# Task learning reveals signatures of sample-based internal models in rodent prefrontal cortex

Abhinav Singh[1], Adrien Peyrache[2], Mark D. Humphries[1]

1. Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT, United Kingdom

2. The Neuroscience Institute, School of Medicine, New York University, New York, New York, USA.

**Contact:** Correspondence should be addressed to M.D.H (mark.humphries@manchester.ac.uk)

**Abstract**

The idea that brains use probabilistic internal models of the world is a powerful explanation for a range of behavioural phenomena. Unclear is whether or how neurons represent such models in higher cortical regions, learn them, and use them in behaviour. To address these issues, we sought evidence for the learning of internal models by cortical neurons during a behavioural task. Using a sampling framework, we predicted that trial-evoked and sleeping population activity represent the inferred and expected probabilities generated from an internal model of the task, and would become more similar as the task was learnt. To test these predictions, we analysed population activity from rodent prefrontal cortex before, during, and after sessions of learning rules on a maze. Distributions of activity patterns converged between trials and post-learning sleep during successful rule learning. Learning induced changes were greatest for patterns predicting correct choice and expressed at the choice point of the maze, consistent with an updated internal model of the task. Our results suggest sample-based internal models are a general computational principle of cortex.

## Introduction

How do we know what state the world is in? Behavioural evidence suggests brains solve this problem using probabilistic reasoning (Kording and Wolpert, 2004; Pouget et al., 2013). Such reasoning implies that brains represent and learn internal models for the statistical structure of the external world (Wolpert et al., 1995; Dayan and Abbot, 2001; Kording and Wolpert, 2004). With these models, neurons could represent uncertainty about the world with probability distributions, and update those distributions with new knowledge using the rules of probabilistic inference. Theoretical work has elucidated potential mechanisms for how cortical populations represent and compute with probabilities (Zemel et al., 1998; Ma et al., 2006; Buesing et al., 2011; Pouget et al., 2013; Kappel et al., 2015), and shown how computational models of inference predict aspects of cortical activity in sensory and decision-making tasks (e.g. Beck et al., 2008; Pouget et al., 2013). But experimental evidence for the neural basis of probabilistic reasoning, and the underlying internal models, is lacking.

An experimentally-accessible proposal is the recent inference-by-sampling hypothesis (Fiser et al., 2010; Berkes et al., 2011). This proposes that cortical population activity

at some time $t$ is a sample from an underlying probability distribution, which can be reconstructed by integrating over samples. Cortical activity evoked by external stimuli represents sampling from the model-generated "posterior" distribution that the world is in a particular state. Spontaneous cortical activity represents sampling of the model in the absence of external stimuli, forming a model-generated "prior" for the expected properties of the world. A key prediction is that the evoked and spontaneous population activity should converge over repeated experience, as the internal model adapts to match the relevant statistics of the external world. Just such a convergence has been observed in small populations from ferret V1 over development (Berkes et al., 2011). Unknown is whether neural inference is a general computational principle for cortex: whether it can be observed during learning, or in higher-order cortices, or during ongoing behaviour.

A natural candidate to address these issues is the medial prefrontal cortex (mPFC). Medial PFC is necessary for learning new rules or strategies (Ragozzino et al., 1999; Rich and Shapiro, 2007), and changes in mPFC neuron firing times correlates with successful rule learning (Benchenane et al., 2010), suggesting that mPFC coding of task-related variables changes over learning. We thus hypothesised that mPFC encodes an internal model of a task, which is updated by task performance, and from which the statistical distributions of population activity are generated. To test these hypotheses, we analysed previously-recorded population activity from the medial prefrontal cortex of rats learning rules in a Y-maze task (Peyrache et al., 2009).

# Results

Rats with implanted tetrodes learnt one of three rules on a Y-maze: go left, go right, or go to the randomly-lit arm (Figure 1A). Each recording session was a single day containing 3 epochs totalling typically 1.5 hours: pre-task sleep/rest, behavioural testing on the task, and post-task sleep/rest. We focussed on ten sessions where the animal reached the learning criteria for a rule mid-session (Experimental Procedures; 15-55 neurons per session). In this way, we sought to isolate changes in population activity solely due to rule-learning.

## Theory sketch

Here we outline our theoretical predictions for changes in population activity, derived from the inference-by-sampling hypothesis; a full account is given in Supplementary File 1. We sought to test the idea that the mPFC contains at least one internal model related to task performance, such as representing the relevant decision-variable (here, left or right) or the rule-dependent outcomes. Learning of the task should therefore update the internal model based on feedback from each trial's outcome. We theorised that mPFC population activity on each trial was sampling from the posterior distribution generated from this model; and that "spontaneous" activity in slow-wave sleep (SWS), occurring in the absence of task-related stimuli and behaviour, samples the corresponding prior distribution (Figure 1B). Consequently, updating the internal model from task feedback should be reflected in changes to the posterior and prior distributions generated from that model.

By restricting our analyses to sessions with successful learning, we expected the post-task SWS activity to be sampling from an internal model that has learnt the correct rule. To compare posterior distribution samples from the same internal model, we considered population activity during correct trials after the learning criteria were met – we call this distribution $P(R)$. Our main prediction was thus that the distribution $P(R)$ of activity

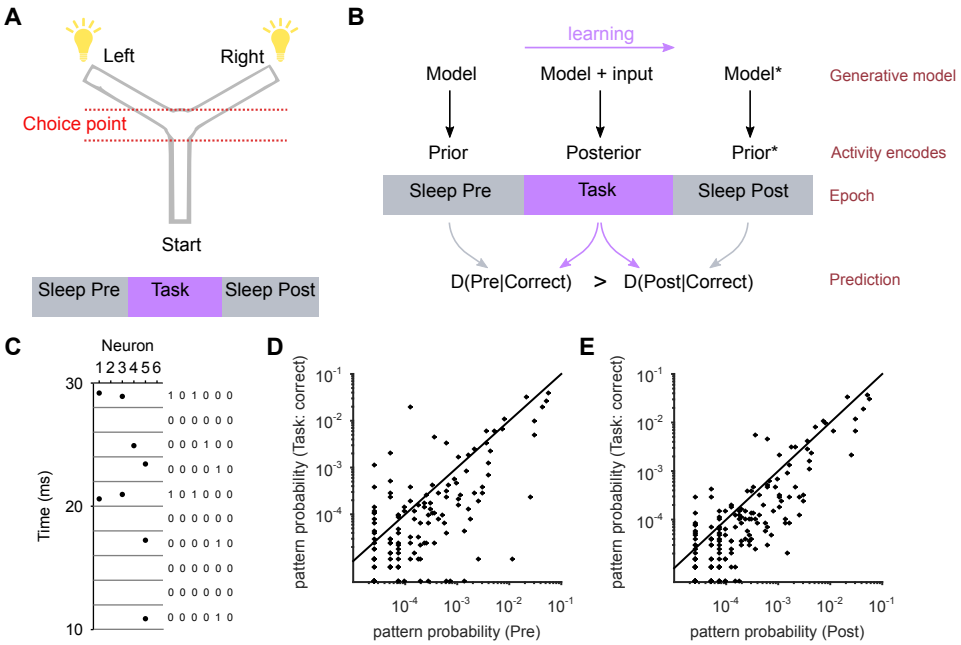


**Figure 1: Activity pattern distributions during rule-learning**. (A) Y-maze task set-up (top); each session included the epochs of pre-task sleep/rest, task trials, and post-task sleep/rest (bottom) - Figure 4A gives a breakdown per session. One of three target rules for obtaining reward was enforced throughout a session: go right; go left; go to the randomly-lit arm. (B) Schematic of theory. If prefrontal cortex encodes an internal model of the task, then activity during the task is derived from the internal model plus the relevant external inputs: the distribution of activity is thus the posterior distribution over the encoded task variables. During sleep, the distribution of activity is derived entirely from the internal model, and thus is the prior distribution over the encoded task variables. Updates to the internal model by task learning (creating Model*) will then change the prior distribution encoded during sleep (to Prior*). The theoretical prediction is then that the activity distribution in post-session sleep, derived from the model of the correct rule, will be closer to the distribution on the correct trials, compared to the pre-session sleep. (C) The population activity of simultaneously recorded spike trains was represented as a binary activity pattern in some small time-bin (here 2 ms). (D) Scatter plot of the joint frequency of every occurring pattern in pre-task SWS (distribution $P(Pre)$) and task (distribution $P(R)$) epochs for one session. (E) For the same session as (D), scatter plot of the joint frequency of every occurring pattern in post-task SWS [$P(Post)$] and task [$P(R)$] epochs.

during the correct trials would be more similar to the distribution in post-task SWS [$P(Post)$] than in pre-task SWS [$P(Pre)$]. Such a convergence of distributions would be evidence that a task-related internal model in mPFC was updated by feedback.

