Nature Neuroscience Article Submission

Metastable attractors explain the variable timing of stable behavioral action sequences

Stefano Recanatesi[1,a,b] **Ulises Pereira**[1,c,d] **Masayoshi Murakami**[e,f] **Zachary Mainen**[2,f] **Luca Mazzucato**[2,b,g]

[a] *University of Washington, Center for Computational Neuroscience and Swartz Center, Seattle*
[b] *Institute of Neuroscience, University of Oregon, Eugene.*
[c] *Department of Statistics, The University of Chicago, Chicago*
[d] *Center for Neural Science, New York University, New York*
[e] *Department of Neurophysiology, University of Yamanashi, Japan.*
[f] *Champalimaud Centre for the Unknown, Lisbon, Portugal.*
[g] *Departments of Biology and Mathematics, University of Oregon, Eugene.*

E-mail: `zmainen at neuro dot fchampalimaud dot org; lmazzuca at uoregon dot edu`

Abstract: Natural animal behavior displays rich lexical and temporal dynamics, even in a stable environment. The timing of self-initiated actions shows large variability even when they are executed in reliable, well-learned sequences. To elucidate the neural mechanism underlying this mix of reliability and stochasticity, we trained rats to perform a stereotyped sequence of self-initiated actions and recorded neural ensemble activity in secondary motor cortex (M2), known to reflect trial-by-trial action timing fluctuations. Using hidden Markov models, we established a dictionary between ensemble activity patterns and actions. We then showed that metastable attractors, with a reliable sequential structure yet high transition timing variability, could be produced by coupling a high-dimensional recurrent network and a low-dimensional feedforward one. Transitions between attractors in our model were generated by correlated variability arising from the feedback loop between the two networks. This mechanism predicted aligned, low-dimensional noise correlations that were empirically verified in M2 ensembles. Our work establishes a novel framework for investigating the circuit origins of self-initiated behavior based on correlated variability.

[1] Co-first authors.
[2] Co-corresponding authors.

Contents

1 Introduction

When interacting with a complex environment, animals generate naturalistic behavior in the form of self-initiated action sequences, originating from the interplay between external cues and the internal dynamics of the animal. Self-initiated behavior exhibits variability both in its temporal dimension (when to act) and in its spatial features (which actions to choose, in which order) [1–3]. Large trial-to-trial variability has been observed in action timing, where transitions between consecutive actions are well described by a Poisson process [4]. Recent studies in C. Elegans [5], Drosophila [1] and rodents [2, 3] demonstrated that the spatiotemporal dynamics of self-initiated action sequences can be captured by state space models, based on an underlying Markov process. These analyses revealed a repertoire of behavioral motifs typically numbering in the hundreds, leading to a combinatorial explosion in the number of action sequences. Such a large behavioral landscape poses a formidable challenge for investigating the neural underpinnings of behavioral variability. A promising approach to tame the curse of dimensionality is to reduce the lexical variability in the behavioral repertoire, by using a task where the set of actions is rewarded when executed in a fixed order, yet retaining variability in action timing [6, 7], a hallmark of self-initiated behavior [4].

Previous studies in rodents have identified the secondary motor cortex (M2) as part of a distributed network involved in motor planning, working memory [8] and self-initiated tasks [6, 7]. During delay periods in decision-making tasks, trial-averaged population activity in M2 displays clear features of attractor dynamics, with two discrete attractors encoding the animal's upcoming choice [9]. Are attractor dynamics in M2 specific to delay period activity? Here, we investigate the alternative hypothesis that attractor dynamics represent the intrinsic operational regime of M2 neural circuits throughout sequences of self-initiated behavior. In particular, we sought to uncover a correspondence between M2 attractors and upcoming self-initiated actions.

Because self-initiated action sequences are characterized by large trial-to-trial temporal variability in transition timing, they cannot be directly aligned across trials, hampering the applicability of traditional trial-averaged measures of neural activity. A principled framework to tackle this issue is to model neural population dynamics using hidden Markov models (HMMs) [10]. These state space models can identify hidden states from population activity patterns in single trials, and have been successfully deployed in a variety of tasks and species from C. Elegans [5] to rodents [11–14], primates [15–18] and humans [19, 20]. Hidden Markov models segment single-trial population activity into sequences in an unsupervised manner by inferring hidden states from multi-neuron firing patterns. Within each pattern, neurons fire at an approximately constant firing rate for intervals typically lasting hundreds of milliseconds.

Previous work showed that the activity patterns, revealed by hidden Markov models, can be interpreted as metastable attractors, arising from recurrent dynamics in local cortical circuits [12, 21]. Metastable attractors are produced in biologically plausible network models [22, 23] and have been used to elucidate features of sensory processing [12, 21], working memory [24] and expectation [25], and to explain state-dependent modulations of neural variability [22, 23, 26]. However, while previous models are capable of generating sequential activity [21, 27–29], they are hindered by a fundamental trade-off between sequence reproducibility and trial-to-trial temporal variability. Namely, they can endogenously generate either reliable sequences without temporal variability [27–29] or, instead, sequences with large temporal variability but unreliable order [12, 23]. Thus, existing models are incapable of generating reproducible sequences of metastable attractors, characterized by large trial-to-trial variability in attractor dwell times.

