- although we cannot rule out the contribution of other factors like contextual modulations (as discussed above) or the passage of time (but see (Nilchian et al., 2022)), which might be more relevant to modulation of activity in regions like CA1. Drawing further conclusions about stimulus-dependences of representational drift in visual cortex – and other sensory cortices – thus needs a critical evaluation by teasing apart the contribution of different components (stimulus-induced and stimulus-independent).

Another recent study reported stimulus-dependent representational drift in the visual cortex, whereby responses to natural images experienced large representational drift over weeks compared to responses to drifting gratings (Marks and Goard, 2021). In line with the finding of this study, we found here that responses to drifting gratings were more robust to behavioural variability in general. However, we also observed that different stimulus types can induce variable patterns of behaviour, thus highlighting the combined contribution of behaviour and stimulus to representational drift. Notably, the mentioned study (Marks and Goard, 2021) found a dependence

of representational drift on the pupil size (see Supplementary Fig. 8c in (Marks and Goard, 2020)), with larger decreases in pupil size over time correlating with more representational drift for both stimulus types (see Supplementary Fig. 11d in (Marks and Goard, 2021)). Such consistent changes of behaviour may contribute to representational drift over longer timescales (days to weeks), by recruiting similar mechanisms as we described here for shorter intervals (e.g., changes in setpoint similarity). Mapping behavioural changes over longer times and per individual animal can shed light on the specific contribution of behaviour to representational drift. It would for instance be interesting to see if the large variability of representational drift across different animals (see Supplementary Fig. 5 in the same study (Marks and Goard, 2021)) might be linked to their behavioural variability.

Behavioural variability might be more pertinent to other modalities for which active sensing is less constrained during experiments. While eye movements are minimized in head-fixed mice, in other modalities (like olfaction) it might be more difficult to control for the behavioural variability arising from active sensing (e.g., sniffing) over time and across animals. A recent study demonstrated significant representational drift over weeks in the primary olfactory cortex of mouse (Schoonover et al., 2021). The surprising finding that sensory representations are not stable in a sensory cortex was hypothesized to be linked to the different structure of piriform cortex compared to other sensory cortices with more structured connectivity. It would be interesting to see if, and to which extent, other factors like changes in the gating of olfactory bulb by variable top-down modulations (Boyd et al., 2012; Markopoulos et al., 2012), or changes in the sniffing patterns of animals, may contribute to this. Similar to the general decline over time of the pupil size reported in the visual cortex (Marks and Goard, 2021), animals may change their sniffing patterns during experiments, which can in turn lead to a general or specific suppression or amplification of odours, depending on the level of interest and engagement of individual animals in different sessions.

Behavioural modulation might be more systematically present depending on the design of specific tasks, for instance if a sensory task is novel or otherwise engages behavioural states. Interpretation of the results of neural response to novel or surprising stimuli might, therefore, be compromised, if one ascribes the changes in neural activity to local computations only, without the analysis of behaviour and without controlling for more global signals (e.g., arising from arousal, running, whisking, or licking). Low-dimensional signals associated with behavioural and internal state of the animal have in fact been suggested to create a potential confound for multi-sensory processing, with sound eliciting eye and body movements that modulate visual cortex, independent of projections from auditory cortex (Bimbard et al., 2021).

How can the animals perform reliable sensory processing in the face of extensive modulation of neural activity by behavioural changes? Our results suggest that the reliable units, which are stably encoding sensory information across behavioural states, could serve as a stable core for the brain to rely on for coping with changing neuronal responses (Figure 7). However, the distribution of stimulus reliability was surprisingly skewed towards lower reliability values (Figure 2H). In fact, at the population level, units were more reliably modulated by behaviour (Figure 2H). It is therefore likely that the brain would rely on both stimulus- and behaviour-related information, rather than focusing on a small subset of reliable units, to cope with changing representations.

Beyond sensory processing, variability of internal state can also contribute to other cognitive processes in various cortices (Joshi and Gold, 2020). A recent study in monkey found that changes in the perceptual behaviour was modulated by a slow drift in its internal state, as measured by pupil size (Cowley et al., 2020). This was correlated with a slow drift of activity in V4 and PFC, along with changes in the impulsivity of the animal (as reflected in the hit rates), which overrode the sensory evidence. These results, in another species, are in agreement with our findings here on the contribution of behavioural drift to changes in neural representations. Interestingly, the sensory bias model in the study could not capture the effect of the slow drift on decoding accuracy; instead, an alternative impulsivity model, which introduced the effect of slow drift as an independent behavioural parameter, matched with the data (Fig. 6 in (Cowley et al., 2020)).

Another study in monkey M1 found that learning a new BCI task was modulated along the dimension of neural engagement of the population activity, which in turn was correlated with pupil size (Hennig et al., 2021). Neural engagement increased abruptly at the beginning, and decreased gradually over the course of learning, where output-null and output-potent components of neural engagement differentially attuned for different targets. Notably, exploiting behavioural perturbations in this study enabled an interactive interrogation of the neural code during learning. Behavioural perturbations, combined with large-scale recording and perturbation of neural activity (Emiliani et al., 2015; Yizhar et al., 2011; Zhang et al., 2018), which are more feasible in mice, can pave the way for a more precise (and potentially causal) interrogation of the neural mechanisms underlying representational drift. It would specifically be interesting to see how the bidirectional modulation of activity by behaviour we observed here emerges and which circuit mechanisms (Cohen-Kashi Malina et al., 2021; Ferguson and Cardin, 2020; Fu et al., 2014; Pakan et al., 2016) contribute to it.

In summary, our analysis reveals new insights on representational drift from the viewpoint of behaviour. Conceptually, it argues for the primacy of internal parameters (Buzsáki, 2019), and suggests that representational similarity could be better understood and characterized in a

multidimensional parameter space where the contributions of both external and internal parameters are considered. Computationally, it argues for an independent mixing of stimulus-evoked and behavioural signals, rather than a simple gain modulation of sensory inputs by behaviour. Technically, it asks for further controls and analysis of behavioural variability in the characterisation of representational drift. Future studies will hopefully probe the multidimensional code underlying representations in the brain by combining large-scale recordings of neural activity with simultaneous measurement and quantification of behaviour.

Methods

Curation and preprocessing of the data

Data curation. Publicly available data provided by the Allen Brain Observatory (de Vries et al., 2020; Siegle et al., 2021) was accessed via AllenSDK (https://allensdk.readthedocs.io). We analysed recording sessions in which neuronal responses to visual stimuli were measured via electrophysiology techniques by Neuropixels probes (https://portal.brain-map.org/explore/circuits/visual-coding-neuropixels). The data composed of 58 sessions/mice in two separate datasets: brain observatory dataset (Dataset1; n=32) and functional connectivity (Dataset2; n=26) (Supplementary Table 1). Similar stimuli (including natural moves and drifting gratings) were shown to the animals, with different length and sequence of presentations in each dataset (https://allensdk.readthedocs.io/en/latest/_static/neuropixels_stimulus_sets.png; see Figure 5—figure supplement 3a,b for illustration of different stimulus sets). We used the spiking activity of units which was already extracted by Kilosort2 (Stringer et al., 2019), and we included units in our analysis which passed the default quality criteria. Invalid intervals were treated as Not a Number (NaN) values. For further details on the preparation of animals, visual stimulation, data acquisition and default pre-processing of data, see the Technical White Paper from the Allen Brain Observatory on Neuropixels Visual Coding.

Pre-processing of data. For our analysis here, we rendered the spiking activity of units in bins of 1 second. When analysis was focused on specific stimulus types (e.g. presentation of natural movie 1 as in Figure 1d,g), the activity was rendered from the onset of presentation of each block of the stimulus. When the analysis was across all stimuli and involved the activity during the whole session (e.g. data shown in Figure 1c,f), the activity was rendered from the beginning of the session or an arbitrary time (e.g. time frames specified in Figure 5—figure supplement 3). Behavioural information was obtained in similar time frames. Locomotion was quantified for all animals as the average running speed. Level of arousal was quantified by pupil size, as measured by pupil width (whenever pupillometry was available; Supplementary Table 1).

To normalize the parameters (e.g., to normalize for different size of pupil across animals), we calculated their z-score values. For parameter x (units' activity, pupil size or running speed), it was obtained as $z = (x - \mu_x)/\sigma_x$, where μ_x and σ_x are the mean and standard deviation of x during the entire session or a specified time window.

Data analysis

Representational similarity. Representational similarity of population activity was quantified by calculating the correlation of responses to different repeats of the same stimulus (Figure 1--figure

supplement 1c). Let v be a vector of responses of N recorded units to M 1-second long chunks of a natural movie (the natural movie is broken down to M chunks, or frames, each lasting for 1 second, corresponding to the bin width the neural activity is rendered in). v is a $1 \times \mathrm{N}M$ population vector composed of the concatenated activity of units (either the actual activity, i.e. average spiking activity, or the z-scored activity of each unit). Denote v_i and v_j as vectors of responses to two repeats of the same natural movie. Representational similarity is quantified as the Pearson correlation coefficient of these two population vectors:

$$\rho_{ij} = \frac{\mathrm{cov}(v_i, v_j)}{\sigma_{v_i}\sigma_{v_j}}$$

Representational drift index (RDI) is calculated from representational similarity. Similar to previous metrics(Marks and Goard, 2021), we defined RDI between two blocks of stimulus presentation as:

$$RDI = (CC_{ws} - CC_{bs})/(CC_{ws} + CC_{bs})$$

where CC_{ws} and CC_{bs} are, respectively, the average correlation coefficient of population vectors within and between sessions of presentation (see Figure 1—figure supplement 5). CC_{bs} was obtained from the average of ρ_{ij}, for all pairs of repeats with i-th repeat in the first block and the j-th repeat in the second block of presentation. CC_{ws} is obtained from the average of ρ_{ij}, for all non-identical pairs of repeats where the i-th and j-th repeats are both within the same block, respectively.

Note that RDI can also be defined, in principle, for a single pair of movie presentations (corresponding to blocks of presentation with 1 movie in each block). In this case, CC_{ws} is 1, by definition, and we therefore have:

$$rdi = \frac{1 - \rho_{ij}}{1 + \rho_{ij}}$$

showing the relation between representational drift and representational similarity in its most simplified case.