## Activity distributions converge between task and post-task sleep

To test these hypotheses, we compared the statistical distributions of activity patterns between task and sleep epochs. Activity patterns were characterised as a binary vector (or "word") of active and inactive neurons with a binsize of 2 ms (Figure 1C). Each recorded population of $N$ neurons had the same sub-set of all $2^N$ possible activity patterns in all epochs (Figure 1 - figure supplement 1) (Luczak et al., 2009; Wohrer et al., 2013). Such a common set of patterns is consistent with their being samples generated from the same form of internal model across both behaviour and sleep.

For each pair of epochs, we computed the distances between the two corresponding distributions of activity patterns (Figure 1D,E). We first used the information-theory based

Kullback-Liebler divergence to measure the distance $D(P|Q)$ between distributions $P$ and $Q$ in bits (Berkes et al., 2011). We found that in 9 of the 10 sessions the distribution $P(R)$ of activity during the trials was closer to the distribution in post-task SWS $[P(Post)]$ than in pre-task SWS $[P(Pre)]$ (Figure 2A).

On average the task-evoked distribution of patterns was $18.7 \pm 6.2\%$ closer to the post-task SWS distribution than the pre-task SWS distribution (Figure 2B), showing a convergence between task-evoked and post-task SWS distributions. Further, we found a robust convergence even at the level of individual sessions (Figure 2C).

While the Kullback-Liebler divergence provides the most complete characterisation of the distance between two probability distributions, estimating it accurately from limited sample data has known issues (Panzeri et al., 2007). To check our results were robust, we re-computed all distances using the Hellinger distance, a non-parametric measure that provides a lower bound for the Kullback-Liebler divergence. Reassuringly, we found the same results: the distribution $P(R)$ of activity during the trials was consistently closer to the distribution in post-task SWS $[P(Post)]$ than in pre-task SWS $[P(Pre)]$ (Figure 2F-H; the mean convergence between task-evoked and post-task SWS distributions was $21 \pm 2.8\%$).

The convergence between the task $P(R)$ and post-task SWS $P(Post)$ distributions was also robust to both the choice of activity pattern binsize (Figure 2 - figure supplement 1) and the choice of correct trials in the task distribution $P(R)$ (Figure 2 - figure supplement 2) .

Together, these results are consistent with the convergence over learning of the posterior and prior distributions represented by mPFC population activity. They imply that mPFC encodes a task-related internal model that is updated by task feedback.

## Convergence is a consequence of changes to correlations, not firing rates

Population firing rate differences between waking and sleep states, and increases in SWS firing after task-learning, could potentially account for the convergence of distributions (Okun et al., 2012; Fiser et al., 2013). To control for this, we used the "raster" model (Okun et al., 2012) to generate surrogate sets of spike-trains that matched both the mean firing rates of each neuron, and the distribution of total population activity in each time-bin ($K = 0, 1, \ldots, N$ spikes per bin). Consequently, the occurrence rates of particular activity patterns in the raster model are those predicted to arise from neuron and population firing rates alone.

We found that firing rates could not account for the convergence between task and post-task SWS distributions. The data-derived distance $D(Post|R)$ was always smaller than the distance $D(Post - model|R)$ predicted by the raster model (Figure 3A). This was true whether we used Kullback-Liebler divergence or the Hellinger distance (Figure 3C) to measure distances between distributions.

Our activity patterns are built from single units, unlike previous work using multi-unit activity (Schneidman et al., 2006; Berkes et al., 2011; Okun et al., 2012; Tkacik et al., 2014; Ganmor et al., 2015), so we expect our patterns to be sparse with rare synchronous activity. Indeed our data are dominated by activity patterns with $K = 0$ and $K = 1$ spikes (Figure 3 - figure supplement 1 . If all patterns were $K = 0$ or $K = 1$, the raster model spike trains would be exactly equivalent to the data. It is all the more surprising then that we found such a consistent lower distance for our data-derived distributions.

It follows that the true difference between data and model is in the relative occurrence of co-activation patterns with $K \geq 2$ spikes. To check this, we applied the same analysis to distributions built only from these co-activation patterns, drawn from data and from

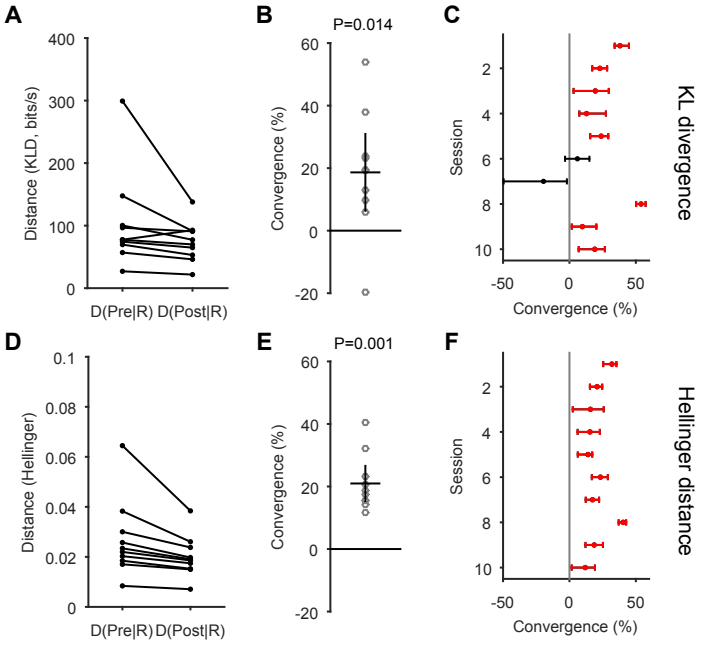


**Figure 2: Convergence of activity pattern distributions between the task and post-task sleep**. (A) Distances between the distributions of pattern frequencies in sleep and task epochs; one dot per session. $D(X|Y)$: distance between pattern distributions in epochs $X$ and $Y$: Pre: pre-task SWS; Post: post-task SWS; R: correct task trials. (B) Scatter of convergence across all sessions (circles). Convergence is $D(Pre|R) - D(Post|R)/D(Pre|R)$. A value greater than zero means that the activity pattern distribution in the task is closer to the distribution in post-task SWS than the distribution in pre-task SWS. Black lines give mean $\pm$ 2 s.e.m. (C) Data (dot) and 95% bootstrapped confidence interval (line) for the convergence of task and post-task SWS activity pattern distributions for each session. Red: sessions with CIs above 0. (D) - (F) As (A)-(C), using Hellinger distance. All $P$-values from 1-tailed Wilcoxon signrank test, with $N$=10 sessions.

the raster model fitted to the complete data. We found that the data-derived distance $D(Post|R)$ was always smaller than the distance $D(Post - model|R)$ predicted by the raster model (Figure 3B,D). Across all sessions, the model-predicted distance $D(Post - model|R)$ was between 3% and 46% greater than the data-derived distance $D(Post|R)$ using Kullback-Liebler divergence, indicating that much of the convergence between task and SWS distributions could not be accounted for by firing rates alone. Consequently, the changed distributions of activity patterns are due to changes in the correlations between neurons.

Reassuringly, for these $K \geq 2$ activity pattern distributions, all convergence results held (Figure 3 - figure supplement 2) despite the order-of-magnitude fewer sampled patterns.

## Convergence is not a recency effect

We examined periods of SWS in order to most likely observe the sampling of a putative internal model in a static condition, with no external inputs and minimal learning. But as correct task trials more likely occur towards the end of a session, this raises the possibility that the closer match between task and post-task SWS distributions are a recency effect, due to some trace or reverberation in sleep of the most recent task activity.

The time-scales involved make this unlikely. Bouts of SWS did not start until typically

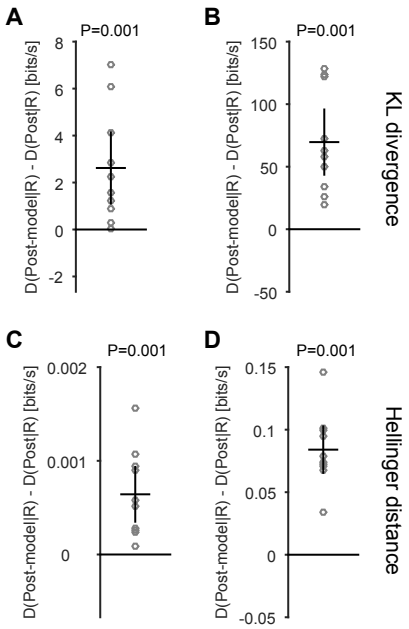


**Figure 3: Convergence is caused by changes in correlation, not firing rate**. (A) The distance between the task and post-task sleep distributions $D(Post|R)$ is always smaller than predicted by firing rate changes during sleep alone $D(Post - model|R)$, as given by the raster model. Black lines give mean $\pm$ 2 s.e.m in all panels. (B) As in (A), using only activity patterns with $K \geq 2$ spikes from data and model. (C)-(D) As (A)-(B), using Hellinger distance. All $P$-values from 1-tailed Wilcoxon signrank test, with $N$=10 sessions.