Here, we addressed these issues in a waiting task [6, 7] in which freely moving rats performed many repetitions of a sequence of self-initiated actions leading to a water reward. The identity and order of actions in the sequence was fixed by the task reward contingencies (i.e., producing out-of-sequence actions yielded no rewards), yet action timing retained large trial-to-trial variability [6, 7]. We found that M2 population activity during the task could be well modeled by an HMM that established a dictionary between self-initiated actions and neural patterns. To explain the neural mechanism generating reproducible yet temporally variable sequences of patterns, we propose that transitions between attractors are driven by low-dimensional correlated variability. This can be produced by reciprocally connecting a high dimensional recurrent network and a low-dimensional feedforward network. Attractors in the high-dimensional network represent the neural patterns inferred from M2 population activity. Previous experiments showed that recurrent circuits between cortical areas like M2 and subcortical areas such as thalamus [30, 31] and basal ganglia nuclei [32–34] are necessary to sustain attractor dynamics and produce motor sequences, and we suggest that cortical-subcortical circuits might correspond to our high- and low-dimensional network interaction. This mechanistic model predicts low-dimensional and sequentially aligned noise correlations, which we confirmed in the empirical data. While previous work showed that low-dimensional (differential) correlations may be detrimental for accurately encoding external stimuli [35], our results demonstrate that, surprisingly, they are also essential for circuits to produce stable yet temporally variable self-initiated action sequences.