Stimulus reliability. We also quantified the reliability of how single units respond individually to repetitions of the stimuli (Figure 1--figure supplement 1d). To quantify that, we calculated a stimulus reliability metric, which is obtained as the average correlation coefficient of each unit's activity vector (r) across repetitions of the stimulus (e.g., the natural movie). Let ${r^k}_i$ and ${r^k}_j$ be the vectors of response of the k-th unit to the i-th and j-th repetitions of the natural movie. Similarity of the unit's response between these two repeats can be quantified by the Pearson correlation coefficient of the responses as before:

$$\rho_{ij}^k = \frac{\mathrm{cov}\left(r_i^k, r_j^k\right)}{\sigma_{r_i^k}\sigma_{r_j^k}}$$

Stimulus reliability of the unit k is obtained as the average correlation across all pairs of (non-identical) repetitions of the stimulus:

$$\rho^k = \frac{1}{N_r(N_r-1)}\sum_{i=1}^{N_r}\sum_{j\neq i}\frac{\mathrm{cov}\left(r_i^k, r_i^k\right)}{\sigma_{r_i^k}\sigma_{r_i^k}}$$

where N_r is the number of repetitions of the stimulus. Note that to each single unit we can ascribe a stimulus reliability index, since this is calculated from the individual vectors of single units' responses (r^{k}); on the other hand, representational similarity is calculated from the population vector of responses (v) and indicates a single population metric ascribed to the activity of a group of neurons (e.g., V1 units or all recorded units).

Behavioural tuning. To obtain a measure of how single units are modulated by behaviour, we calculated the correlation of units' responses with behavioural parameter, β:

$$\rho_i(\beta) = \frac{\mathrm{cov}(r_i, \beta)}{\sigma_{r_i}\sigma_\beta}$$

Here, r_i is the vector of response of the i-th unit, and β is the vector of respective behavioural parameter (either pupil size or running speed) rendered during the same time window and with the same bin width as unit's activity.

To obtain a measure of reliability of this modulation by behaviour, we calculated bootstrap correlations. The activity of each unit was shuffled for 100 times and the correlation with behaviour was calculated. The mean (μ_{sh}) and std (σ_{sh}) of the distribution of shuffled correlations were then used to obtain the z-scored, bootstrapped correlation:

$$\mathrm{Z} = \frac{\rho(\beta) - \mu_{sh}}{\sigma_{sh}}$$

where $\rho(\beta)$ is the unshuffled correlation of the unit with behaviour.

Principal component analysis. To analyse the low-dimensional patterns of activity, we performed principal component (PC) analysis on the vectors of population activity. The vectors of neuronal activity from which response similarity was calculated were of size $N \times N_r \times M$, where N is the number of recorded units, N_r is the number of repeats of the stimulus, and M is the length of rendered activity in chunks of 1-second (corresponding to image frames of the natural movie). We concatenated these vectors to obtain a $N \times N_r\ M$, with each row denoting the total activity of each unit to all the repeats. PCs resulting from the PC analysis, therefore, represented a vector of length $N_r\ M$, denoting the low-dimensional activity of the population in the same time frame (each data point in Figure 3b,h corresponding to 1 s of activity). PC loadings (of size N for each PC) are used to represent the individual units in the space of PCs (Figure 3c,f,i,l).

Modelling

Gain models. To gain mechanistic insight on the contribution of behavioural changes to modulation of representational similarity, we explored two different models. First, we developed a gain model, in which the integration of the signal and the noise by neurons was differently modulated by behaviour (Figure 2—figure supplement 1a). For a population of N_p neurons, let u be the $1 \times N_p$ vector of responses of neurons upon presentation of a stimulus. This is assumed to be composed of signal (S) and noise (N) components. Change in the behavioural parameters (for instance, pupil size) is supposed to change a gain parameter, g, which in turn differently modulate the signal (S) and noise (N). The vector of population activity, u, is obtained, as a linear combination of weighted components by the behavioural/gain parameter. If the signal and the noise are both scaled by the behavioural parameter, it is given as $u = \mathrm{gS} + \mathrm{gN}$. If either the noise or the signal is scaled, it is given as $u = \mathrm{S} + \mathrm{gN}$ or $u = \mathrm{gS} + \mathrm{N}$, respectively (Figure 2—figure supplement 1b).

S and N are both vectors of size $1 \times N_p$, where each element is drawn from a random uniform distribution between $[0,1]$. The population activity is simulated for N_r repeats of the stimulus. The stimulus signal, S, remains the same for all the repeats (frozen noise drawn from the same range as before, $[0,1]$), while the noise component, N, is instantiated randomly on each repeat (from the same range, $[0,1]$, as the signal). The behavioural parameter (e.g., pupil size) is assumed to change on every repeat too, which changes the gain parameter, g, as a result. g was therefore assumed to be a random number uniformly drawn from the range $[0.5, 2]$ for each repeat. We chose $N_p = 1000$ and $N_r = 100$.

Representational similarity for different models was calculated, similar to the procedure in analysing the experimental data, as:

$$\rho_{ij} = \frac{\mathrm{cov}(u_i, u_j)}{\sigma_{u_i}\sigma_{u_j}}$$

where u_i and u_j are population responses to the i-th and j-th repeat of the stimulus, obtained from different gain models. This value is plotted against the relative gain (obtained as the ratio of the gains in the two repeats, g_i/g_j or g_j/g_i) in Figure 2—figure supplement 1b.

Extended gain model. To match better with the experimental data on a single unit level, we extended the previous signal gain model to have stimulus tuning for individual units (Figure 2a). Whereas before the stimulus was assumed to be a single, fixed value between $[0,1]$ for each neuron, now the stimulus itself is extended (corresponding to different frames of the natural movie or different orientations of drifting gratings). The stimulus, s, is assumed to be a vector of fixed random values between [0,1] with size $1 \times N_s$. Each neuron, k, has a different stimulus drive/tuning, T_k, with which the stimulus vector is multiplied. T is a vector of size $1 \times N_p$ (number of neurons in

the population), randomly drawn from $[0,1]$. Response of the k-th neuron to each repeat of the stimulus is composed of its stimulus signal ($S = T_k s$), which is multiplied by the behavioural gain (g), and an added noise term (N), which is independently drawn for each stimulus and repeat from the range $[0,1]$. $N_s = 10$, $N_p = 1000$, $N_r = 200$.

Independent model. We also developed an alternative model, whereby the effect of behaviour on population responses was modelled as an independent signal (Figure 2b). Here, instead of scaling signal or noise components of the input, behaviour enters as an independent parameter:

$$u = S_S + N + S_B$$

where S_S and S_B are stimulus-evoked and behavioural signals and N is the noise. S_S and N were instantiated as before, while S_B was determined based on two factors. First, the behavioural parameter, β, which was changing on every repeat, and was simulated, similarly as the behavioural gains before, by a random number between $[0.5, 2]$ for each repeat. Second, the vector of tuning (T_B) of different neurons in the population with the behavioural parameter, which was modelled as a random number between $[0,1]$ for each neuron. The behavioural signal was obtained as: $S_B = \beta T_B$. Representational similarity was computed as before for the population vectors and plotted against the relative behavioural parameters.

Decoding model. To directly compare the stimulus-induced information available in different blocks of stimulus presentation, we developed a decoding model (Figure 7a). A linear decoder is trained on the neural activity (composed of the average activity of units in response to different repeats of the natural movie) during the first block of presentation to discriminate different frames (1 second long) of the natural movie (Figure 7a, upper). The weights of the readout (W) for each target frame were optimized to maximize its classification (C=1) against other, non-target frames (C=0). The decoder is then tested on the data in the second block (Figure 7a, lower). The population activity in response to each frame (the vector of average responses of neurons to a single frame across different repeats) is passed through the decoder to determine whether it is the target (D=1) or not (D=0). Performance of the decoder is quantified by calculating the discriminability (d′) as

$$d' = \frac{\mu_s - \mu_n}{\sqrt{\sigma_s^2 + \sigma_n^2}}$$

where μ_s and σ_s are the average and std of D for target frame across repetitions (within the second block), and μ_n and σ_n are similar values for non-target frames. The discrimination matrix (Figure 7b,e) then shows the discriminability (d′) of each movie frame as a target when presented against all other frames.

Theoretical analysis

Gain models. Representational similarity for the responses in the gain models can be calculated as follows.

In the absence of any scaling of the signal or the noise, $u = S + N$, the representational similarity is obtained as the correlation coefficient of responses to a pair of stimulus repeats:

$$\rho_{ij} = \frac{\mathrm{cov}(u_i, u_j)}{\sigma_{u_i}\sigma_{u_j}}$$

where $u_i = \mathrm{S} + N_i$ and $u_j = \mathrm{S} + N_j$. Assuming that S and N have zero means, we can write:

$$\rho_{ij} = \frac{\sigma_{\mathrm{S}}^2}{\sigma_S^2 + \sigma_N^2}$$

where σ_S and σ_N are the std of S and N, respectively. This indicates that representational similarity can be expressed as a function of the relative variability of the signal and the noise. If modulation of the responses due to signal is dominant over the noise, $\sigma_S \gg \sigma_N$, $\rho_{ij} \to 1$.

If both the signal and the noise are scaled by the behavioural parameter, by the gain factor g, as $\mathrm{u} = \mathrm{gS} + \mathrm{gN}$, we obtain:

$$\rho_{ij} = \frac{g_i g_j \sigma_S^2}{g_i g_j (\sigma_S^2 + \sigma_N^2)} = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_N^2}$$

where g_i and g_j are the gains in the i-th and j-th repeat of the stimulus, respectively. Representational similarity, therefore, remains the same under similar scaling of S and N.

If only the noise is scaled by behaviour, we obtain:

$$\rho_{ij} = \frac{\sigma_S^2}{\sqrt{(\sigma_S^2 + g_i^2\sigma_N^2)(\sigma_S^2 + g_j^2\sigma_N^2)}}$$

showing that the larger the gain, the smaller the representational similarity.

Similarly, if only the signal is scaled, representational similarity can be obtained as follows:

$$\rho_{ij} = \frac{g_i g_j \sigma_S^2}{\sqrt{(g_i^2\sigma_S^2 + \sigma_N^2)(g_j^2\sigma_S^2 + \sigma_N^2)}}$$

which, if rewritten as:

$$\rho_{ij} = \frac{\sigma_S^2}{\sqrt{(\sigma_S^2 + \sigma_N^2/g_i^2)(\sigma_S^2 + \sigma_N^2/g_j^2)}}$$

shows that larger gains effectively decrease the significance of noise, and hence enhance representational similarity. Specifically, in the limit of very large gains for both repetitions ($g_i \gg 1,\ g_j \gg 1$), we have: $\rho_{ij} \to 1$.

For the specific case where gains are the same between the two repeats ($g_i = g_j = g$), the equation simplifies to:

$$\rho_{ij} = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_N^2/g^2}$$

Thus, for similar behavioural states (and hence gains) between the two repeats of the stimulus, representational similarity increases if $g > 1$ and decreases if $g < 1$.