8 minutes after the end of the task (mean 397s; S.D. 188 s; Figure 4A). Any reverberation would thus have to last at least that long to appear in the majority of post-task SWS distributions.

The intervening period before the first bout of SWS contains quiet wakefulness and early sleep stages. If convergence was a recency effect, then we would expect that distributions $[P(Rest)]$ of activity patterns in this more-immediate "rest" epoch would also converge. We did not find this: across sessions, there was no evidence that the distribution in post-task rest $[P(Rest)]$ consistently converged on the distribution during task trials $[P(R)]$ (Figure 4B,C; mean convergence was $-8.7 \pm 18.7\%$). Not only is the observed convergence inconsistent with a recency effect, it seems also selective for activity in SWS.

## Distributions are updated by task-relevant activity patterns

The above analysis rests on the idea that the distributions of activity patterns are derived from an internal model of the task. This predicts that individual patterns should correlate with some aspect of the task. We sought an unbiased way of testing this prediction, so considered the following. In our theory, the changes to the internal model over learning should be directly reflected in the differences between the prior distributions before and after learning. Consequently, if we compare the sampling of activity patterns in pre-task sleep to sampling in post-task sleep, then any patterns with changes in their sampling should be from the updated model. This means that these patterns should encode some aspect of the task.

Remarkably, this is exactly what we found. For each co-activation pattern, we found its ability to predict a trial's outcome by its rate of occurrence on that trial (Fig 5A). When

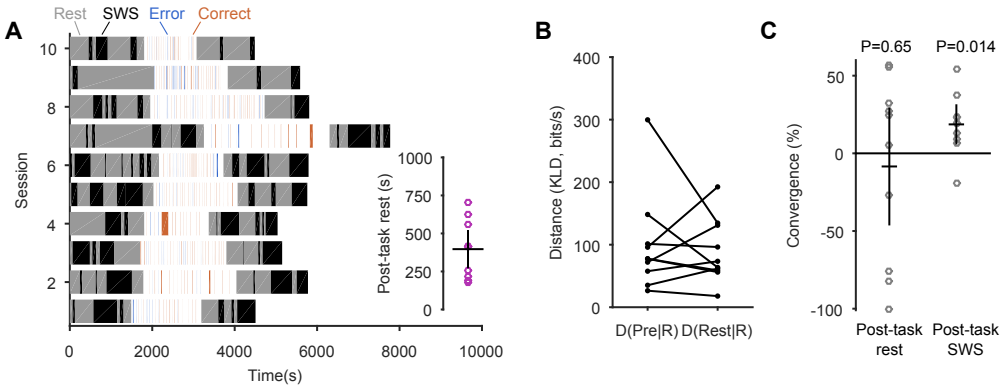


**Figure 4: Convergence is not a recency effect**. (A) Breakdown of each session into the duration of its state components. The task epoch is divided into correct (red) and error (blue) trials, and inter-trial intervals (white spaces). Trial durations were typically 2-4 seconds, so are thin lines on this scale. The pre- and post-task epochs contained quiet waking and light sleep states ("Rest" period) and identified bouts of slow-wave sleep ("SWS"). Inset: duration of the Rest period between the end of the last trial and the start of the first SWS bout (lines give mean $\pm$ 2 s.e.m.) (B) Distances between the distributions of pattern frequencies in different epochs; one dot per session. $D(X|Y)$: distance between pattern distributions in epochs $X$ and $Y$: Pre: pre-task SWS; Rest: immediate post-task rest period; R: correct task trials. Compare to Figure 2A. (C) Results from panel (B) expressed as the convergence between the distributions in the task and post-task rest period. We also re-plot here the convergence between the task and post-task SWS distributions. ($P$-values from 1-tailed Wilcoxon signrank test, with $N$=10 sessions).

we compared this outcome prediction to the change in sampling between pre- and post-task sleep, we found a strong correlation between the two (Figure 5B-D). This correlation was highly robust (Figure 5E-G). The learnt internal model, as evidenced by the updated patterns sampled from it, was seemingly encoding the task.

### Outcome-predictive patterns occur around the choice point

Consistent with the internal model being task-related, we further found that the outcome-predictive activity patterns preferentially occurred around the choice point of the maze (Figure 6A,B). Particularly striking was that patterns strongly predictive of outcome rarely occurred in the starting arm (Figure 6A). Together, the selective changes over learning to outcome-specific (Figure 5) and location-specific (Figure 6) activity patterns show that the convergence of distributions (Figure 1) is not a statistical curiosity, but is evidence for the updating of a behaviourally-relevant internal model.

## Discussion

Prefrontal cortex has been implicated in both planning and working memory during spatial navigation (Baeg et al., 2003; Fujisawa et al., 2008; Ito et al., 2015; Spellman et al., 2015), and executive control in general (Miller, 2000; Sul et al., 2010). Our results suggests a probabilistic basis for these functions. We find that moment-to-moment patterns of mPFC population activity change their sampling rates during learning of a spatial navigation task. Consequently, the statistical distributions of patterns in spontaneous and task-evoked activity converge. Our analyses thus suggest mPFC encodes a probabilistic internal model of a task, which is updated by behavioural outcomes, and uses population-activity sampling as the basis for inference.

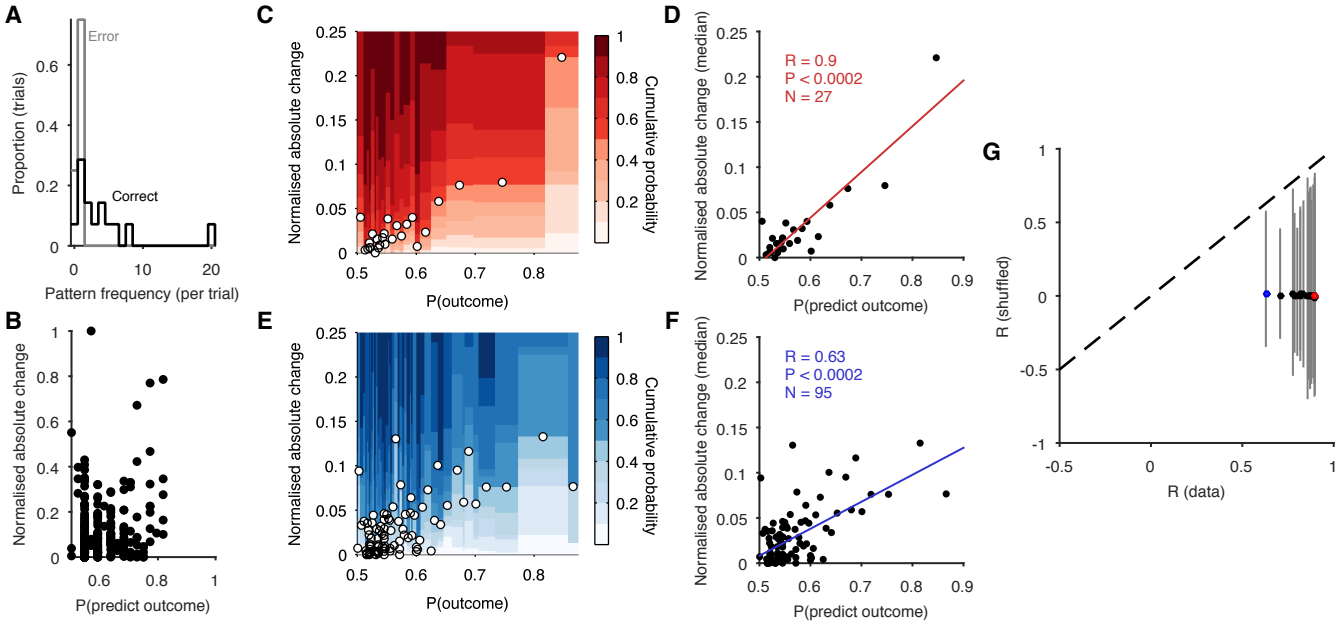


**Figure 5: Coding of trial outcome by sampled activity patterns**. (A) Example distributions of a pattern's frequency conditioned on trial outcome from one session. (B) For all co-activation patterns in one session, a scatter plot of outcome prediction and (absolute) change in pattern frequency between pre- and post-task SWS. Change is normalised to the maximum change in the session. (C) Distribution of change in pattern frequency according to outcome prediction over all ten sessions. Colour intensity gives the cumulative probability of at least that change. Circles give the median absolute change for each distribution. In this example, distributions were built using bins with 90 data-points each. Unbinned data are analysed in Figure 5 - figure supplement 1. (D) Correlation of outcome prediction and median change in pattern occurrence between sleep epochs from (C), over all ten sessions. Red line is the best-fit linear regression ($P < 0.0002$, permutation test). (E-F) As (C)-(D), for the worst-case correlation observed, with 25 data-points per bin. (G) Robustness of correlation results. Solid dots plot the correlation coefficient $R$ between outcome prediction and median change in pattern frequency obtained for different binnings of the data. Coloured dots correspond to panels C-D and E-F. Lines each give the entire range of $R$ obtained from a 5000-repeat permutation test; none reach the equivalent data point (dashed line shows equality), indicating all data correlations had $P < 0.0002$.