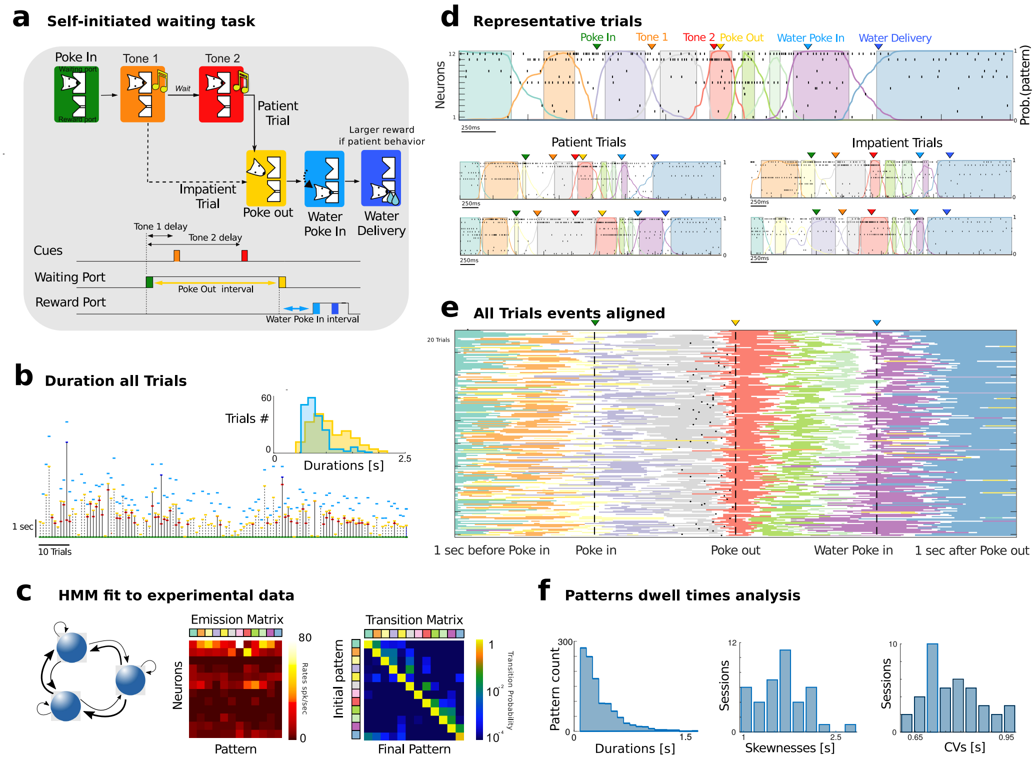

Figure 1. Waiting task and M2 pattern sequences. a) Schematic of task events. A rat self-initiated the waiting task by poking into a Wait port (Poke In), where tone 1 was played (after 400 ms), and, after a variable delay, a different tone 2 was played. The animal could decide to Poke Out of the Wait port at any time (after tone 2 in patient trials; between tone 1 and 2 in impatient trials) and move to the Reward port (Water Poke In) to receive a water reward (large and small for patient and impatient trials, respectively). Bottom: schedule of trial events. Three events (PI, PO, WPI) are triggered by self-initiated actions with respective interevents interval highlighted. b) Waiting behavior in a representative session. Vertical bars indicate waiting times for patient or impatient trials (full and dashed lines, respectively). Tick marks represent event times (color-coded as in a). Inset: Interevent interval distribution for self-initiated actions ([PO - PI] and [WPI - PO], yellow and cyan, respectively). c) Neural pattern inference via Hidden Markov Model (HMM). An HMM (left, schematics) is fit to a representative session in d, returning a set of neural patterns (Emission Matrix, center) and a Transition Probability Matrix (TPM, right). Each pattern is a population firing rate vector (columns in the Emission Matrix). The TPM returns the probability for a transition between two patterns to occur. d) Representative trials from one ensemble of 12 simultaneously recorded M2 neurons during patient (top and bottom left) and impatient (bottom right) trials. Top: spike rasters with latent patterns extracted via HMM (colored curves represent pattern posterior probability; colored areas indicate intervals where a pattern was detected with probability exceeding 80%). e) All trials from the representative session (each row corresponds to a trial). Individual trials have been time-stretched to align to five different events (1 s before Poke In, Poke In, Poke Out, Water Poke In, 1 s after Poke Out). All trials display a stereotyped pattern sequence. Color-coded lines represent stretched intervals where patterns were detected (same as colored intervals in d). Black tick marks represent tone 2 onset in patient trials only. f) Left: Histograms of pattern dwell times across trials in the representative session reveal right-skewed distributions (we excluded the first and last pattern in the sequence, whose duration artificially depends on trial interval segmentation). Skewness and coefficient of variability (CV) of pattern dwell time distributions reveal large trial-to-trial variability (41 sessions).

2 Results

2.1 Ensemble activity in M2 unfolds through reliable pattern sequences

To elucidate the circuit mechanism underlying self-initiated actions we trained animals on a waiting task. In the waiting task, freely moving rats were trained to perform a sequence of self-initiated actions to obtain a water reward. Animals engaged in the trial by inserting their snout into a wait port, where, after a 400 ms delay, a first auditory tone signaled the beginning of the waiting epoch. Two alternative options were made available: i) waiting for a second tone, delivered at random times, then move to the Reward port to collect a large water amount (henceforth referred to as “patient" trials); or ii) terminating the trial at any moment before the second tone, then move to a reward port to collect a small amount of water (henceforth referred to as “impatient" trials). In either case, rewards were collected by withdrawing the snout from the wait port and poking into the reward port; thus, patient and impatient trials shared the same action sequence (Fig. 1a). The intervals between consecutive actions show large trial-to-trial variability with right-skewed distributions (Fig. 1b and Fig. S1a), suggestive of a potential stochastic mechanism underlying their action timing [4].

To uncover the neural correlates of self-initiated actions, we recorded ensemble spike trains from the secondary motor cortex (M2, from $N = 5-20$ neurons per session, 9.1 ± 0.5 on average across 41 recorded sessions) of rats engaged in the waiting task [6, 7]. We found that single-trial ensemble neural activity in M2 consistently unfolded through reliable sequences of hidden or latent neural patterns, inferred via a Poisson hidden Markov model (HMM, see Fig. 1c and Fig. S2). This latent variable model posits that ensemble activity in a given time bin is determined (and emitted) by one of a few unobservable latent activity patterns, represented by a vector of ensemble firing rates (depicted column-wise in the “emission matrix”). In the next time bin, the ensemble may either dwell in the current pattern or transition to a different pattern, with probabilities given by rows of the “transition matrix.” Stochastic transitions between patterns occur at random times according to an underlying Markov chain, and neurons discharge as Poisson processes with pattern-dependent firing rates. The number of patterns in each session was selected via an unsupervised cross-validation procedure (10.3 ± 4.1 across 41 sessions, which ranged from 4 to 21 patterns, Fig. S2; see Methods). The identity and order of inferred patterns were remarkably consistent within each session even across patient and impatient trials, (Fig. 1e and Fig. S3). The average pattern dwell time was 0.50 ± 0.28 s (Fig. 1f), in agreement with previous findings in other cortical areas [12]. Such long dwell times, which are greater than typical single neuron time constants, suggest that the observed patterns may be an emergent property of the collective circuit dynamics within M2 and reciprocally connected brain regions. Crucially, even though the identity and order of patterns within a sequence were highly consistent across trials, pattern dwell times showed large-trial to trial variability, characterized by right-skewed distributions (Fig. 1f, coefficient of variability CV=0.78 ± 0.10 and skewness 1.67 ± 0.46). This temporal heterogeneity suggests that a stochastic mechanism may contribute to driving transitions between consecutive patterns within a sequence.