Independent model. For the model in which the stimulus and the behaviour contributes independently to neural responses, representational similarity in response to the same stimulus can be expressed as:

$$\rho_{ij} = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_B^2 + \sigma_N^2}$$

where σ_S, σ_B, and σ_N denote the variability of the population response induced by stimulus, behaviour and noise components, respectively. In deriving the above equation, we have assumed that the stimulus and behavioural components of the signal are independent, i.e. $< S_S . S_B >= 0$ (in addition to the noise term being independent of S_S and S_B, respectively). We also assumed that the behavioural signal, $S_B = \beta T$, remained the same between the two repeats (that is, the behavioural parameter was the same: $\beta_i = \beta_j = \beta$). If the behavioural parameter changes between the repeats, the equation can, in turn, be written as:

$$\rho_{ij} = \frac{\sigma_S^2}{\sqrt{\left(\sigma_S^2 + \beta_i^2\sigma_T^2 + \sigma_N^2\right)\left(\sigma_S^2 + \beta_j^2\sigma_T^2 + \sigma_N^2\right)}}$$

Note that, when representational similarity is only characterized in terms of the stimulus part of the signal, the contribution of behavioural variability is similar to a noise term – decreasing ρ_{ij} for larger values of β. Changes in the behavioural state can, thus, not be distinguished from random variability of the "signal".

Relation between representational similarity and stimulus reliability. As explained above, representational similarity and stimulus reliability are calculated to quantify the similarity of population and single units' responses, respectively, to the repeats of the same stimulus. In fact, representational similarity of a population vector composed of one single unit is the same as the stimulus reliability of that unit. Similarly, if all the units in a population of neurons had the same response profile in response to the stimulus, the stimulus reliability of units would be the same as the representational similarity of the population responses. Although these two measures are related (similar to lifetime sparseness and population sparseness (Froudarakis et al., 2014)), they are, however, not always directly equivalent to each other.

Consider a single unit, k, which has a constant baseline firing rate of r_b and a component which is modulated by the stimulus, r_m: $r = r_b + r_m$. If the stimulus-modulated component of the response is randomly changing between different repeats of the stimulus, the neuron would have a stimulus reliability of zero: $\rho^k = 0$. A population of units with this behaviour would have an average stimulus reliability of zero. However, the representational similarity of the responses of this population is not necessarily zero. In fact, we may obtain high values of population-level representational similarity, if the baseline component of the responses is significantly larger than their modulation ($r_b \gg r_m$). Under this scenario, representational similarity is calculated from the baseline component of the population responses (v_b), which indeed remains constant across repeats, hence $\rho_{ij} \to 1$.

Supplementary Figure Legends

Figure 1—figure supplement 1: Characterization and quantification of representational similarity and representational drift.

a, Left: Illustration of response of a population of neurons (#1 to #n; upper) to a stimulus (lower), composed of binary values (ON: red; OFF: white). **b**, The population response to two other repetitions of the same stimulus can remain the same (upper), demonstrating a stable and reliable code, or it can change from the original pattern (lower), leading to a drift of representations. Representational drift can be quantified by a representational drift index (RDI), which compares the correlation of population responses within the same session/block of presentation (CC_{ws}) with the correlation of population responses across sessions/blocks (CC_{bs}; see Methods). **c**, The degree of change or constancy of representations can be assayed by comparing the population responses to repeats of the same stimulus. The degree of similarity is quantified by representational similarity (RS), which is quantified by the corelation coefficient (CC) of the concatenated (across neurons) vector of population responses to two repeats (PV(i) and PV(j)). **d**, Stimulus reliability (SR) is calculated for each unit individually, from the CC of the vector of responses of that unit to two stimulus presentations.

Figure 1—figure supplement 2: Relation between behavioural changes and representational similarity in other sessions and for all units.

a, Same as Figure 1d,g for four other example recording sessions. Left: Examples from Neuropixels dataset1; Session numbers and the number of V1 units (#), respectively: 762602078 (#75), 750332458 (#63), 760345702 (#72), 751348571 (#49). Left: Examples from Neuropixels dataset2; Session numbers and the number of V1 units (#), respectively: 766640955(#52), 787025148(#68), 771990200(#54), 829720705(#52). Only sessions with #>40 units are included in the analysis. **b**, Same as Figure 1e,h when all recorded units are included (instead of only V1 units).

Figure 1—figure supplement 3: Dependence of representational similarity on behavioural change when calculated from z-scored activity and from activity rendered in longer time bins.

a, Same as Figure 1e,h when population vectors are composed of z-scored activity of units in V1 (upper) or all regions (lower). Z-scored activity of unit i is calculated as $z_i = (r_i - \mu_i)/\sigma_i$, where μ_i and σ_i are the average and std of the activity of unit (r_i) during the two blocks of presentation of natural movie 1. Left: Neuropixels dataset1; Right: Neuropixels dataset2. **b**, Same as Figure 1e,h

when representational similarity is calculated from the activity of units rendered in time bins of 2 seconds (instead of original 1 second; see Methods).

Figure 1—figure supplement 4: Dependence of representational similarity on behavioural change in wild type mice and for male/female animals separately.

Same as Figure 1—figure supplement 2 when (**a**) only wild typed (WT) mice are included in the analysis, or (**b**) when the analysis is performed for V1 units in female and male mice separately (see Supplementary Table 1 for details). Left: Neuropixels dataset1; Right: Neuropixels dataset2.

Figure 1—figure supplement 5: Representational drift between the two blocks of stimulus presentation in animals with different levels of behavioural variability.

a, Matrices of representational similarity (calculated from population vectors of V1 units) between repeats of Natural Movie 1 (similar to Figure 1d,g, left) for sessions across the two datasets (Upper: Neuropixels dataset1; lower: Neuropixels dataset2; The colour code is the same as the example shown in (c).). The numbers on the bottom denote the average change in the pupil width between the two blocks (average of the absolute normalized change between all pairs of repeats in the first and the second block), and the sessions are sorted according to that (from the least change to the highest change). The numbers on the top denote the representational drift index (RDI; see Methods) between the two blocks, which is calculated as: $RDI = (CC_{ws} - CC_{bs}) / (CC_{ws} + CC_{bs})$, where CC_{ws} and CC_{bs} represent the average representational similarity within sessions and between sessions of the two blocks, respectively. CC_{ws} is obtained as the average of average CC within the first block and the second block. **b**, Relationship between the average change in pupil size (numbers denoted on the bottom of plots in (a)) and representational drift for different sessions in the two datasets. Red lines show the best fitted regression lines, with values of R^2 denoted in red. **c**, The example session, highlighted with the red box in (a), is singled out (top), as being an outlier in terms of the relation between pupil change and RD between the two blocks (red circle in (b)). The average pupil width and running speed for each movie repeat is shown for this session on the bottom. **d,e**, Same as (b,a), respectively, for all units (instead of V1 units).

Figure 2—figure supplement 1: Dependence of representational similarity on behaviour in different gain models and in different experimental datasets.

a, A model neuron which integrates the signal and the noise components in its inputs. The signal has the same pattern over multiple repetition (rep#) of the stimulus, while the noise changes in each repeat. **b**, Representational similarity as a function of relative gain of the repeats (g_i/g_j,

where g_i and g_j are the gains in the i-th and j-th repeats) for three models, where both signal and noise (left), only noise (middle), or only signal (right) components of the input are scaled by behaviour (see Methods for details). Change in the gain did not change representational similarity when both signal and noise were scaled (left), consistent with our theoretical analysis (see Methods). When arousal scaled noise only, there was a small decrease in the average representational similarity (middle). The most prominent effect was observed when arousal scaled signal only. For this scenario, a general increase in the average representational similarity was obtained, with the maximum increase happening at equal gains ($g_i = g_j$) (right). **c**, Representational similarity as a function of relative pupil size (obtained by the division of the average pupil sizes in a pair of movie repeats) for all recorded units. **d**, The average representational similarity of all mice shown in (**c**) for datasets 1 (red) and 2 (blue) separately. Bottom: Same when V1 units are only included, from the sessions with more than 40 units (the inclusion criterion). **e**, Same as (c) when V1 units are only included in the analysis. There are fewer individual sessions here because not all sessions contained more than 40 V1 units. **f**, Same as (d) for V1 units. **g**, Same as Figure 1g, for a stronger value of noise (x2 N; left) or a weaker value of behavioural signal (x0.5 B; right).

Figure 2—figure supplement 2: Wide and mixed distribution of stimulus and behavioural modulations.

a, Distribution of stimulus reliability for all V1 units from all sessions in Neuropixels dataset1 (red) and dataset2 (blue). The average stimulus selectivity for the two datasets are (0.32, 0.36), respectively. Gray lines show the distribution of stimulus reliability when it is calculated for each block separately and then average across the two blocks. The average values obtained in this manner are (0.36, 0.40) for the two datasets, respectively. **b**, Sample activity of V1 units with high stimulus reliability (indicated by the numbers on the top) from each dataset. Top: The activity in response to each movie repeat; bottom: average activity in each block of presentation. **c**, Distribution of behavioural modulation of all V1 units for the two datasets. Behavioural modulation is obtained as the correlation coefficient (CC) of each unit's activity with pupil size. **d**, Sample activity of V1 units with strong modulation by pupil size (numbers indicated on the top). Top: tuning of unit's activity with pupil size. Bottom: the activity of units in response to repeats of the natural movie, showing different levels of modulation by stimulus within and across blocks of presentation (denoted by the value of stimulus reliability on top). **e,f**, Activity of units can be weakly or strongly modulated by stimulus or behaviour, giving rise to four possible quadrants. Sample V1 units from each quadrant are shown for Neuropixels dataset1 (e) and dataset2 (f). For each sample, z-score

activity of the unit across different repetitions of the movie is plotted (left), with the number on top denoting stimulus reliability of the unit. The average activity of each unit as a function of average pupil size (during each movie repeat) is plotted on the right, with the number on the top denoting behavioural modulation of the unit (CC with pupil size).

Figure 2—figure supplement 3: Stimulus-independent behavioural modulation of CA1.

a, Top: Average stimulus reliability across units in V1 and CA1 for different mice in each dataset. Bottom: Same for the average (across units) of the absolute value of behavioural modulation. Filled circles: average across mice. Red: Neuropixels dataset1; Blue: Neuropixels dataset2. **b**, Top: Sample activity of CA1 units (from Neuroixels dataset1) with considerable modulation by pupil size (numbers indicated on the top). Bottom: The activity of units in response to repeats of the natural movie. **c**, Same as (**b**) for Neuroixels dataset2. **d**, Schematic representation of population responses with stimulus-evoked (red) and behaviourally induced (green) components to the repeats of the same stimulus. Even if the stimulus-evoked component is different between repeats (red), the population vector of responses (see Figure 1--figure supplement 1c) can have some similarity due to the constancy of the component set by the behaviour (green). **e**, Average representational similarity as a function of change in pupil width (similar to Figure 1e,h, right) for V1 (left) and CA1 units (right). Red: Neuropixels dataset1; Blue: Neuropixels dataset2.