Remarkably we observed the convergence of distributions using precise activity patterns down to 2 ms resolution. Using surrogate models, we showed that the convergence is due to changes in correlations between neurons, rather than changes in firing rates. Previous work observed fine structure in stimulus-evoked population activity patterns in retina (e.g. Schneidman et al., 2006; Tkacik et al., 2014) and V1 (Berkes et al., 2011). We extend these results to show that such fine time-scale correlation structure can be observed in cortical regions for executive control, and be evoked by tasks. Unexpectedly, we have shown that, despite their high temporal resolution, task information can be decoded from these patterns.

How a cortical region encodes an internal model is an intriguing open question. A strong candidate is the relative strengths of the synaptic connections both into and within the encoding cortical circuit (Fiser et al., 2010; Buesing et al., 2011; Habenschuss et al., 2013; Kappel et al., 2015). The activity of a cortical circuit is strongly dependent on the pattern and strength of the connections between its neurons (e.g. Cossell et al., 2015; Okun et al., 2015). Consequently, defining the underlying model as the circuit's synaptic network allows both model-based inference through synaptically-driven activity and model

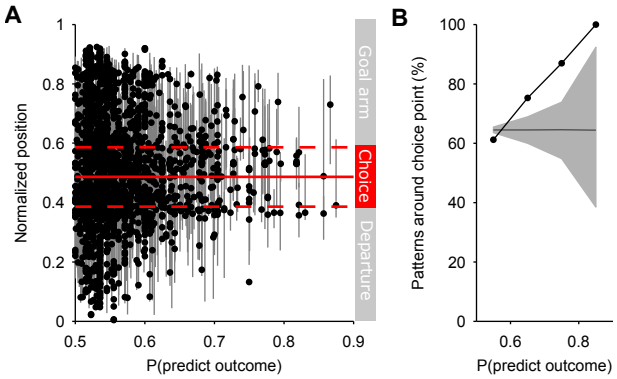


**Figure 6: Outcome predicting activity patterns are sampled in the choice area.** (A) Scatter plot of each pattern's outcome prediction and sample locations in the maze (dot is median position; grey line is interquartile range); all positions given as a proportion of the linearised maze from start of departure arm. Red lines indicate the approximate centre (solid) and boundaries (dashed) of the maze's choice area (cf Fig 1A). (B) Proportion of activity patterns whose interquartile range of sample locations enters the choice area (black dots and line). Grey region shows mean (line) and 95% range (shading) of proportions from a permutation test. The data exceed the upper limit of expected proportions for all outcome-predictive patterns.

learning through synaptic plasticity (Fiser et al., 2010).

Our results are distinct from previous observations of task-specific replay during sleep in prefrontal cortex (Euston et al., 2007), including reports (Peyrache et al., 2009) using the same data analysed here. In contrast to the work here, replay accounts do not consider the statistical distributions of the observed patterns, nor identify the changed patterns (beyond example templates in Euston et al., 2007), nor relate them to task behaviour; moreover, replay is described for coincident activity on coarse time-scales greater than those used here by a factor of 50 (Peyrache et al., 2009) up to a factor of 10000 (Euston et al., 2007). They thus do not address the statistical changes to population-wide activity predicted by theories of probabilistic population coding.

Our theory proposes that spontaneous neural activity during sleep is sampling a prior distribution generated from an internal model. We found that the set of activity patterns was remarkably conserved between sleeping and behaviour (Figure 1 - figure supplement 1), despite the different global dynamics of cortex between these states (Destexhe et al., 1999; Steriade et al., 2001), consistent with activity being generated from the same internal model in both states. This theory predicts that manipulating synaptic weights during sleep, changing the internal model, should change both the prior and the posterior distributions over task variables. Recent work has shown that inducing task-specific reward signals during sleep, likely altering synaptic weights, indeed immediately alters task behaviour on waking (de Lavilleon et al., 2015). Our results thus suggest that casting sleeping and waking activity as prior and posterior distributions generated from the same internal model could be a fruitful computational framework for relating cortical dynamics to behaviour.

# Materials and methods

**Task and electrophysiological recordings** The data analysed here were from ten recording sessions in the study of (Peyrache et al., 2009). For full details on training, spike-sorting, and histology see (Peyrache et al., 2009). Four Long-Evans male rats with

implanted tetrodes in prelimbic cortex were trained on the Y-maze task (Figure 1A). Each recording session consisted of a 20-30 minute sleep or rest epoch (pre-task epoch), in which the rat remained undisturbed in a padded flowerpot placed on the central platform of the maze, followed by a task epoch, in which the rat performed for 20-40 minutes, and then by a second 20-30 minute sleep or rest epoch (post-task epoch). Every trial started when the rat reached the departure arm and finished when the rat reached the end of one of the choice arms. Correct choice was rewarded with drops of flavoured milk. Each rat had to learn the current rule by trial-and-error, either: go to the left arm; go to the right arm; go to the lit arm. To maintain consistent context across all sessions, the extra-maze light cues were lit in a pseudo-random sequence across trials, whether they were relevant to the rule or not.

We analysed here data from the ten sessions in which the previously-defined learning criteria were met: the first trial of a block of at least three consecutive rewarded trials after which the performance until the end of the session was above 80%. In later sessions (not analysed here) the rats reached the criterion for changing the rule: ten consecutive correct trials or one error out of 12 trials. Thus each rat learnt at least two rules.

Tetrode recordings were spike-sorted only within each recording session for conservative identification of stable single units. In the ten sessions we analyse here, the populations ranged in size from 15-55 units.

**Activity pattern distributions** For a population of size $N$, we characterised population activity from time $t$ to $t + \delta$ as an $N$-length binary vector with each element being 1 if at least one spike was fired by that neuron in that time-bin, and 0 otherwise. In the main text we use a binsize of $\delta = 2$ ms throughout, and report the results of using larger binsizes in (Figure 2 - figure supplement 1) . We build patterns using the number of recorded neurons $N$, up to a maximum of 35 for computational tractability. The probability distribution for these activity patterns was compiled by counting the frequency of each pattern's occurrence and normalising by the total number of pattern occurrences.

**Comparing distributions** We quantified the distance $D(P|Q)$ between probability distributions $P$ and $Q$ using both the Kullback-Liebler divergence (KLD) and the Hellinger distance.

The KLD is an information theoretic measure to compare the similarity between two probability distributions. Let $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$ be two discrete probability distributions, for $n$ distinct possibilities – for us, these are all possible individual activity patterns. The KLD is then defined as $d(P|Q) = \sum_{i=1}^{n} p_i ln(\frac{p_i}{q_i})$. This measure is not symmetric, so that in general $d(P|Q) \neq d(Q|P)$. Following prior work (Berkes et al., 2011; Okun et al., 2012), we thus compute and report the symmetrised KLD: $D(P|Q) = (d(P|Q) + d(Q|P))/2$.

There are $2^N$ distinct possible activity patterns in a recording with $N$ neurons. Most of these activity patterns are never observed, so we exclude the activity patterns that are not observed in either of the epochs we compare. The empirical frequency of the remaining activity patterns is biased due to the limited length of the recordings (Panzeri et al., 2007). To counteract this bias, we use the Bayesian estimator and quadratic bias correction exactly as described in (Berkes et al., 2011). The Berkes estimator assumes a Dirichlet prior and multinomial likelihood to calculate the posterior estimate of the KLD; we use their code (github.com/pberkes/neuro-kl) to compute the estimator. We then compute a KLD estimate using all $S$ activity patterns, and using $S/2$ and $S/4$ patterns randomly sampled without replacement. By fitting a quadratic polynomial to these three

KLD estimates, we can then use the intercept term of the quadratic fit as an estimate of the KLD if we had access to recordings of infinite length (Strong et al., 1998; Panzeri et al., 2007).

The Hellinger distance for two discrete distributions $P$ and $Q$ is $D(P|Q) = \frac{1}{2}\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2$. To a first approximation, this measures for each pair of probabilities $(p_i, q_i)$ the distance between their square-roots. In this form, $D(P|Q) = 0$ means the distributions are identical, and $D(P|Q) = 1$ means the distributions are mutually singular: all positive probabilities in $P$ are zero in $Q$, and vice-versa. The Hellinger distance is a lower bound for the KLD: $2D(P|Q) \leq KLD$.