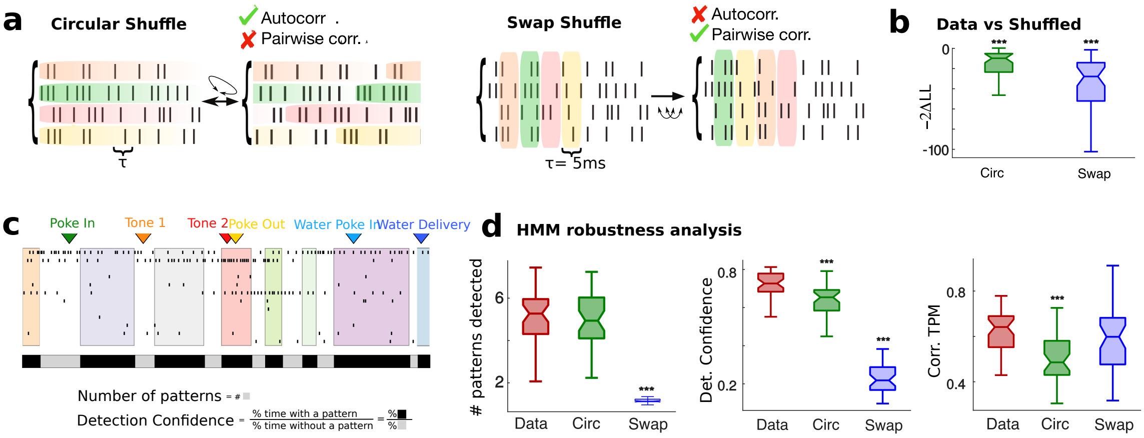

Figure 2. Robustness of pattern inference. a) Schematic of shuffled procedure to create surrogate datasets: Circular Shuffle (left) preserved single-cell autocorrelations and destroyed pairwise correlations; Swap Shuffle (right) preserved pairwise correlations and destroyed autocorrelations. b) Difference in $-2\times$Log-likelihood between HMM fit of empirical dataset and of the two surrogate datasets. c) Representative trial showing detection confidence measure (same color-coded notation as in Fig. 1; black and grey bars: fraction of trial duration where patterns were detected with probability larger or smaller than 80%, respectively. d) HMM robustness analyses. Left: Number of patterns detected across sessions for empirical and surrogate datasets. Center: Pattern detection confidence, estimated as fraction of trials were patterns were detected with probability exceeding 80%. Right: Consistence of pattern sequence, estimated as Pearson correlation coefficients between single-trial estimates of "symbolic" TPMs encoding the sequence identity (see Methods section 4.4). b-d: signed-rank tests between empirical and shuffled datasets, $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$.

2.2 Pattern sequences capture dynamics beyond auto- and pairwise correlations

Before investigate this mechanism (section 2.5, below), we show a number of analyses directed at quantifying how well the HMM-inferred pattern sequences captured ensemble spiking activity beyond single-cell autocorrelations and pair-wise correlations. To do so, first we performed a cross-validation analysis comparing the data to two surrogate datasets (Fig. 2a) [13]. In the "circular-shuffled" surrogate dataset, we circularly shifted bins for each neuron within a trial (i.e., row-wise), thus destroying pairwise correlations but preserving single-cell autocorrelations. In the "swap-shuffled" surrogate dataset, we randomly permuted population activity across bins within a trial (i.e. column–wise), thus preserving instantaneous pairwise correlations but destroying autocorrelations. We found that the cross-validated likelihood of held-out trials for an HMM trained on the real dataset was significantly larger compared to an HMM trained on surrogate datasets (Fig. 2b, empirical vs. circular shuffled: $p = 9 \times 10^{-8}$; vs. swap shuffled: $p = 0.051$, signed-rank test). When we destroyed autocorrelations, the model entirely failed to detect pattern transitions, leaving only one pattern (Figs. 2c to 2d, $p = 2.4 \times 10^{-8}$). When destroying pairwise correlations, the model still detected multiple patterns whose number was in the same range as the model trained on the empirical data (Fig. 2d, $p = 0.11$). However, pattern detection was significantly less confident than in empirical data (Figs. 2c and 2d, $p = 1.4 \times 10^{-7}$); moreover, inferred pattern sequences were significantly sparser and more similar across trials in the data compared to the surrogate datasets (Fig. 2d, $p = 9 \times 10^{-8}$, Fig. S2c-Fig. S2d). We concluded that pattern sequences captured the single-trial dynamics of population activity beyond autocorrelations and pairwise correlations.

2.3 Patterns arise from dense and distributed neural representations

How do pattern sequences emerge from neural activity? Patterns formed separate clusters tiling population activity space, with between-cluster distances being significantly larger than within-cluster distances (Fig. 3a, Wilcoxon rank-sum test, $p < 10^{-20}$). Most neurons were active in several patterns, leading to dense neural representations, where overlaps between patterns (0.41±0.22, defined as Pearson correlation between firing rate vectors) were significantly larger than expected solely on the basis of the underlying firing rate distribution (Fig. 3b, $p < 3 \times 10^{-18}$, t-test). We found that the vast majority of neurons (88±2%) were "multistable", with their firing rates significantly modulated across patterns (Fig. 3c), in agreement with previous findings [12]. In particular, neurons attained on average 3.2±1.2 different firing rates across patterns. We concluded that most M2 neurons participated in the pattern sequences, suggesting that M2 neural populations can support dense and distributed representations characterized by mixed selectivity to multiple patterns.