Figure 3—figure supplement 1: Relation of the principal components of neural activity to the average activity of units and their stimulus reliability.

a, Strong correlation between PC1 and the behavioural state of animal, as assayed by either pupil size (left) or running speed (right). Both pupil size and running speed are normalized by z-scoring across the entire session. **b**, Projections of the activity of units in example sessions over the first three PCs (cf. Figure 3a,b,g,h), with the average activity of each unit indicated by the pseudo colour code. **c**, Projection of units' activity over PC1/PC2 versus the average activity of the unit. The best fitted regression lines and R^2 values in each case are shown. **d**, Similar to Figure 3c,i for individual sessions. R^2 values of regression lines fitted to the projection of units' activity over PC1/PC2 versus stimulus reliability of the respective units is plotted for individual sessions. Upper: Neuropixels dataset1; Lower: Neuropixels dataset2.

Figure 3—figure supplement 2: PC1 and its relation to difference in pupil size and time.

a, Relationship between distance on PC1 (cf. Figure 3e,k), change in pupil width and passage of time. Datapoints highlighted in blue show the cluster with smaller passage of time (corresponding to movie repeats within the same blocks of presentation). Note that the two clusters arise from two

blocks of stimulus presentation, which are separated from each other for >80 minutes in both datasets. **b**, Distance on PC1 versus change in pupil width for data points with small passage of time (blue data points in (a)). The best fitted regression line and its R^2 value denoted in blue. **c**, Distance on PC1 versus passage of time. The best fitted regression line and its R^2 value denoted in red.

Figure 4—figure supplement 1: Average activity of units is modulated by behavioural state.

a, Average population activity (left y-axis) and running speed (right y-axis) as a function of pupil size, for the example shown in Figure 4a from Neuropixels dataset1. **b**, Average population activity of V1 units during each movie presentation as a function of pupil size, from all recorded sessions. Pupil size for each repeat is normalized (within each session) by subtracting the mean value (across repeats) and dividing by it. **c**, For the example session in Figure 4, the average (across movie frames) activity of V1 units is calculated and their mean and std across movie repetitions in each block is shown. Units are sorted in both blocks according to the mean in the 1st block. **d**, Average activity (across movie frames and repeats) of units during the 2nd block versus the 1st. Note the logarithmic scales. **e**, R^2 values of the regression fits to the data like Figure 4c, when the population vectors are composed of the average activity of units during presentation of each individual frame (1 second long) of the natural movie. **f-j**, Same as (**a-e**) for Neuropixels dataset2.

Figure 4—figure supplement 2: Nonmonotonic relationship between setpoint similarity and behaviour.

a, Relationship between setpoint similarity and change in pupil size for the same example session as Figure 4g. **b**, Same data as (a), when setpoint similarity is plotted as a function of both changes in pupil size and change in running speed between movie repeats. Red datapoints indicated movie repeats with small difference of average running speed between them (Δ run. speed < 10 cm/s), and blue datapoints indicate movie repeats with large changes of average pupil width (|Δ pupil width| > 1.25). Black: other data points. **c**, Same as (a) for red datapoints in (b) (for movie repeats with small changes of average running speed between them). **d**, Average pupil width (normalized by z-scoring) during each movie presentation in the first (repeat #1-30) and the second (repeat #31-60) blocks of presentations. Red and blue bars show, respectively, the movie presentations above (> -0.8; red dashed line) and below (< -1.2; blue dashed line) the average of the pupil size in the first block. **e**, Same as (a) when setpoint similarity between specific pairs of movie repeats are colour coded. For pairs of movie repeats with one in the first block and the other in the second block, setpoint similarity is colour coded according to the average pupil width during the movie repeat in the first block, with red and blue circles corresponding to blue and red bars in (d), respectively. **f**, Shift in the

eye position across movie presentations. ΔX and ΔY denote deviation of the average centre of the pupil (X and Y) from their grand average value during the whole session.

Figure 4—figure supplement 3: Changes in the pupil centre, and their relation to setpoint similarity and changes in pupil width (for sessions in Neuropixels dataset1).

First column: Same as Figure 4—figure supplement 2f, showing shift in the eye position across movie presentations for each session (session # indicated on the top).

Second and third columns: Same as Figure 4—figure supplement 2a, when setpoint similarity is plotted against changes in the average eye position between movie repeats.

Fourth column: Relation between changes in eye position and changes in pupil width.

Figure 4—figure supplement 4: Changes in the pupil centre, and their relation to setpoint similarity and changes in pupil width (for sessions in Neuropixels dataset2).

Same as Figure 4—figure supplement 3 for Neuropixels dataset2.

Figure 4—figure supplement 5: Overall average activity of units and average pupil width gradually increase during the recording session.

Average activity of units (across all recorded units and all sessions/animals) across time during the recording session. Each recording session (~2.5 hours) is broken to 10 time bins and the mean value within that time bin is calculated. The average z-scored value of pupil width (averaged across animals and within the same time bins) is shown on the right (red bars). (**a**) Neuropixels dataset1, (**b**) Neuropixels dataset2. Note that for both datasets the average activity increases gradually from ~6 spikes/s to ~9 spikes/s at the end (~50% increase). Also, note the similar, gradual increase of pupil width.

Figure 5—figure supplement 1: Significant modulation of units by behaviour.

a, Left: Distribution of correlation of V1 units' activity with running during first and second blocks of presentation of natural movie 1. Right: Distribution of bootstrapped correlations with running. Correlation coefficient (CC) of each unit's activity with 100 randomly shuffled versions of the running speed is calculated. The z-score of bootstrapped correlation (Z) is calculated by subtracting the mean of this distribution from the unshuffled CC and dividing it by the std of the distribution (see Methods for details). Bootstrapped correlations are calculated during the first (grey) and second (black) blocks separately. Significant correlations are taken as units for which $|Z| > 2$ (indicated by dashed red lines). Fractions of significant correlations during the first and second

blocks are indicated on the top, respectively. **b**, same as (**a**) for all recorded units. **c,d**, Same as (**a,b**) for dataset2.

Figure 5—figure supplement 2: Consistent modulation of neuronal responses by behaviour across blocks of presentation of drifting gratings, and across stimuli.

a, Left: Distribution of correlations with running during the 1st and 2nd blocks of presentation of drifting gratings across all sessions. Right: Distribution of the z-score of bootstrapped correlations (Z) with running (see Methods and Figure 5—figure supplement 1). Significant correlations with running are defined as $|Z| > 0.2$. **b**, Correlation with running of units during the 2nd block against the 1st block, for all units and sessions (left; red), and for selected units (right; magenta), where sessions with similar levels of running between the two blocks and units with significant correlations are selected. **c**, Upper: Average running during the 1st and 2nd blocks for all sessions (all; red) and for selected units (sel; magenta). Lower: Correlation of all units with pupil versus their correlation with running, during the 1st (grey) and 2nd (black) blocks. Magenta: regression fits for selected units only. **d**, Correlation of all units with running speed during the presentation of drifting gratings versus correlations with running obtained during the presentation of natural movie 1, in the 1st and 2nd blocks of presentations, respectively. **e-h**, Same as (**a-d**) for dataset2.

Figure 5—figure supplement 3: Consistent modulation of neuronal responses by behaviour across stimuli, regions, and datasets.

a,b, Analysis of the reliability of behavioural tuning in two parts of each session in the two datasets. The composition of stimulus sets in each session type is shown, with the type, sequences, and the length of each stimulus presentation indicated. In both datasets, correlation of units with running speed is calculated in two parts: 1st part from 30 to 90 minutes, and the 2nd part from 90 to 150 minutes. **c,d**, Similar to Figure 5—figure supplement 2a-c for the 1st and 2nd part of the sessions in dataset1 (**c**) and dataset2 (**d**). **e**, Upper: Distribution of correlations with running for units recorded from different regions. Results for the two parts of the sessions (lighter lines denoting the first part) and both datasets (red: dataset1; blue: dataset2) are overlayed. Sessions where the average running between the two parts are too different are excluded (exclusion criteria: $|Z2-Z1|>0.3$, where Z1 and Z2 are the average of the z-scored value of running speed in the 1st and 2nd parts, respectively). Lower: Correlation with running in the 2nd part against the 1st part in each region, for dataset1 (red) and dataset2 (blue), respectively. Lines show the best fitted least-square regression lines, with numbers denoting the R^2 values of the fit in each case. **f**, Tuning reliability (average R^2 values in (**e**)), for different regions across the two datasets. Regions key: [visual cortex, VIS] VISp: primary visual

cortex; VISl: lateromedial area; VISrl: rostrolateral area; VISal: anterolateral area; VISam: posteromedial area; VISam: anteromedial area. [Hippocampal formation] CA1: cornu ammonis 1; CA3: cornu ammonis 3; DG: dentate gyrus; SUB: subiculum; ProS: prosubiculum. [Thalamus] LGd: lateral geniculate nucleus; LP: lateral posterior nucleus. [Midbrain] APN: anterior pretectal nucleus.

Figure 7—figure supplement 1: Decoding natural images does not improve by focusing on behaviourally modulated units.

Same as Figure 7c,d when reliable (Rel.) units are chosen as units with strong behavioural modulation (correlation with running speed of more than 0.5), instead of units with strong stimulus reliability (cf. Figure 7). Relation between average d' and the number of units available for decoding in each session (all units or behaviourally reliable units) is plotted on the bottom.

Figure 7—figure supplement 2: Decoding behavioural states from the population activity.

a, A linear decoder is trained to predict the pupil width (upper) and the running speed (lower) of the animal from the activity of all recorded units in each session. It is trained on the data from half of randomly chosen movie presentations from both blocks (not shown), and tested on the other half of movie presentations (orange traces). The target values (test) for a specific randomization are shown in blue (note that the order of movie repeats is randomized, while the order of movie frames within each movie presentation is preserved). The quality of prediction is quantified by the correlation coefficient (CC) of test and predicted traces in each case (indicated for the example sessions on top of each behaviour, respectively). Example sessions are the same as in Figure 1 and Figure 7. **b**, Average correlation of the predictions with behaviour for multiple repeats of the decoding with different random choices of train and test sets, for the example sessions shown in (a) (100 repeats) and for all sessions (10 repeats per each session). The mean and std of the correlations across repeats are plotted. **c**, The optimal weights obtained from the decoders in (b) are used to weight the neural activity and obtain a readout (similar to predictions in (a)). The readout activity is then used to assay the stimulus-related content of the behavioural decoder. A matrix of correlation coefficient of the readout activity between different repeats of the movie is calculated. The average of the off-diagonal entries of the matrix is taken as the correlation of reconstructed activity across multiple stimulus presentations. The mean and std of this across multiple randomisations (as in (b)) are plotted in (c). Left: Neuropixels dataset1; Right: Neuropixels dataset2.

Supplementary Table 1: Information of recording sessions in different datasets.

Figure 1—source data 1

Related to Figure 1.