However we computed the distances between pairs of distributions, to compare those distances between sessions we computed a normalised measure of "convergence". The divergence between a given pair of distributions could depend on many factors that differ between sessions, including that each recorded population was a different size, and how much of the relevant population for encoding the internal model we recorded. Consequently, the key comparison between the divergences $D(Pre|R) - D(Post|R)$ also depends on these factors. To compare the difference in divergences across sessions, we computed a "convergence" score by normalising by the scale of the divergence in the pre-task SWS: $\left((D(Pre|R) - D(Post|R)\right)/D(Pre|R)$. We express this as a percentage. Convergence greater than 0% indicates that the distance between the task ($R$: correct trials) and post-task SWS ($Post$) distributions is smaller than that between the task and pre-task SWS ($Pre$) distributions.

**Statistics** Quoted measurement values are means ± s.e.m. All hypothesis tests used the non-parametric Wilcoxon signtest for a one-sample test that the sample median for the population of sessions is greater than zero. In all cases $N$=10 sessions. Throughout we plot mean values and their approximate 95% confidence intervals given by ± 2 s.e.m.

Bootstrapped confidence intervals (in Figure 2C,F) for each session were constructed using 1000 bootstraps of each epoch's activity pattern distribution. Each bootstrap was a sample-with-replacement of activity patterns from the data distribution $X$ to get a sample distribution $X^*$. For a given pair of bootstrapped distributions $X^*, Y^*$ we then compute their distance $D^*(X^*|Y^*)$. Given both bootstrapped distances $D^*(Pre|R)$ and $D^*(Post|R)$, we then compute the bootstrapped convergence $\left(D^*(Pre^*|R^*) - D^*(Post^*|R^*)\right)/D^*(Pre^*|R^*)$.

**Raster model** To control for the possibility that changes in activity pattern occurrence were due solely to changes in the firing rates of individual neurons and the total population, we used the raster model exactly as described in (Okun et al., 2012). For a given data-set of spike-trains $N$ and binsize $\delta$, the raster model constructs a synthetic set of spikes such that each synthetic spike-train has the same mean rate as its counterpart in the data, and the distribution of the total number of spikes per time-bin matches the data. In this way, it predicts the frequency of activity patterns that should occur given solely changes in individual and population rates.

For Fig 3 we generated 1000 raster models per session using the spike-trains from the post-task SWS in that session. For each generated raster model, we computed the distance between its distribution of activity patterns and the data distribution for correct trials in the task $D(Post - model|R)$. This comparison gives the expected distance between task and post-task SWS distributions due to firing rate changes alone. We plot the difference between the mean $D(Post - model|R)$ and the data $D(Post|R)$ in Figure 3.

**Outcome prediction** We examined the correlates of activity pattern occurrence with behaviour. To rule out pure firing rate effects, we excluded all patterns with $K = 0$ and $K = 1$ spikes, considering only co-activation patterns with two or more active neurons.

To check whether individual activity patterns coded for the outcome on each trial, we used standard receiver-operating characteristic (ROC) analysis. For each pattern, we computed the distribution of its occurrence frequencies separately for correct and error trials (as in the example of Figure 5A). We then used a threshold $T$ to classify trials as error or correct based on whether the frequency on that trial exceeded the threshold or not. We found the fraction of correctly classified correct trials (true positive rate) and the fraction of error trials incorrectly classified as correct trials (false positive rate). Plotting the false positive rates against the true positive rates for all values of $T$ gives the ROC curve. The area under the ROC curve gives the probability that a randomly chosen pattern frequency will be correctly classified as from a correct trial; we report this as $P$(predict outcome).

**Relationship of sampling change and outcome prediction** Within each session, we computed the change in each pattern's occurrence between pre- and post-task SWS. These were normalised by the maximum change within each session. Maximally changing patterns were candidates for those updated by learning during the task. Correlation between change in pattern sampling and outcome prediction was done on normalised changes pooled over all sessions. Change scores were binned using variable-width bins of $P$(predict outcome), each containing the same number of data-points to rule out power issues affecting the correlation. We regress $P$(predict outcome) against median change in each bin, using the mid-point of each bin as the value for $P$(predict outcome). Our main claim is that prediction and change are dependent variables (Figure 5C-G). To test this claim, we compared the data correlation against the null model of independent variables, by permuting the assignment of change scores to the activity patterns. For each permutation, we repeat the binning and regression. We permuted 5000 times to get the sampling distribution of the correlation coefficient $R^*$ predicted by the null model of independent variables. To check robustness, all analyses were repeated for a range of fixed number of data-points per bin between 20 and 100.

**Relationship of location and outcome prediction** The location of every occurrence of a co-activation pattern was expressed as a normalized position on the linearised maze (0: start of departure arm; 1: end of the chosen goal arm). Our main claim is that activity patterns strongly predictive of outcome occur predominantly around the choice point of the maze, and so prediction and overlap of the choice area are dependent variables (Figure 6B). To test this claim, we compared this relationship against the null model of independent variables, by permuting the assignment of location centre-of-mass (median and interquartile range) to the activity patterns. For each permutation, we compute the proportion of patterns whose interquartile range overlaps the choice area, and bin as per the data. We permuted 5000 times to get the sampling distribution of the proportions predicted by the null model of independent variables: we plot the mean and 95% range of this sampling distribution as the grey region in Figure 6B.

**Author Contributions** M.D.H and A.S. designed the analyses; A.S. and M.D.H. analysed the data; all authors discussed the results; M.D.H wrote the paper with contributions from A.S. and A.P.

**Acknowledgements** We thank the Humphries lab for discussions; Rasmus Petersen for comments on the manuscript; and P. Berkes and M. Okun for making their KL

divergence and raster model code publicly available. A.S. and M.D.H were supported by a Medical Research Council Senior non-Clinical Fellowship award to M.D.H. A.P. was supported by Human Frontier Science Program Fellowship LT000160/2011-l and National Institute of Health Award K99 NS086915-01.

# References

Baeg, E. H., Kim, Y. B., Huh, K., Mook-Jung, I., Kim, H. T. and Jung, M. W. (2003). Dynamics of population code for working memory in the prefrontal cortex. Neuron *40*, 177–188.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E. and Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. Neuron *60*, 1142–1152.

Benchenane, K., Peyrache, A., Khamassi, M., Tierney, P. L., Gioanni, Y., Battaglia, F. P. and Wiener, S. I. (2010). Coherent theta oscillations and reorganization of spike timing in the hippocampal- prefrontal network upon learning. Neuron *66*, 921–936.

Berkes, P., Orbán, G., Lengyel, M. and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science *331*, 83–87.

Buesing, L., Bill, J., Nessler, B. and Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. PLoS Comput Biol *7*, e1002211.

Cossell, L., Iacaruso, M. F., Muir, D. R., Houlton, R., Sader, E. N., Ko, H., Hofer, S. B. and Mrsic-Flogel, T. D. (2015). Functional organization of excitatory synaptic strength in primary visual cortex. Nature *518*, 399–403.

Dayan, P. and Abbot, L. F. (2001). Theoretical Neuroscience. MIT Press, Cambridge, MA.

de Lavilleon, G., Lacroix, M. M., Rondi-Reig, L. and Benchenane, K. (2015). Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. Nat Neurosci *18*, 493–495.

Destexhe, A., Contreras, D. and Steriade, M. (1999). Spatiotemporal analysis of local field potentials and unit discharges in cat cerebral cortex during natural wake and sleep states. J Neurosci *19*, 4595–4608.

Euston, D. R., Tatsuno, M. and McNaughton, B. L. (2007). Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. Science *318*, 1147–1150.

Fiser, J., Berkes, P., Orbán, G. and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci *14*, 119–130.

Fiser, J., Lengyel, M., Savin, C., Orbán, G. and Berkes, P. (2013). How (not) to assess the importance of correlations for the matching of spontaneous and evoked activity. , arXiv:1301.6554.

Fujisawa, S., Amarasingham, A., Harrison, M. T. and Buzsáki, G. (2008). Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. Nat Neurosci *11*, 823–833.

Ganmor, E., Segev, R. and Schneidman, E. (2015). A thesaurus for a neural population code. Elife *4*, e06134.

Habenschuss, S., Jonke, Z. and Maass, W. (2013). Stochastic computations in cortical microcircuit models. PLoS Comput Biol *9*, e1003311.

Ito, H. T., Zhang, S.-J., Witter, M. P., Moser, E. I. and Moser, M.-B. (2015). A prefrontal-thalamo-hippocampal circuit for goal-directed spatial navigation. Nature *522*, 50–55.

Kappel, D., Habenschuss, S., Legenstein, R. and Maass, W. (2015). Network Plasticity as Bayesian Inference. PLoS Comput Biol *11*, e1004485.

Kording, K. P. and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. Nature *427*, 244–247.