2.4 Pattern onsets predict self-initiated actions

What kind of information about self-initiated behavior can be decoded from M2 pattern sequences? The statistical structures of neural patterns and action sequences shared remarkable similarities: single-trial consistency of identity and order of actions/patterns within a session, yet right-skewed distributions of timing intervals across trials (Fig. 1 and Fig. S3). We thus hypothesized that the onset of specific neural patterns could be causally involved with and therefore predictive of the timing and identity of upcoming self-initiated actions.

To test this hypothesis, we aimed to establish a cross-validated dictionary between actions and neural patterns, which we did by tagging the onset of specific patterns with the actions they most strongly predicted (Fig. 4a). This automated tagging method showed that, even though both pattern onsets and actions occurred at highly variable times in different trials, action onset times were reliably preceded by specific patterns onset on a sub-second scale (168 ± 175 ms, interval between pattern onset and tagged action). To assess the significance of the action/pattern dictionary, we tested whether pattern onsets could correctly predict actions performed during incorrect trials (32.5 ± 15.8%, see Fig. 4b and Fig. S4), based on the dictionary learned solely from correct trials (defined as those in which a visit to the waiting port was following directly by a movement to the reward port; both patient and impatient types, 67.4 ± 15.8% fraction of trials per session, see Fig. 1a); other action sequences, where the animal behaved erratically, were deemed as incorrect trials (Fig. S4). When using the cross-validated action/pattern dictionary learned on correct trials, we were able to correctly predict which actions the animal would perform in incorrect trials (Figs. 4b to 4c).

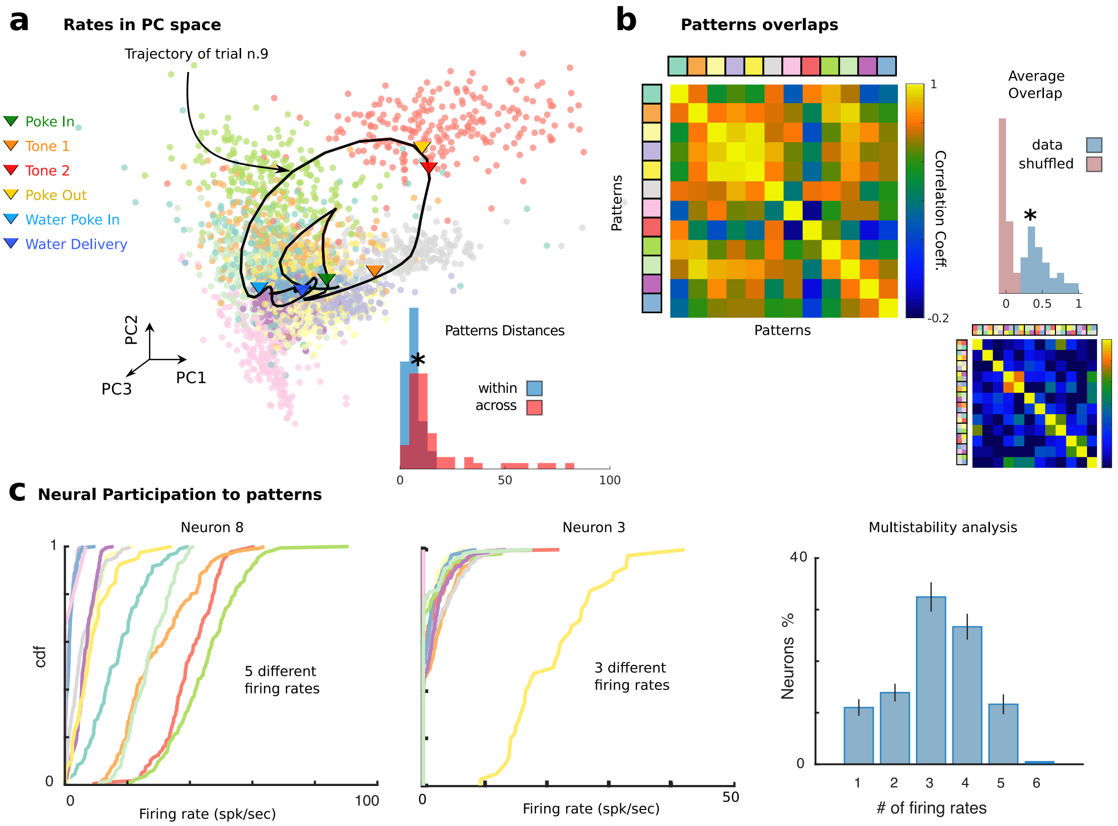

Figure 3. **Dense and distributed population code in M2.** a) Neural patterns cluster in Principal Component space (all trials from the representative session in Fig. 1; color-coded dots represent patterns in single trials; one representative trial smoothed trajectory overlaid where arrows show events onsets along trajectory). Inset: Distribution of within- and across-cluster distances between patterns (ranksum test $p < 2.0 \times 10^{-7}$) b) Pearson correlation matrix between patterns reveals significantly larger overlaps in the empirical data (top left: representative session) compared to those found when drawing random patterns from the empirical firing rate distribution (bottom right). Inset: Distribution of pattern correlations for empirical (blue) versus shuffled datasets (red). c) Single neuron firing rates are modulated by pattern sequences. Left: Cumulative firing rate distributions conditioned on patterns (color-coded as a) and Fig. 1d) for two neurons from the representative ensemble, revealing 5 and 2 significantly different firing rates across patterns, respectively (see Methods section 4.5. Right: Number of different firing rates per neuron revealed multistable dynamics where $88 \pm 2\%$ of neurons had activities modulated by patterns.