Source data (for normalized changes in pupil width and representational similarity between pairs of movie repeats) for individual sessions across the two Neuropixels datasets.

Figure 2—source data 1

Related to Figure 2.

Source data (for stimulus reliability of V1 units and their modulation by pupil size) for individual sessions across the two Neuropixels datasets.

Figure 4—source data 1

Related to Figure 4.

Source data (for the change in average z-scored pupil width and setpoint similarity between pairs of movie repeats) for individual sessions across the two Neuropixels datasets.

Figure 6—source data 1

Related to Figure 6.

Source data (for the average z-scored values of pupil width and running speed in each block) for individual sessions across the two Neuropixels datasets.

Source Codes for Figures 1-4 are provided.

References

Arieli a, Sterkin a, Grinvald a, Aertsen a. 1996. Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses., Science (New York, N.Y.).

Aschauer DF, Eppler JB, Ewig L, Chambers AR, Pokorny C, Kaschube M, Rumpel S. 2022. Learning-induced biases in the ongoing dynamics of sensory representations predict stimulus generalization. *Cell Rep* **38**:110340. doi:10.1016/J.CELREP.2022.110340

Bimbard C, Sit TP, Lebedeva A, Harris KD, Carandini M. 2021. Behavioral origin of sound-evoked activity in visual cortex. *bioRxiv* 2021.07.01.450721. doi:10.1101/2021.07.01.450721

Boyd AM, Sturgill JF, Poo C, Isaacson JS. 2012. Cortical Feedback Control of Olfactory Bulb Circuits. *Neuron* **76**. doi:10.1016/j.neuron.2012.10.020

Bradley MM, Miccoli L, Escrig MA, Lang PJ. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* **45**. doi:10.1111/j.1469-8986.2008.00654.x

Buzsáki G. 2019. The Brain from Inside Out, The Brain from Inside Out. doi:10.1093/oso/9780190905385.001.0001

Clopath C, Bonhoeffer T, Hübener M, Rose T. 2017. Variance and invariance of neuronal long-term representations. *Philos Trans R Soc B Biol Sci*. doi:10.1098/rstb.2016.0161

Cohen-Kashi Malina K, Tsivourakis E, Kushinsky D, Apelblat D, Shtiglitz S, Zohar E, Sokoletsky M, Tasaka G, Mizrahi A, Lampl I, Spiegel I. 2021. NDNF interneurons in layer 1 gain-modulate whole cortical columns according to an animal's behavioral state. *Neuron*. doi:10.1016/j.neuron.2021.05.001

Cowley BR, Snyder AC, Acar K, Williamson RC, Yu BM, Smith MA. 2020. Slow Drift of Neural Activity as a Signature of Impulsivity in Macaque Visual and Prefrontal Cortex. *Neuron* **108**. doi:10.1016/j.neuron.2020.07.021

de Vries SEJ, Lecoq JA, Buice MA, Groblewski PA, Ocker GK, Oliver M, Feng D, Cain N, Ledochowitsch P, Millman D, Roll K, Garrett M, Keenan T, Kuan L, Mihalas S, Olsen S, Thompson C, Wakeman W, Waters J, Williams D, Barber C, Berbesque N, Blanchard B, Bowles N, Caldejon SD, Casal L, Cho A, Cross S, Dang C, Dolbeare T, Edwards M, Galbraith J, Gaudreault N, Gilbert TL, Griffin F, Hargrave P, Howard R, Huang L, Jewell S, Keller N, Knoblich U, Larkin JD, Larsen R, Lau C, Lee E, Lee F, Leon A, Li L, Long F, Luviano J, Mace K, Nguyen T, Perkins J, Robertson M, Seid S, Shea-Brown E, Shi J, Sjoquist N, Slaughterbeck C, Sullivan D, Valenza R, White C, Williford A, Witten DM, Zhuang J, Zeng H, Farrell C, Ng L, Bernard A, Phillips JW, Reid RC, Koch C. 2020. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nat Neurosci* **23**. doi:10.1038/s41593-019-0550-9

Deitch D, Rubin A, Ziv Y. 2021. Representational drift in the mouse visual cortex. *Curr Biol*. doi:10.1016/J.CUB.2021.07.062

Dipoppa M, Ranson A, Krumin M, Pachitariu M, Carandini M, Harris KD. 2018. Vision and Locomotion Shape the Interactions between Neuron Types in Mouse Visual Cortex. *Neuron* **98**. doi:10.1016/j.neuron.2018.03.037

Driscoll LN, Pettit NL, Minderer M, Chettih SN, Harvey CD. 2017. Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell* **170**. doi:10.1016/j.cell.2017.07.021

Emiliani V, Cohen AE, Deisseroth K, Häusser M. 2015. All-Optical Interrogation of Neural Circuits. *J Neurosci* **35**:13917–13926. doi:10.1523/JNEUROSCI.2916-15.2015

Ferguson KA, Cardin JA. 2020. Mechanisms underlying gain modulation in the cortex. *Nat Rev Neurosci* **21**:80–92. doi:10.1038/s41583-019-0253-y

Fiser J, Chiu C, Weliky M. 2004. Small modulation of ongoing cortical dynamics by sensory input during natural vision. *Nature* **431**:573.

Froudarakis E, Berens P, Ecker AS, Cotton RJ, Sinz FH, Yatsenko D, Saggau P, Bethge M, Tolias AS. 2014. Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nat Neurosci 2014 176* **17**:851–857. doi:10.1038/nn.3707

Fu Y, Tucciarone JM, Espinosa JS, Sheng N, Darcy DP, Nicoll RA, Huang ZJ, Stryker MP. 2014. A cortical circuit for gain control by behavioral state. *Cell* **156**:1139–1152. doi:10.1016/j.cell.2014.01.050

Garcia del Molino LC, Yang GR, Mejias JF, Wang X-J. 2017. Paradoxical response reversal of top-down modulation in cortical circuits with three interneuron types. *Elife* **6**. doi:10.7554/eLife.29742

Gonzalez WG, Zhang H, Harutyunyan A, Lois C. 2019. Persistence of neuronal representations through time and damage in the hippocampus. *Science (80-)* **365**:821–825. doi:10.1126/SCIENCE.AAV9199/SUPPL_FILE/AAV9199S9.MP4

Hennig JA, Oby ER, Golub MD, Bahureksa LA, Sadtler PT, Quick KM, Ryu SI, Tyler-Kabara EC, Batista AP, Chase SM, Yu BM. 2021. Learning is shaped by abrupt changes in neural engagement. *Nat Neurosci* **24**. doi:10.1038/s41593-021-00822-8

Joshi S, Gold JI. 2020. Pupil Size as a Window on Neural Substrates of Cognition. *Trends Cogn Sci*. doi:10.1016/j.tics.2020.03.005

Jun JJ, Steinmetz NA, Siegle JH, Denman DJ, Bauza M, Barbarits B, Lee AK, Anastassiou CA, Andrei A, Aydin Ç, Barbic M, Blanche TJ, Bonin V, Couto J, Dutta B, Gratiy SL, Gutnisky DA, Häusser M, Karsh B, Ledochowitsch P, Lopez CM, Mitelut C, Musa S, Okun M, Pachitariu M, Putzeys J, Rich PD, Rossant C, Sun WL, Svoboda K, Carandini M, Harris KD, Koch C, O'Keefe J, Harris TD. 2017. Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**. doi:10.1038/nature24636

Kenet T, Bibitchkov D, Tsodyks M, Grinvald A, Arieli A. 2003. Spontaneously emerging cortical representations of visual attributes. *Nature* **425**:954–6. doi:10.1038/nature02078

Lütcke H, Margolis DJ, Helmchen F. 2013. Steady or changing? Long-term monitoring of neuronal population activity. *Trends Neurosci*. doi:10.1016/j.tins.2013.03.008

Markopoulos F, Rokni D, Gire DH, Murthy VN. 2012. Functional Properties of Cortical Feedback Projections to the Olfactory Bulb. *Neuron* **76**. doi:10.1016/j.neuron.2012.10.028

Marks TD, Goard MJ. 2021. Stimulus-dependent representational drift in primary visual cortex. *Nat Commun 2021 121* **12**:1–16. doi:10.1038/s41467-021-25436-3

Marks TD, Goard MJ. 2020. Stimulus-dependent representational drift in primary visual cortex. *bioRxiv*.

Montijn JS, Meijer GT, Lansink CS, Pennartz CMA. 2016. Population-Level Neural Codes Are Robust to Single-Neuron Variability from a Multidimensional Coding Perspective. *Cell Rep* **16**. doi:10.1016/j.celrep.2016.07.065

Musall S, Kaufman MT, Juavinett AL, Gluf S, Churchland AK. 2019. Single-trial neural dynamics are dominated by richly varied movements. *Nat Neurosci* **22**. doi:10.1038/s41593-019-0502-4

Niell CM, Stryker MP. 2010. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron* **65**:472–479.

Niell CM, Stryker MP. 2008. Highly selective receptive fields in mouse visual cortex. *J Neurosci* **28**:7520–36. doi:10.1523/JNEUROSCI.0623-08.2008

Nilchian P, Wilson MA, Sanders H. 2022. Animal-to-Animal Variability in Partial Hippocampal Remapping in Repeated Environments. *J Neurosci* **42**:5268–5280. doi:10.1523/JNEUROSCI.3221-20.2022

Pakan JMP, Lowe SC, Dylda E, Keemink SW, Currie SP, Coutts CA, Rochefort NL. 2016. Behavioral-state modulation of inhibition is context-dependent and cell type specific in mouse visual cortex. *Elife* **5**. doi:10.7554/eLife.14985

Rubin A, Geva N, Sheintuch L, Ziv Y. 2015. Hippocampal ensemble dynamics timestamp events in long-term memory. *Elife* **4**. doi:10.7554/ELIFE.12247

Rule ME, O'Leary T, Harvey CD. 2019. Causes and consequences of representational drift. *Curr Opin Neurobiol*. doi:10.1016/j.conb.2019.08.005

Saleem AB, Ayaz A, Jeffery KJ, Harris KD, Carandini M. 2013. Integration of visual motion and locomotion in mouse visual cortex. *Nat Neurosci* **16**:1864–1869.