Luczak, A., Barth, P. and Harris, K. D. (2009). Spontaneous events outline the realm of possible sensory responses in neocortical populations. Neuron *62*, 413–425.

Ma, W. J., Beck, J. M., Latham, P. E. and Pouget1, A. (2006). Bayesian inference with probabilistic population codes. Nat Neurosci *9*, 1432–1438.

Miller, E. K. (2000). The prefrontal cortex and cognitive control. Nat Rev Neurosci *1*, 59–65.

Okun, M., Steinmetz, N. A., Cossell, L., Iacaruso, M. F., Ko, H., Barth, P., Moore, T., Hofer, S. B., Mrsic-Flogel, T. D., Carandini, M. and Harris, K. D. (2015). Diverse coupling of neurons to populations in sensory cortex. Nature *521*, 511–515.

Okun, M., Yger, P., Marguet, S. L., Gerard-Mercier, F., Benucci, A., Katzner, S., Busse, L., Carandini, M. and Harris, K. D. (2012). Population rate dynamics and multineuron firing patterns in sensory cortex. J Neurosci *32*, 17108–17119.

Panzeri, S., Senatore, R., Montemurro, M. A. and Petersen, R. S. (2007). Correcting for the sampling bias problem in spike train information measures. J Neurophysiol *98*, 1064–1072.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I. and Battaglia, F. P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nat Neurosci *12*, 916–926.

Pouget, A., Beck, J. M., Ma, W. J. and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. Nat Neurosci *16*, 1170–1178.

Ragozzino, M. E., Detrick, S. and Kesner, R. P. (1999). Involvement of the prelimbic-infralimbic areas of the rodent prefrontal cortex in behavioral flexibility for place and response learning. J Neurosci *19*, 4585–4594.

Rich, E. L. and Shapiro, M. L. (2007). Prelimbic/infralimbic inactivation impairs memory for multiple task switches, but not flexible selection of familiar tasks. J Neurosci *27*, 4747–4755.

Schneidman, E., Berry, M. J., Segev, R. and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. Nature *440*, 1007–1012.

Spellman, T., Rigotti, M., Ahmari, S. E., Fusi, S., Gogos, J. A. and Gordon, J. A. (2015). Hippocampal-prefrontal input supports spatial encoding in working memory. Nature *522*, 309–314.

Steriade, M., Timofeev, I. and Grenier, F. (2001). Natural waking and sleep states: a view from inside neocortical neurons. J Neurophysiol *85*, 1969–1985.

Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R. and Bialek, W. (1998). Entropy and Information in Neural Spike Trains. Phys Rev Lett *80*, 197–200.

Sul, J. H., Kim, H., Huh, N., Lee, D. and Jung, M. W. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron *66*, 449–460.

Tkacik, G., Marre, O., Amodei, D., Schneidman, E., Bialek, W. and Berry, 2nd, M. J. (2014). Searching for collective behavior in a large network of sensory neurons. PLoS Comput Biol *10*, e1003408.

Wohrer, A., Humphries, M. D. and Machens, C. (2013). Population-wide distributions of neural activity during perceptual decision-making. Prog Neurobiol *103*, 156–193.

Wolpert, D. M., Ghahramani, Z. and Jordan, M. I. (1995). An internal model for sensorimotor integration. Science *269*, 1880–1882.

Zemel, R. S., Dayan, P. and Pouget, A. (1998). Probabilistic interpretation of population codes. Neural Comput *10*, 403–430.

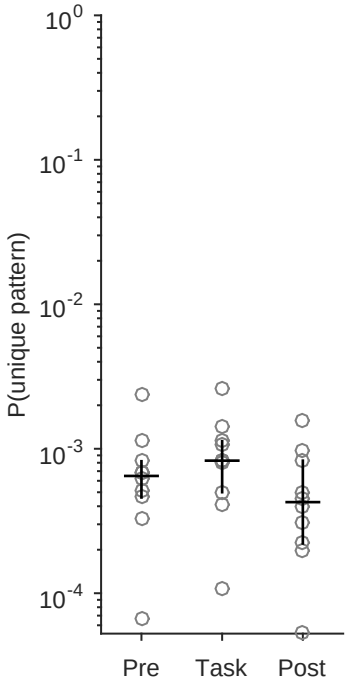


**Figure 1 - figure supplement 1**.

**Consistent sampling of activity patterns across session epochs.** Each circle is the proportion of activity patterns that appeared only in that epoch of the session. Black bar and line give the median and interquartile range across the 10 sessions. Note the log-scale, showing that the median proportion of unique patterns was less then 0.001 in all three epochs of the session.

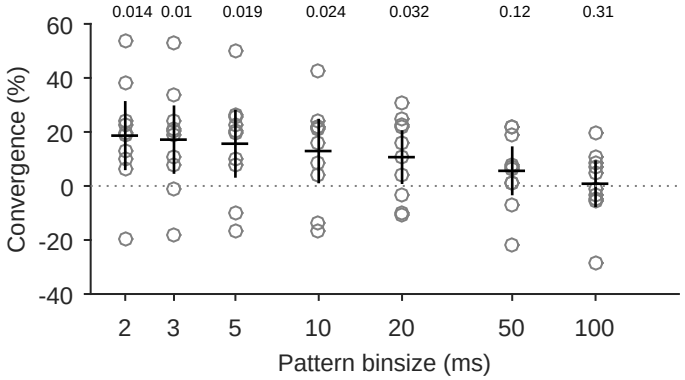


**Figure 2 - figure supplement 1**.

**Convergence of distributions over choice of pattern binsize.** We plot here the dependence of the convergence of task and post-task SWS distributions on the binsize used for constructing the activity patterns. We see that the convergence of the distribution of patterns on correct task trials is robust to an order of magnitude increase in binsize (the distribution at the binsize of 2ms is plotted in Fig. 2B). Above a binsize of 50 ms, convergence is statistically indistinguishable from zero, meaning that the pre- and post-task SWS distributions are equidistant, on average, from the task distribution. This suggests there is statistical structure in fine time-scale activity patterns that is not present on larger time-scales. Circles are convergence in individual sessions using Kullback-Liebler divergence; black lines give mean $\pm$ 2 s.e.m. Above each distribution is the P-value from a 1-tailed Wilcoxon signrank test, with $N$=10 sessions.

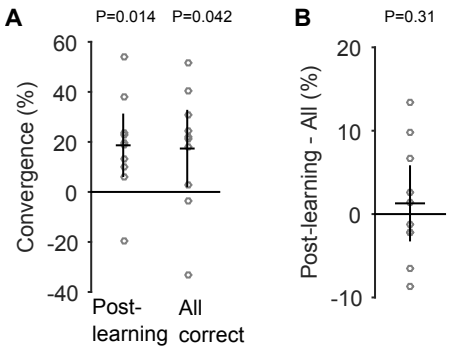


**Figure 2 - figure supplement 2**.

**Effect on convergence of including all correct trials.** In the main text, we construct all task-related distributions by considering only correct trials after the learning criterion trial (see Supplementary Note - Theory). We examine here whether our results were strongly contingent on that choice. (A) If we include all correct trials of a session, we find that convergence between task and post-task SWS distributions is still present ($P$-values from 1-tailed Wilcoxon signrank test, with $N$=10 sessions). Note that the convergence between task and post-task SWS distributions is if anything greater for post-learning trials, even though that task distribution is built from fewer samples, and so might be expected to be noisier. Black lines give mean $\pm$ 2 s.e.m. (B) Difference in convergence between using only post-learning or all correct trials for each session. ($P$-values for from 1-tailed Wilcoxon paired-sample ranksum test, with $N$=10 sessions). Black lines give mean $\pm$ 2 s.e.m.