We then restricted our attention to correct trials and further reasoned that, because correct trials entailed the same set of actions in both patient and impatient conditions, if patterns encoded upcoming actions (as oppose to, e.g., reward size expectations) then pattern occurrence would be indistinguishable between the two kinds of trials. We indeed confirmed that the distribution of pattern dwell times was not significantly different between these two conditions in 95% of sessions (Kolmogorov-Smirnov test, $p > 0.05$). We thus concluded that the spatiotemporal variability observed in M2 population activity in single trials is consistent with a mechanism whereby specific pattern onsets anticipate self-initiated actions because they are causally upstream, as expected of motor regions (although we did not establish causality).

2.5 Correlated variability generates sequences of metastable attractors

What is a possible circuit mechanism underlying the observed pattern sequences? We aimed at capturing three main features of the empirical behavioral and neuronal data: (I) Patterns were long lived (0.5 s on average, Fig. 1f), more than one order of magnitude longer than single neuron time constants, suggesting they originate in attractor dynamics as an emergent network property. (II) Pattern dwell time distributions were variable and right-skewed (see Fig. 1f), suggesting that transitions between patterns may be noise-driven (see e.g. [36]). (III) Sequences of patterns were highly reliable, the same sequence consistently recurring across trials (Fig. 1e and Fig. S3). We thus wished to construct a mechanistic model generating reliable sequences of long-lived attractors, where transitions between attractors were driven by noise. First, we demonstrate a computational mechanism explaining the empirical features of the observed sequences, then we map the model components onto a mesoscale circuit, and finally we will test model predictions with experimental data.

The crucial ingredient driving transitions between patterns in the model entails constraining population activity fluctuations along a low-dimensional manifold within a high-dimensional of activity space. We achieved this by embedding a low-rank term in the synaptic couplings, whose structure was as follows. We modeled population activity in M2 as arising from a recurrent network of rate units governed by the following dynamics:

$$\tau \dot{u}_i(t) = -u_i(t) + \sum_{j=1}^{N} J_{ij}^{S} \phi_j(u_j(t)) + \zeta(t) \sum_{j=1}^{N} J_{ij}^{F} \phi(u_j(t)), \tag{2.1}$$

where u_i and $\phi_i(u_i)$ are post-synaptic currents and single-neuron current-to-rate transfer functions representing the activity of M2 neurons. Current-to-rate functions in the model were fit to those estimated from the empirical data in M2 (see Fig. 5a, Fig. S5, Methods and [37]). We hypothesized that patterns originated from p discrete attractors η^{μ}, for $\mu = 1, \ldots, p$, stored in the recurrent synaptic couplings $J_{ij}^{S} \propto \sum_{\mu=1}^{p} f[\eta_i^{\mu}] g[\eta_j^{\mu}]$, connecting a pre-synaptic neuron i and a post-synaptic neuron j in M2 (f and g are threshold functions, see Methods and [38]). This is consistent with experimental evidence supporting discrete attractor dynamics in secondary motor cortex [9, 31, 39]. Because we sought to generate transitions stochastically, the model operates in a regime where the attractors η^{μ} were stable in the absence of the second term J^{F} (Fig. S6a). Transitions between attractors, giving rise to sequences, originate from the asymmetric term $J_{ij}^{F} \propto \sum_{\mu=1}^{p} f[\eta_i^{\mu+1}] g[\eta_j^{\mu}]$ in Eq. (2.1), henceforth referred to as *correlated variability* term. This term generates stochastic dynamics via the noise $\zeta(t)$, with mean $\bar{\zeta}$ and variance σ_{ζ}^{2}. We will discuss below the mechanistic origin of this term.

The correlated variability term constrains population activity fluctuations onto a low-dimensional manifold within activity space, whose dimension is bounded by the number

p of attractors, thus much smaller than the number of neurons N. The effect of this term is to generate population activity fluctuations which are correlated across neurons. We found that, within a large range of parameters (Fig. S6b), the network model met all our objectives: (I) the model generated long-lived attractors (0.98±1.19s, Fig. 5b), whose duration was orders of magnitude longer than single-neuron time constants ($\tau = 20$ms, see Table S1), thus emerging from the network's collective dynamics. (II) Crucially, dwell time distributions were right-skewed with large coefficient of variability (Fig. 5c), capturing the large trial-to-trial variability observed in the empirical distributions of behavioral and neural data (Fig. 1). Since attractors would be stable in the absence of noise $\zeta(t)$ ((Fig. S6a), transitions between attractors were entirely noise-driven in this model. (III) Despite the variability of timing, the sequence of attractors was highly reliable, as in the empirical behavioral and neural data.