Schoonover CE, Ohashi SN, Axel R, Fink AJP. 2021. Representational drift in primary olfactory cortex. *Nature* **594**. doi:10.1038/s41586-021-03628-7

Siegle JH, Jia X, Durand S, Gale S, Bennett C, Graddis N, Heller G, Ramirez TK, Choi H, Luviano JA, Groblewski PA, Ahmed R, Arkhipov A, Bernard A, Billeh YN, Brown D, Buice MA, Cain N, Caldejon S, Casal L, Cho A, Chvilicek M, Cox TC, Dai K, Denman DJ, de Vries SEJ, Dietzman R, Esposito L, Farrell C, Feng D, Galbraith J, Garrett M, Gelfand EC, Hancock N, Harris JA, Howard R, Hu B, Hytnen R, Iyer R, Jessett E, Johnson K, Kato I, Kiggins J, Lambert S, Lecoq J, Ledochowitsch P, Lee JH, Leon A, Li Y, Liang E, Long F, Mace K, Melchior J, Millman D, Mollenkopf T, Nayan C, Ng L, Ngo K, Nguyen T, Nicovich PR, North K, Ocker GK, Ollerenshaw D, Oliver M, Pachitariu M, Perkins J, Reding M, Reid D, Robertson M, Ronellenfitch K, Seid S, Slaughterbeck C, Stoecklin M, Sullivan D, Sutton B, Swapp J, Thompson C, Turner K, Wakeman W, Whitesell JD, Williams D, Williford A, Young R, Zeng H, Naylor S, Phillips JW, Reid RC, Mihalas S, Olsen SR, Koch C. 2021. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**. doi:10.1038/s41586-020-03171-x

Stringer C, Pachitariu M, Steinmetz N, Reddy CB, Carandini M, Harris KD. 2019. Spontaneous behaviors drive multidimensional, brainwide activity. *Science (80-)* **364**. doi:10.1126/science.aav7893

Tsodyks M, Kenet T, Grinvald A, Arieli A. 1999. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. *Science* **286**:1943–6.

Vinck M, Batista-Brito R, Knoblich U, Cardin JA. 2015. Arousal and locomotion make distinct contributions to cortical activity patterns and visual encoding. *Neuron* **86**:740–754.

Yizhar O, Fenno LE, Davidson TJ, Mogri M, Deisseroth K. 2011. Optogenetics in Neural Systems. *Neuron* **71**:9–34. doi:10.1016/j.neuron.2011.06.004

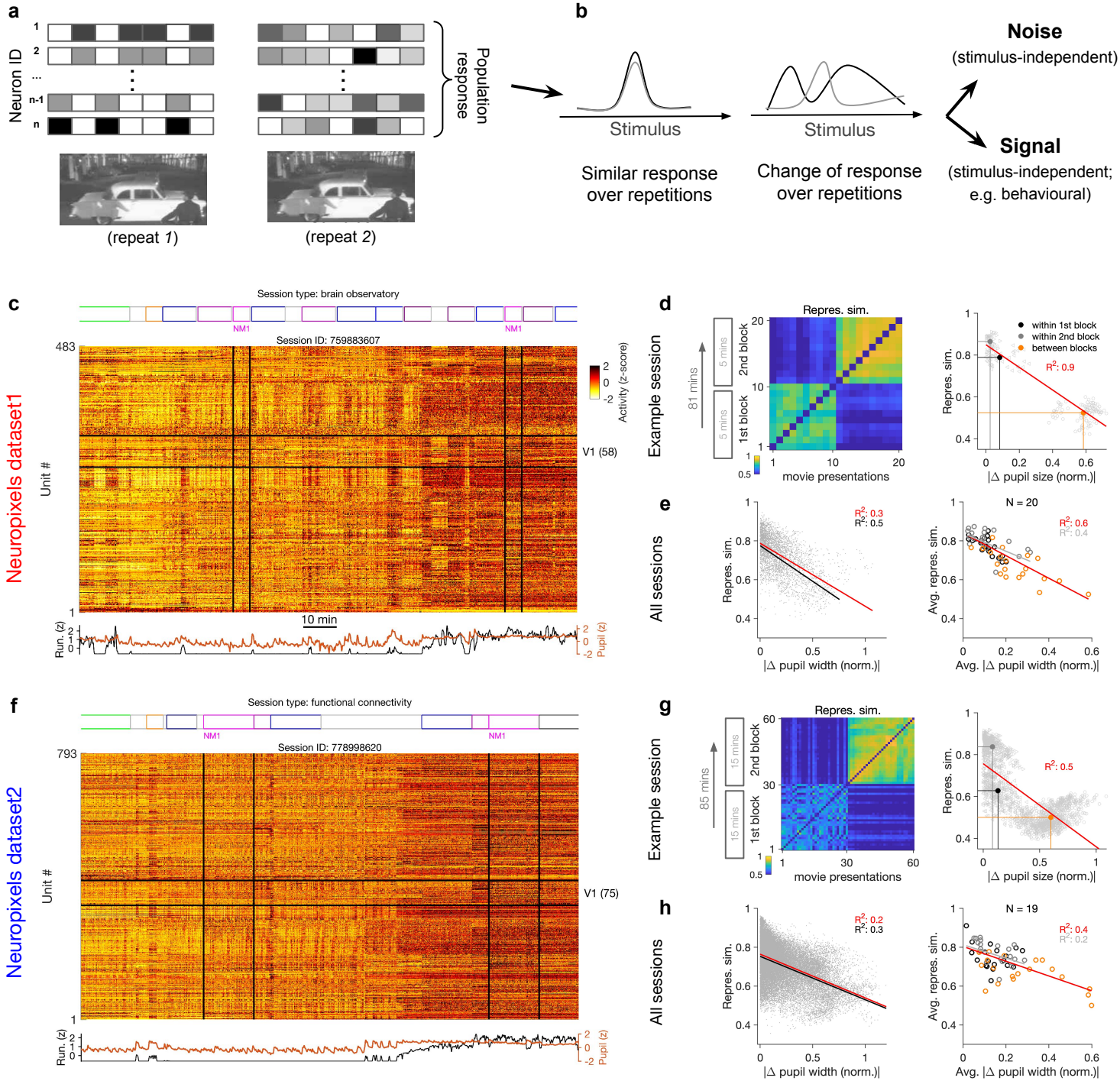
Yoshida T, Ohki K. 2020. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nat Commun* **11**:872. doi:10.1038/s41467-020-14645-x

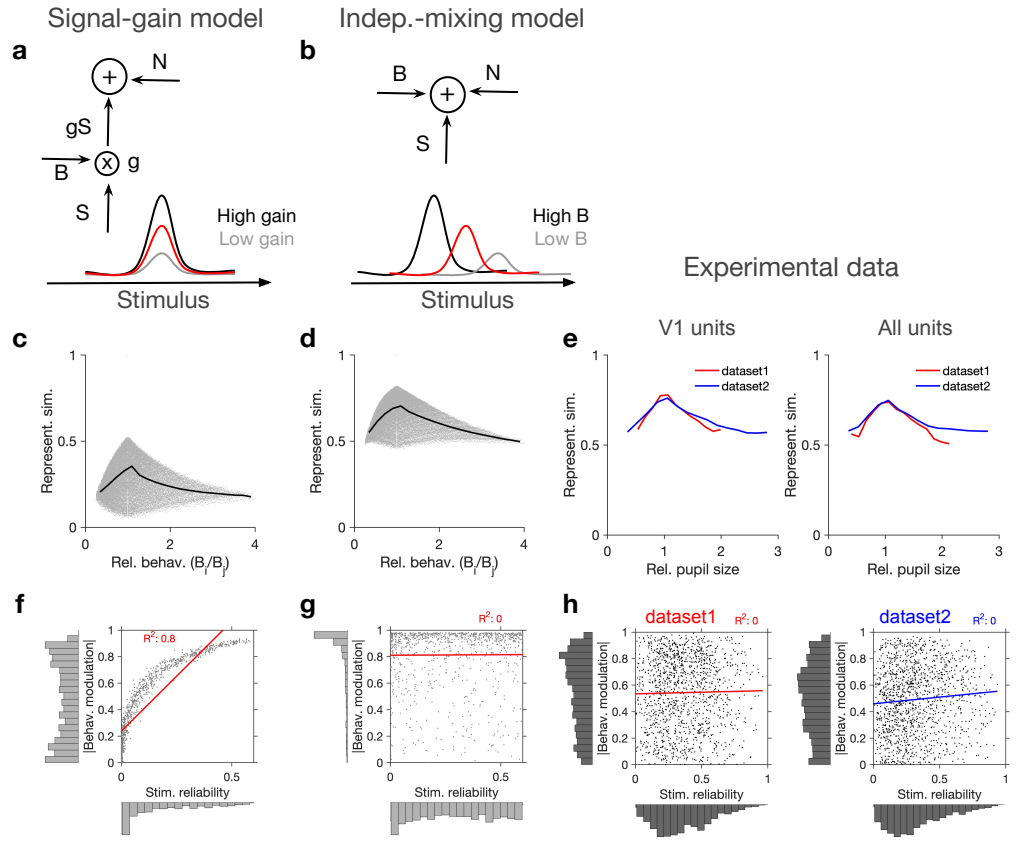
Zhang Z, Russell LE, Packer AM, Gauld OM, Häusser M. 2018. Closed-loop all-optical interrogation of neural circuits in vivo. *Nat Methods* **15**:1037–1040. doi:10.1038/s41592-018-0183-z

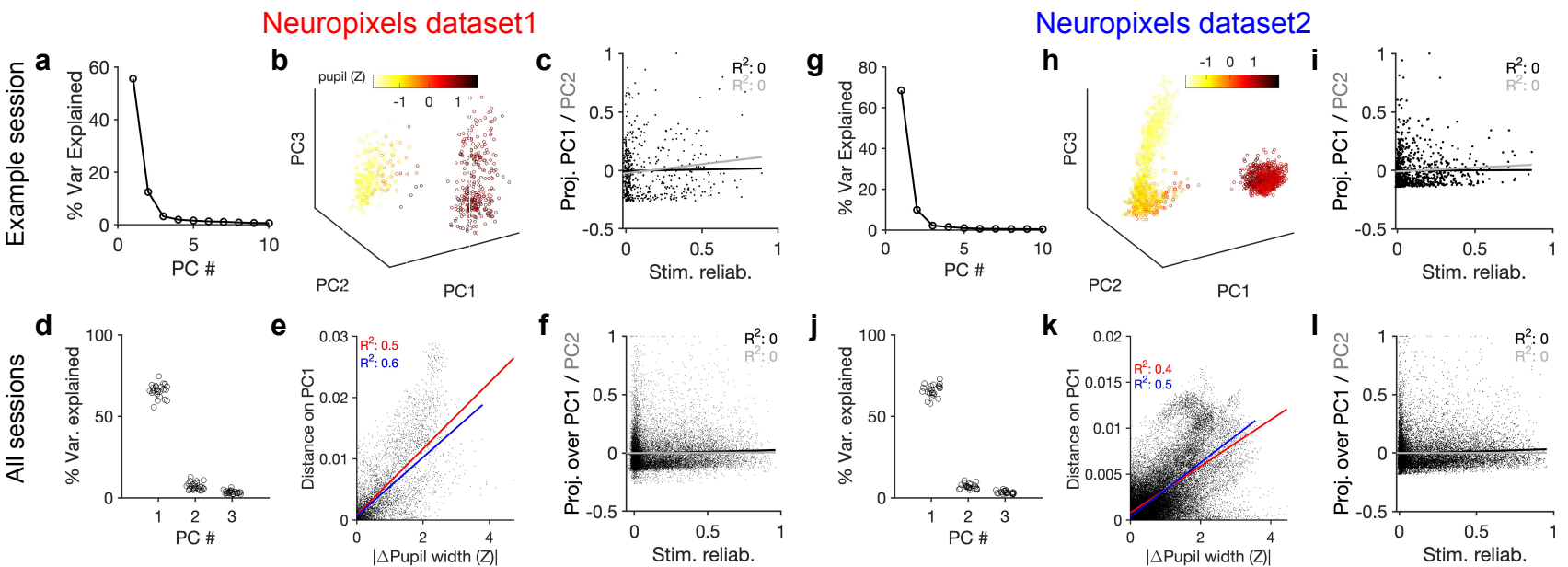
Ziv Y, Burns LD, Cocker ED, Hamel EO, Ghosh KK, Kitch LJ, Gamal A El, Schnitzer MJ. 2013. Long-term dynamics of CA1 hippocampal place codes. *Nat Neurosci* **16**. doi:10.1038/nn.3329