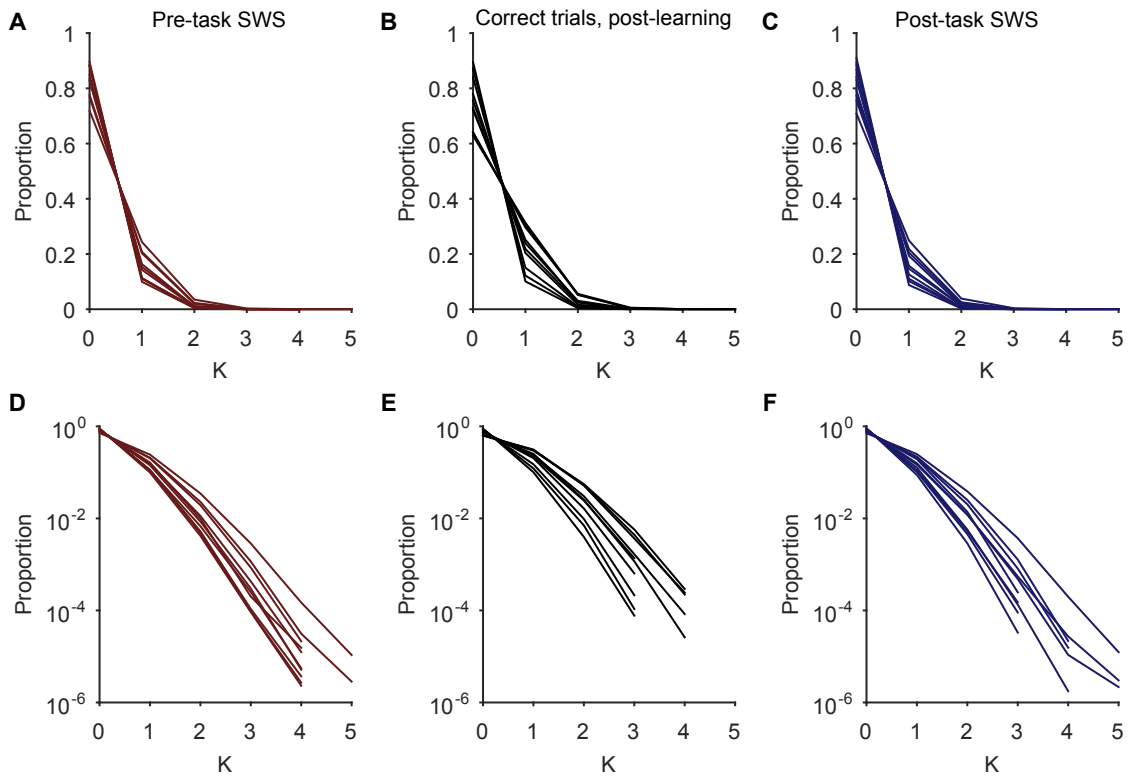


**Figure 3 - figure supplement 1.**
**Distributions of synchronous spiking in all activity patterns.** (A)-(C) Distributions of the number of unique recorded activity patterns containing exactly $K$ spikes, for pre-task SWS (A), correct task trials (B), and post-task SWS (C). Each line is the distribution for one session. (D)-(E) As (A)-(C), plotted on a log-scale to visualise the tails of the distributions. Co-activation patterns ($K \geq 2$ synchronous spikes) form a small proportion of all patterns.

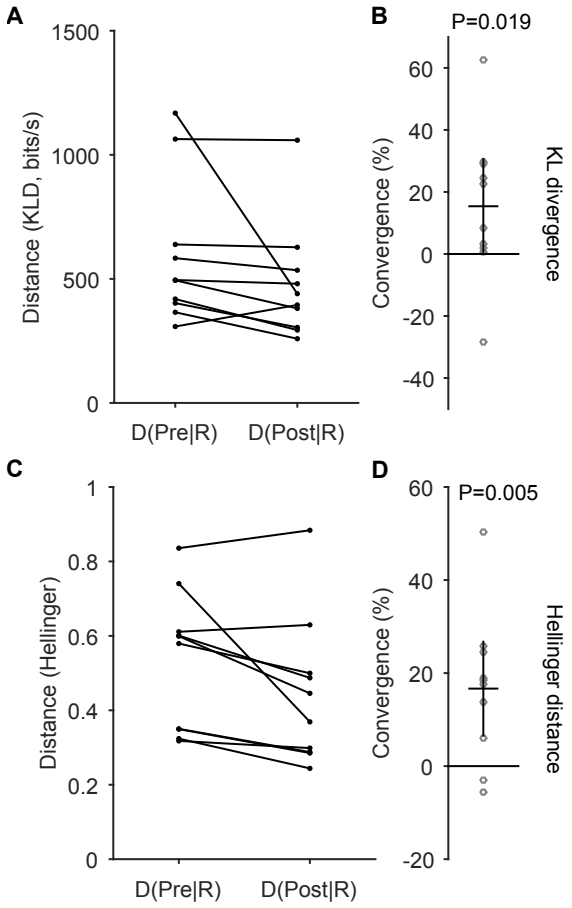


**Figure 3 - figure supplement 2**.

**Convergence between distributions of co-activation patterns**. Analysis of distributions restricted to patterns with two or more co-active neurons. (A) Distances between the distributions of pattern frequencies in sleep and task epochs; one dot per session. $D(X|Y)$: distance between pattern distributions in epochs $X$ and $Y$: Pre: pre-task SWS; Post: post-task SWS; R: correct task trials. (B) Scatter of convergence across all sessions (circles). Convergence greater than zero means that the activity pattern distribution in the task is closer to the distribution in post-task SWS than the distribution in pre-task SWS. Black lines give mean $\pm$ 2 s.e.m. (C) - (D) As (A)-(B), using Hellinger distance. All $P$-values from 1-tailed Wilcoxon signrank test, with $N$=10 sessions.

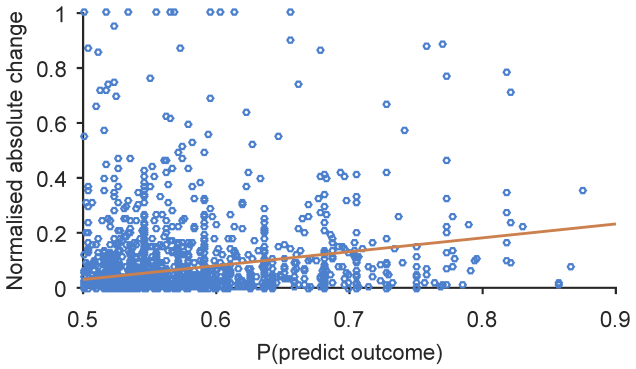


**Figure 5 - figure supplement 1**.
**Joint distribution of outcome prediction and change in sampling**. Here we plot every co-activation pattern's joint values of $P(outcome)$ and the absolute normalised change in sampling between pre- and post-task slow-wave sleep ($N = 2353$ patterns with $K \geq 2$ spikes per pattern across all 10 sessions). The linear regression in red indicates a clear relationship between the two ($R = 0.22$, $P < 10^{-27}$). Nonetheless, the majority of patterns do not markedly change their sampling, nor are they predictive of outcome: 72% (1699/2353) have $P(outcome) \leq 0.6$ and a change of less than 10%. Thus fitting a linear regression is not robust, as it is dominated by fitting to this majority that do not change. Rather, it is clear that there is a distribution of change for each $P(outcome)$, which we analyse in the main text.

# Supplementary File 1 - Theory

Abhinav Singh, Adrien Peyrache and Mark D. Humphries

## Neural inference

How do we know the current state of the world given some input from it? Our input is both limited in time and noisy, so our estimates are inherently uncertain. Consequently, we have an inference problem: what is our best guess of the current state of the world given some finite, noisy input? We can state this problem as being equivalent to inferring the probability distribution

$$P(state|input, model) \tag{1}$$

at some given moment in time $t$; in words, this is the probability of currently being in a given state, out of all possible states, given both the available input and some internal model of the world. Using Bayes' theorem, we can make this dependence on input and model explicit:

$$\underbrace{P(state|input, model)}_{\text{posterior}} \propto \underbrace{P(input|state, model)}_{\text{likelihood}} \underbrace{P(state|model)}_{\text{prior}} \tag{2}$$

The prior is the internal estimate of the current state before the observation *input*, the posterior is the estimate of the current state after observing *input*, and the improvement in the estimate arises from the new information available in *input* that is processed through the likelihood. All these are dependent on the model of the world we are using. This internal model specifies how we interpret the inputs in the likelihood, and generate the prior probabilities. If we change the model, we change these two operations, and so change our estimates of the current state of the world. We can think of the model as specifying what we expect to be relevant in the input, and what states we expect to be in.

One goal of learning is thus to update the internal model to match the statistical properties of the world. The better the model, the better we will be able to predict the state of the external world. But as we can only access directly the inputs generated from those states, formally we say that learning seeks to maximise $P(input|model)$ over all possible inputs at all times $t$ by changing the parameters of the model. A model which always generates maximum values for $P(input|model)$ is the best possible learnt internal model of the external world. Obtaining such a model necessarily means that we have experienced all possible states giving rise to those inputs, so that the prior $P(state|model)$ is always accurate, and we obtain no new information from the likelihood. Consequently, the posterior probability becomes always proportional to the prior probability. A measure of learning is thus how close the prior and posterior distributions have become.

## Inference-by-sampling

The inference-by-sampling theory (Fiser et al., 2010; Berkes et al., 2011) proposes that the model is encoded by the particular set and weight of connections in a neural circuit. In this view, the posterior distribution is encoded by the activity of the circuit evoked by some input. Crucially, it predicts that the prior distribution is encoded by spontaneous activity of the same circuit, as this is solely sampling the model.

If the circuit is the model, then the theory predicts that the circuit's instantaneous population activity is a sample from a probability distribution - from the posterior when

receiving external input, from the prior in spontaneous activity. Some downstream neurons, receiving these samples as a consecutive sequence of inputs, can reconstruct the probability distribution just by summing their inputs over time.