Furthermore, we found that single-neuron firing rate distributions were heterogeneous (Fig. 5d), similar to the empirical ones (Fig. 3c). In particular, most neurons participated in the sequential dynamics, attaining on average 3.8 ± 0.9 different firing rates across patterns, explaining the single-neuron multistability properties observed in M2 neural data (Fig. 5c, see also [12]). We conclude that metastable attractor dynamics in our model captured the lexically stable yet temporally variable features of pattern sequences observed in the empirical data.

2.6 Correlated variability originates in a mesoscale feedback loop

The crucial ingredient driving transitions between patterns in the model (see Eq. (2.1)) entails restricting fluctuations along a low-dimensional manifold within activity space. We achieved this by embedding a low-rank noise term in the synaptic connectivity architecture of the neural circuit. What is the circuit origin of these couplings? We found that this low-rank structure naturally arises from a two-area model, describing a feedback loop between a large recurrent circuit representing M2 and a small feedforward circuit (provisionally denoted as Y):

$$\begin{aligned}\tau \dot{u}_i(t) &= -u_i(t) + \sum_{j=1}^{N} J^S_{ij}\phi_j(u_j(t)) + \sum_{j=1}^{N_Y} W^{(M2\leftarrow Y)}_{ij} r_j \ , \qquad (2.2)\\ \tau_Y \dot{r}_i &= -r_i + \sum_{j=1}^{N} W^{(Y\leftarrow M2)}_{ij}\phi_j(u_j) \ .\end{aligned}$$

Here, u_i represent the activity of M2 neurons (same as in Eq. 2.1), and r_i represent activities of neurons in area Y. The latter area is smaller ($N_Y \ll N$) and faster ($\tau_Y < \tau$), and lacks recurrent couplings, suggesting it may correspond to a subcortical circuit. The asymmetric term J^F in Eq. (2.1), which generates stochastic transitions between otherwise stable M2 attractors, originates from the reciprocal couplings $W^{(Y\leftarrow M2)}$, $W^{(M2\leftarrow Y)}$ between M2 and area Y in Eq. (2.2), its temporal dependence arising from short-term plasticity at these synapses (see Methods [40]). The reciprocal connections W in this two-area model can be integrated out when dynamics in area Y are faster than in M2 ($\tau_Y < \tau$) [41, 42]. The two-area mesoscale model in Eq. (2.2) is then mathematically equivalent to the effective dynamics in Eq. (2.1), whose recurrent couplings are augmented to include an asymmetric term J^F, inherited from the reciprocal loop. In the Methods section we show how the mean and variance of the noise term $\zeta(t)$ in Eq. (2.1) capture, respectively, the strength and the variability of the couplings in the feedback loop between M2 and area Y. Its time dependence arises from short term plasticity in these couplings assuming area Y is small.

2.7 Correlated variability is necessary to explain temporal variability

Is it possible to generate the observed pattern sequences with alternative mechanisms, in the absence of correlated variability? We tested a large class of models where synaptic couplings included both symmetric (J^S) and asymmetric (J^F) terms, but where the asymmetric couplings were constant, i.e., setting $\sigma_\zeta = 0$ in Eq. (2.1). Depending on different ratios between asymmetric to symmetric couplings, this class of models generated either decaying activity, or stable attractors, or sequences of attractors when the ratio was large enough (Fig. S6a). However, in the whole region of parameter space with sequential dynamics, this class of models failed to capture crucial aspects of the data. Namely, dwell times distributions were short and they showed no trial-to-trial variability, thus being incompatible with the observed patterns (Fig. 1f).

We then attempted to rescue these models by driving the network with increasing levels of *private noise*, namely, external noise, independent for each neuron (Fig. S7a, see Methods). This led to small amounts of trial-to-trial variability in dwell times, but was still qualitatively different from the empirical data. Increasing the private noise level beyond a critical value destroyed sequential activity (Fig. S7b).

We reasoned that the difficulty in generating long-lived, right-skewed distributions of dwell times in this alternative class of models was due to the fact that transitions were not driven by noise, but by the deterministic asymmetric term J^F. Adding private noise did not qualitatively change variability, due to the high dimensionality of the stochastic component. Private noise induces independent fluctuations in each neuron; however, in order to drive transitions from one attractor to the next one within a sequence, these fluctuations have to align along one specific direction in the N-dimensional space of activities. The probability that independent fluctuations align in a specific direction vanishes in the limit of large networks, explaining why in the private noise model transitions cannot driven by noise. We thus concluded that correlated variability was necessary to reproduce the right-skewed distribution of pattern dwell time observed in the data.