Acknowledgement

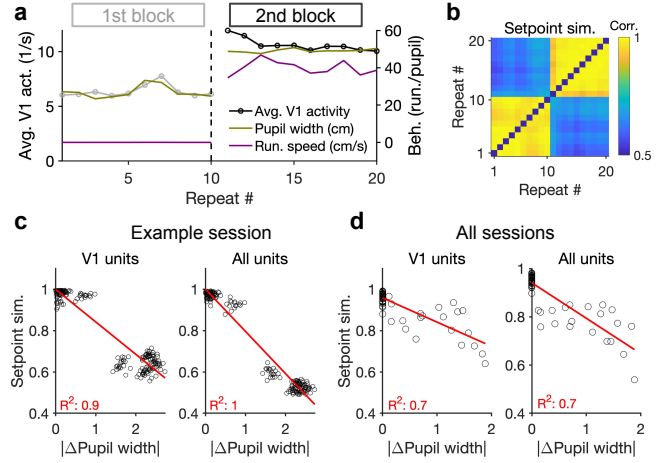
We thank K.A. Wilmes and D.F. Tome for their comments on the manuscript, and all the members of the Clopath Lab for insightful discussions. This work was supported by BBSRC BB/N013956/1, BB/N019008/1, Wellcome Trust 225412/Z/22/Z, Simons Foundation 564408 and EPSRC EP/R035806/1. All data needed to evaluate the conclusions in the paper are presented in the paper and/or the Supplementary Materials. Codes for reproducing the main analyses are provided.


a
Neuron ID
1
2
n-1
n
Population response
(repeat 1)
(repeat 2)
b
Stimulus
Stimulus
Similar response over repetitions
Change of response over repetitions
Noise
(stimulus-independent)
Signal
(stimulus-independent; e.g. behavioural)
c
Neuropixels dataset1
Session type: brain observatory
NM1
NM1
Session ID: 759883607
483
Unit #
V1 (58)
Activity (z-score)
10 min
Run. (z)
Pupil (z)
d
Example session
Repres. sim.
81 mins
5 mins
5 mins
2nd block
1st block
movie presentations
within 1st block
within 2nd block
between blocks
R^2: 0.9
Repres. sim.
|Δ pupil size (norm.)|
e
All sessions
R^2: 0.3
R^2: 0.5
Repres. sim.
|Δ pupil width (norm.)|
N = 20
R^2: 0.6
R^2: 0.4
Avg. repres. sim.
Avg. |Δ pupil width (norm.)|
f
Neuropixels dataset2
Session type: functional connectivity
NM1
NM1
Session ID: 778998620
793
Unit #
V1 (75)
Run. (z)
Pupil (z)
g
Example session
Repres. sim.
85 mins
15 mins
15 mins
2nd block
1st block
movie presentations
R^2: 0.5
Repres. sim.
|Δ pupil size (norm.)|
h
All sessions
R^2: 0.2
R^2: 0.3
Repres. sim.
|Δ pupil width (norm.)|
N = 19
R^2: 0.4
R^2: 0.2
Avg. repres. sim.
Avg. |Δ pupil width (norm.)|

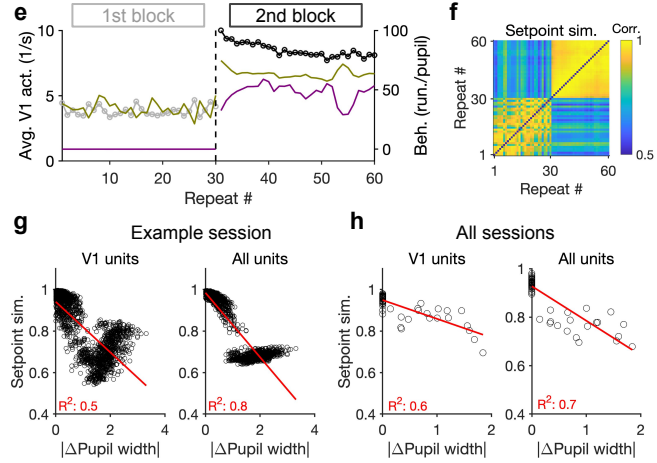

Signal-gain model
Indep.-mixing model
a
N
gS
B
g
S
High gain
Low gain
Stimulus
b
B
N
S
High B
Low B
Stimulus
Experimental data
c
Represent. sim.
Rel. behav. (Bi/Bj)
d
Represent. sim.
Rel. behav. (Bi/Bj)
e
V1 units
All units
dataset1
dataset2
Represent. sim.
Rel. pupil size
f
R2: 0.8
|Behav. modulation|
Stim. reliability
g
R2: 0
|Behav. modulation|
Stim. reliability
h
dataset1
R2: 0
dataset2
R2: 0
|Behav. modulation|
Stim. reliability


Neuropixels dataset1
Neuropixels dataset2
Example session
All sessions
a
b
c
d
e
f
g
h
i
j
k
l
% Var Explained
PC #
pupil (Z)
-1 0 1
PC3
PC2
PC1
Proj. PC1 / PC2
Stim. reliab.
R2: 0
R2: 0
% Var. explained
Distance on PC1
|ΔPupil width (Z)|
R2: 0.5
R2: 0.6
Proj. over PC1 / PC2
R2: 0.4
R2: 0.5