For simplicity, Berkes et al. (2011) consider the instantaneous population activity as some binary vector indicating whether each neuron was active or inactive in a very small time window. This representation makes the distributions easy to measure experimentally.

Learning updates synaptic weights, altering the encoded model. The prediction that posterior and prior distributions converge over learning is thus neurally equivalent to the convergence between the distributions of evoked and spontaneous population activity.

## Evidence for inference-by-sampling in visual cortices

These ideas were developed in the context of visual processing, and particularly with reference to V1. In this context, the "state" of the world is the current view, and the input is the information received by the retina. The proposed purpose of inference in V1 is to infer the most likely low level visual features – edges, for example – present in the current view, given the input to the retina. V1's internal model is then a statistical model of the low-level features, which can be built over a life-time's experience of the world.

Consequently, Berkes and colleagues (Berkes et al., 2011) tested the construction of this internal model by recording from area V1 at different stages of development in the ferret. Natural images were used to probe the current posterior distribution supported by the model, and darkness was used to probe the current prior distribution. Over development, the activity distribution evoked by natural images increased its similarity to the distribution during darkness. This increase was robust to a series of controls for simultaneous changes in firing rate statistics (Berkes et al., 2011; Okun et al., 2012; Fiser et al., 2013). Their results are consistent with the inference-by-sampling interpretation in which the internal model is updated by experience with the world, so that the posterior and prior distributions converge.

## Inference-by-sampling in higher cortices over learning during behaviour

These results could not address learning separately from development. Further, unknown is whether inference-by-sampling can be observed in higher-order cortices, or during ongoing behaviour.

There is no *a priori* reason to expect that inference-by-sampling would be restricted to primary sensory cortices. Much has been written about the generic nature of the cortical microcircuit (Thomson and Lamy, 2007; Harris and Shepherd, 2015), so we might reasonably expect that, if an internal model is encoded by the neural circuit in V1, so other similar cortical circuits in other regions encode other internal models.

Compelling support for this has come from modelling work by Maass and colleagues (Buesing et al., 2011; Habenschuss et al., 2013). Their models have shown how a wide range of plausible cortical circuit models all produce the necessary dynamics to sample from a statistical model encoded by the circuit's connections (Buesing et al., 2011; Habenschuss et al., 2013). Moreover, the models also replicate key properties of the firing statistics in cortex, including the close-to-Poisson irregularity of firing patterns. These suggest that the inference-by-sampling hypothesis is indeed a plausible generic computation for cortex.

Inference of state is also a generic operation. Nothing in Equation 2 limits its application to sensory information. We might consider "state" in the sense used in the reinforcement learning literature (Sutton and Barto, 1998), as a generic description of the current values of variables of the external world. Indeed, in forms of reinforcement learning that depend on simulation of future actions, "state" in this context can even refer to the simulated values of variables in the external world - for which we would use the internal model to simulate possible outcomes. During behaviour, we might thus expect that an internal model is learnt about the statistical dependence of outcomes on decisions in particular contexts.

The power of the inference-by-sampling hypothesis is that we do not need to know the internal model to test for its existence. We need not specify an exact model to test the convergence of distributions in evoked and spontaneous activity, but such a convergence is evidence of an updated internal model.

Consequently, to test the generality of the inference-by-sampling hypothesis, we sought to test the convergence of distributions over learning using data from the medial prefrontal cortex (mPfC) of rats learning rules in a Y-maze task (Peyrache et al., 2009). By looking at these data for a change to some internal model in mPFC, we are assuming only that the model is related to the rule, but not any specific form of model. It could encode the set of task states and their transitions; it could encode the current sequence of required actions; it could be a statistical model of outcomes. Supporting this assumption, we know mPFC is necessary for successful acquisition of new rules (Ragozzino et al., 1999; Rich and Shapiro, 2007), and that mPFC pyramidal neurons change their firing patterns during acquisition of the rules used here (Benchenane et al., 2010).

Even if the interpretation of the convergence of distributions in the inference-by-sampling framework turns out to be incorrect, the observation of such a convergence between waking and spontaneous activity over learning still offers compelling clues to the nature of cortical computation.

## What distributions to compare?

Nonetheless, the inference-by-sampling theory places limits on exactly which activity distributions to compare. In the Berkes et al. (2011) study, this decision was made simple by the elegant experimental design. As they monitored V1 over development, so it was reasonable to expect the internal model to adapt to the statistics of the world over a lifetime. Their tests at different developmental stages were samples of the current posterior and prior distributions supported by the model. We would not expect significant changes to the internal model during their testing, as it was short on the time-scale of the developmental changes, and so they could compare their entire recorded distributions of evoked and spontaneous activity. In other words, they were able to compare two distributions from the same, static model.

Our data on rats learning rules in a Y-maze allow us to address if learning of the internal model can be observed. But learning on short time-scales brings the confounding issue that learning the model is happening online, while we are monitoring activity. So what distributions should we compare?

We chose the 10 training sessions in which the rat clearly acquired the present rule, so we could be reasonably sure that we would observe changes that correlated with learning. We reasoned that neural activity in clearly identified sleep periods before and after the session was a clear candidate for spontaneous activity, as it occurred in the absence of external sensory input. We used slow-wave sleep periods to clearly delineate the presence

of sleep. As the rats acquired the rule in that session then, if mPfC indeed encodes rule acquisition, we expect that the spontaneous activity in sleep after the session is drawn from the internal model related to the correct rule.

We can only be sure that during behaviour this correct-rule model would be sampled on correct trials. This does not imply that mPfC activity is causal for decisions on those trials - even in a monitoring or goal-encoding role, mPfC activity would reflect whether or not the correct decision was made. The mPfC activity on error trials is unconstrained by the theory. Consequently, we can only be sure that, if the inference-by-sampling hypothesis is true, then the distribution of samples on correct trials would converge, on average, to the distribution in sleep after learning.

The final, subtle constraint is that overt behavioural signs of learning likely indicates ongoing synaptic plasticity. For example, on the same Y-maze, some pyramidal neurons in mPfC change the timing of their spikes in relation to the hippocampal theta rhythm, indicating local circuit plasticity (Benchenane et al., 2010). If so, then the internal model is changing during behaviour. But the internal model putatively sampled in the post-session sleep will be stable. To thus minimise the confound of these changes during behaviour, and compare static posterior and prior distributions (as per Berkes et al., 2011), we sought to identify where the internal model updating may have finished. A useful proxy for this is the asymptotic behavioural performance. We thus used the trial at which the rat reached the learning criteria as the indicator of relative stability in the internal model. All correct trials from this trial onwards were then used to construct the activity distribution during the task - we call this distribution $P(R)$ in the main text, and distances measured between it and some other distribution $P(X)$ we call $D(X|R)$.

## References

Benchenane, K., Peyrache, A., Khamassi, M., Tierney, P. L., Gioanni, Y., Battaglia, F. P. and Wiener, S. I. (2010). Coherent theta oscillations and reorganization of spike timing in the hippocampal- prefrontal network upon learning. Neuron *66*, 921–936.

Berkes, P., Orbán, G., Lengyel, M. and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science *331*, 83–87.

Buesing, L., Bill, J., Nessler, B. and Maass, W. (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. PLoS Comput Biol *7*, e1002211.

Fiser, J., Berkes, P., Orbán, G. and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci *14*, 119–130.

Fiser, J., Lengyel, M., Savin, C., Orbán, G. and Berkes, P. (2013). How (not) to assess the importance of correlations for the matching of spontaneous and evoked activity. , arXiv:1301.6554.

Habenschuss, S., Jonke, Z. and Maass, W. (2013). Stochastic computations in cortical microcircuit models. PLoS Comput Biol *9*, e1003311.

Harris, K. D. and Shepherd, G. M. G. (2015). The neocortical circuit: themes and variations. Nat Neurosci *18*, 170–181.

Okun, M., Yger, P., Marguet, S. L., Gerard-Mercier, F., Benucci, A., Katzner, S., Busse, L., Carandini, M. and Harris, K. D. (2012). Population rate dynamics and multineuron firing patterns in sensory cortex. J Neurosci *32*, 17108–17119.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I. and Battaglia, F. P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nat Neurosci *12*, 916–926.

Ragozzino, M. E., Detrick, S. and Kesner, R. P. (1999). Involvement of the prelimbic-infralimbic areas of the rodent prefrontal cortex in behavioral flexibility for place and response learning. J Neurosci *19*, 4585–4594.

Rich, E. L. and Shapiro, M. L. (2007). Prelimbic/infralimbic inactivation impairs memory for multiple task switches, but not flexible selection of familiar tasks. J Neurosci *27*, 4747–4755.

Sutton, R. S. and Barto, A. G. (1998). Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.

Thomson, A. M. and Lamy, C. (2007). Functional maps of neocortical local circuitry. Front Neurosci *1*, 19–42.