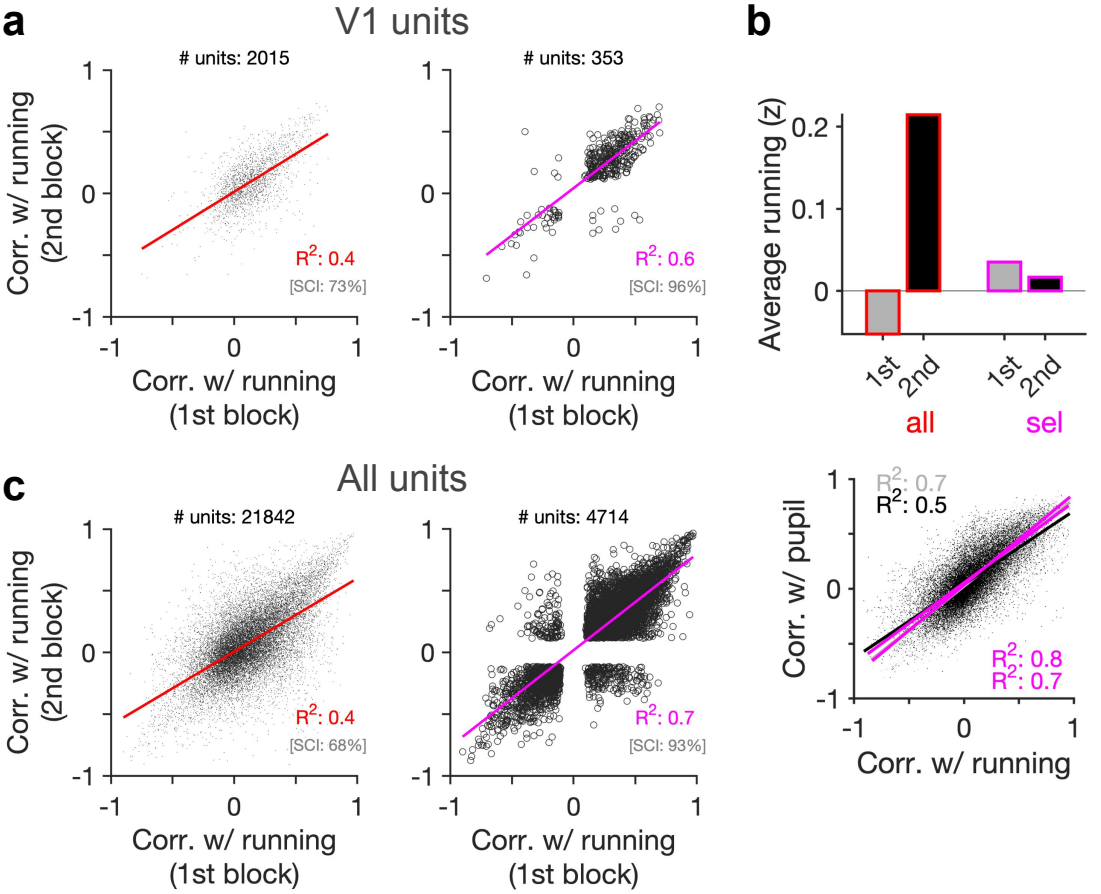
Neuropixels dataset1



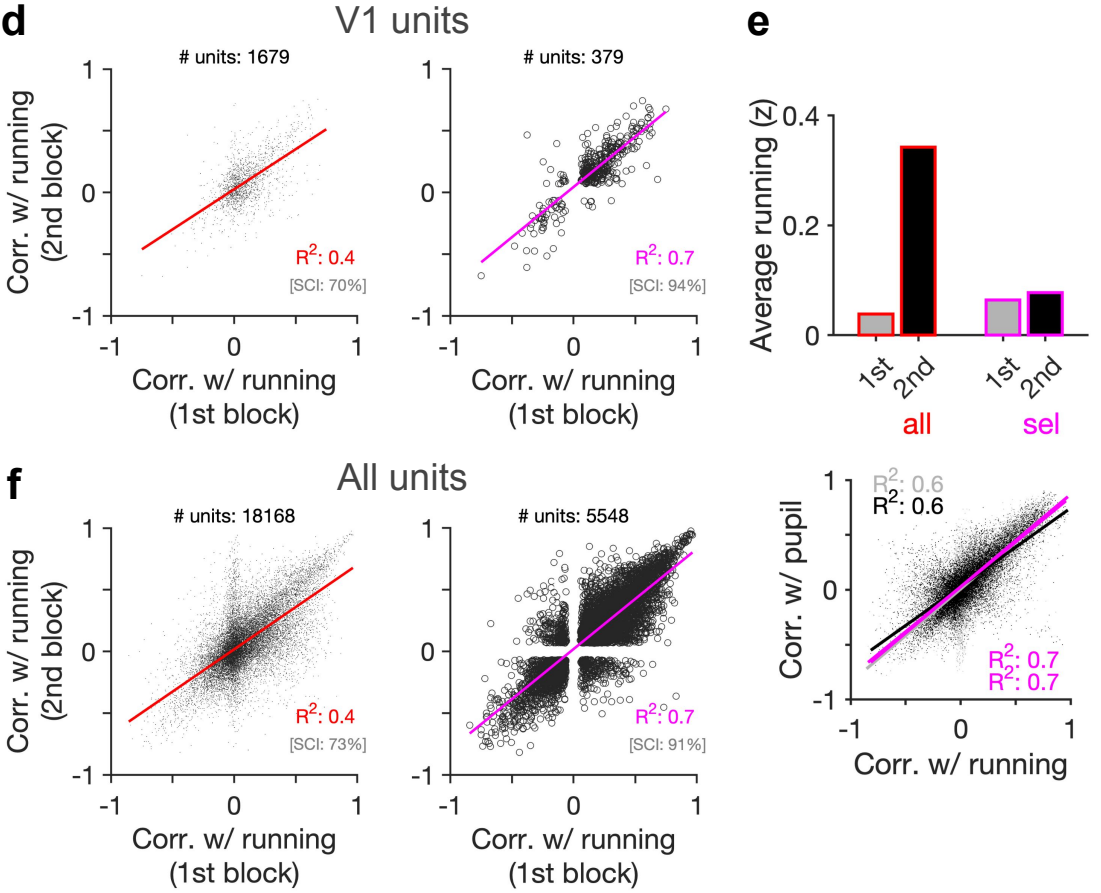
Neuropixels dataset2

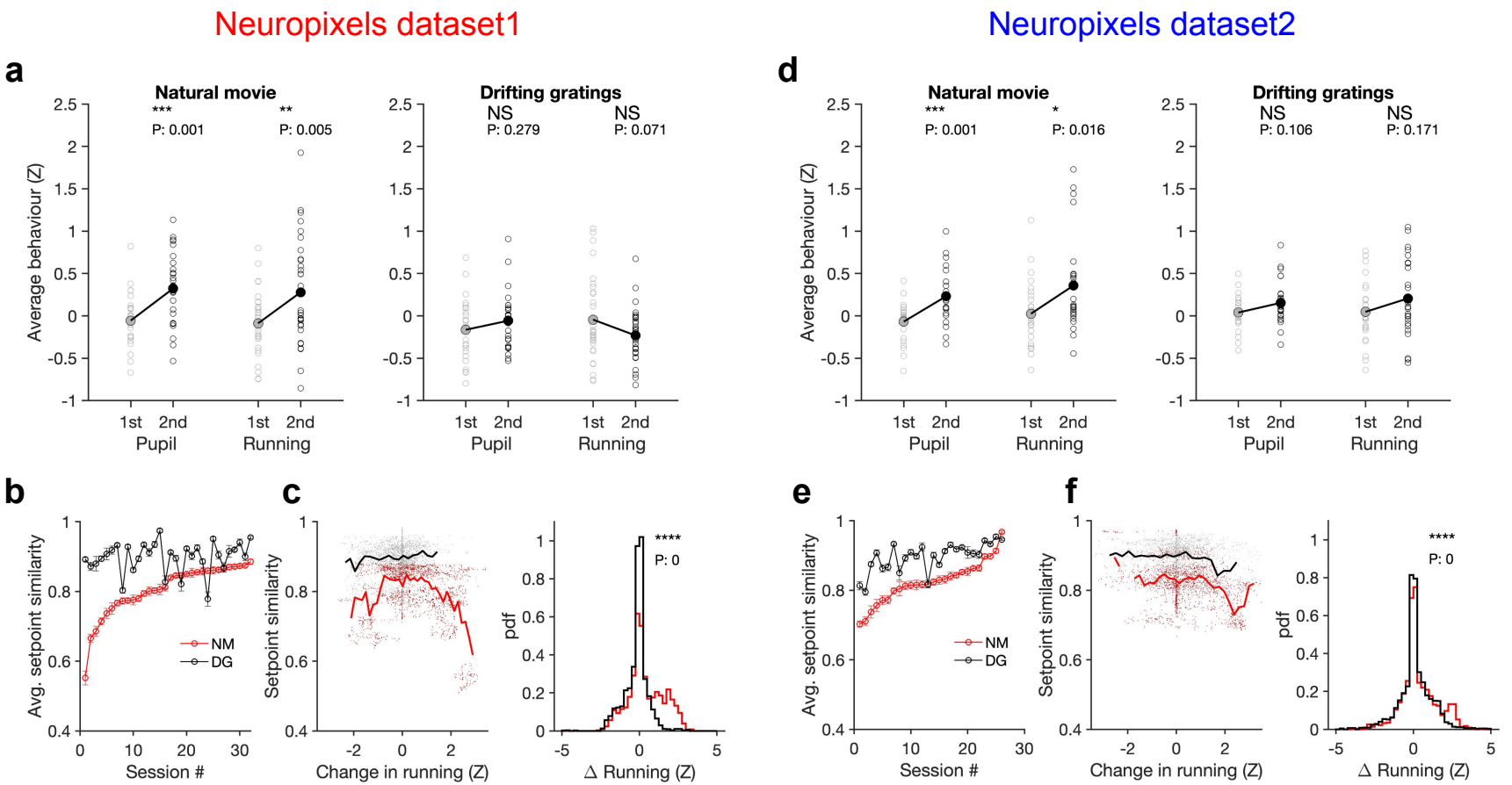


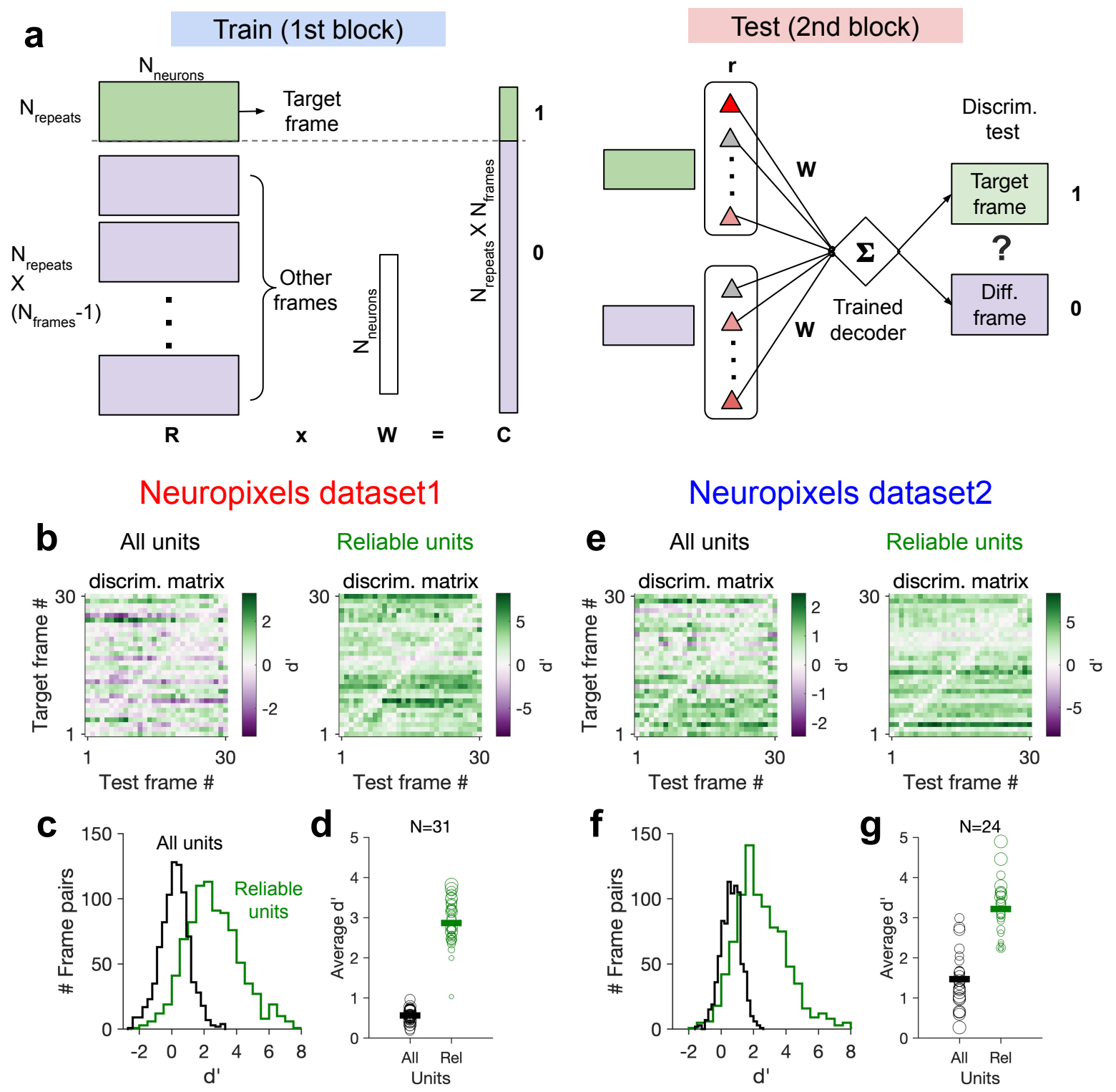
Neuropixels dataset1



Neuropixels dataset2




Neuropixels dataset1
Neuropixels dataset2
a
Natural movie
Drifting gratings
***
P: 0.001
**
P: 0.005
NS
P: 0.279
NS
P: 0.071
Average behaviour (Z)
1st
2nd
Pupil
Running
d
***
P: 0.001
*
P: 0.016
NS
P: 0.106
NS
P: 0.171
b
Avg. setpoint similarity
NM
DG
Session #
c
Setpoint similarity
Change in running (Z)
****
P: 0
pdf
Δ Running (Z)
e
f


a
Train (1st block)
Test (2nd block)
N neurons
N repeats
Target frame
1
N repeats X (N frames -1)
Other frames
N neurons
N repeats X N frames
0
R
x
W
=
C
r
W
Σ
Trained decoder
Discrim. test
Target frame
1
?
Diff. frame
0
Neuropixels dataset1
Neuropixels dataset2
b
All units
Reliable units
discrim. matrix
Target frame #
Test frame #
d'
c
All units
Reliable units
Frame pairs
d'
d
N=31
Average d'
All
Rel
Units
e
f
g
N=24