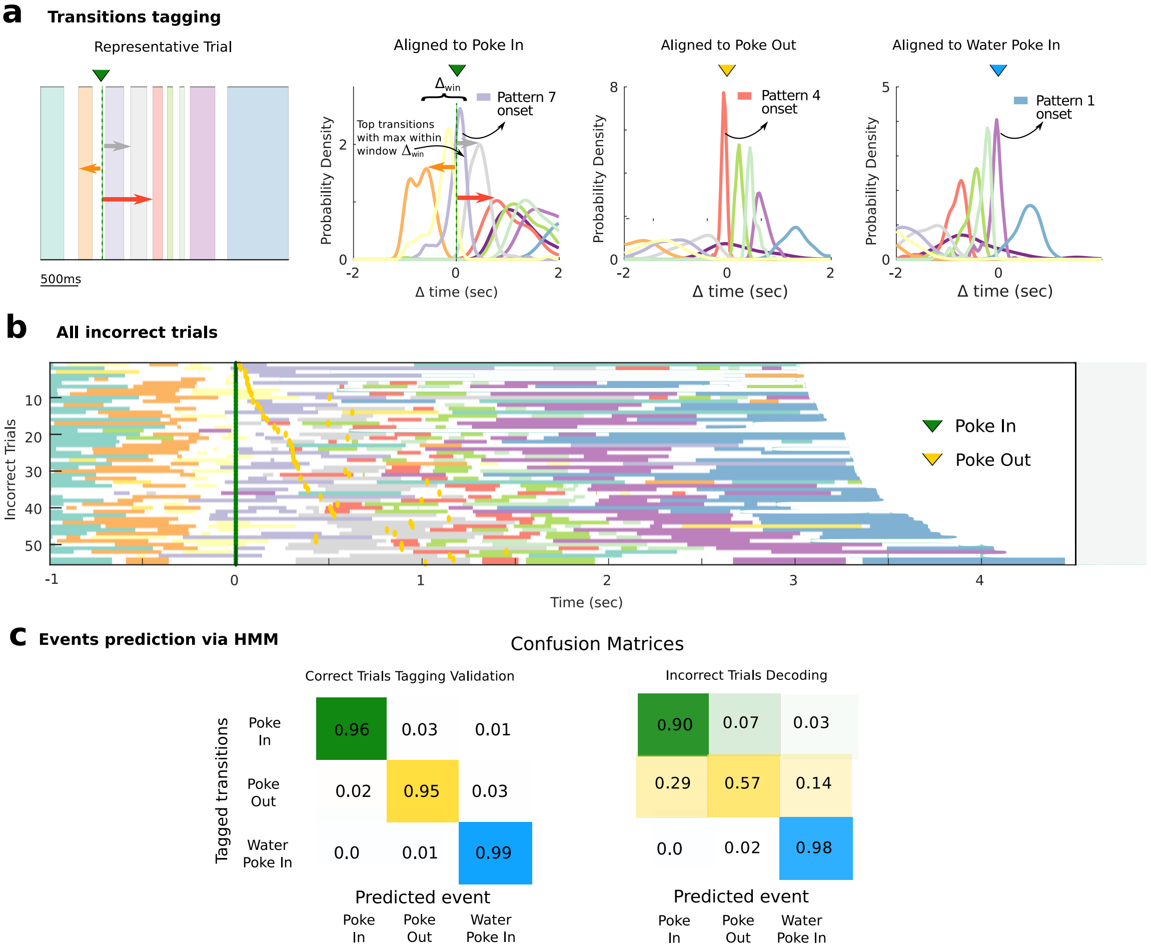

Figure 4. Predicting self-initiated actions from neural pattern onsets a) Schematic of pattern/action dictionary. Left: For each action in a correct trial (left: representative trial from Fig. 1d), pattern onsets are aligned to that action (Poke In in this example). The pattern whose median onset occurs within an interval $\Delta_{win} = [-0.5, 0.1]$ s aligned to the action, and whose distribution has the smallest dispersion, is tagged to that action (color-coded curves are distributions of action-aligned pattern onsets from all correct trials in the representative session in Fig. 1). b) In incorrect trials (55 trials from the same representative session; time $t = 0$ aligned to Poke In), the same patterns as in correct trials are detected (cfr. Fig. 1e), but they concatenate in different sequences. c) Predicting self-initiated actions from pattern onsets. Left panel: In correct trials (split into training and test sets), using a pattern-action dictionary established on the training set (procedure in panel a), action onsets are predicted on test trials (confusion matrix: cross validated prediction accuracy averaged across 41 sessions; hits: correct action predicted within an interval of $[-0.1, 0.5]$s aligned to pattern onset). Right panel: In incorrect trials, actions onsets are predicted based on the cross-validated dictionary established in correct trials.

2.8 Low-dimensional variability of M2 attractors dynamics

Our recurrent network model (see Eq. (2.1)) entails a specific hypothesis for the mechanism underlying the observed sequences: transitions between consecutive attractors are generated by correlated variability. We reasoned that, if this was the mechanism at play in

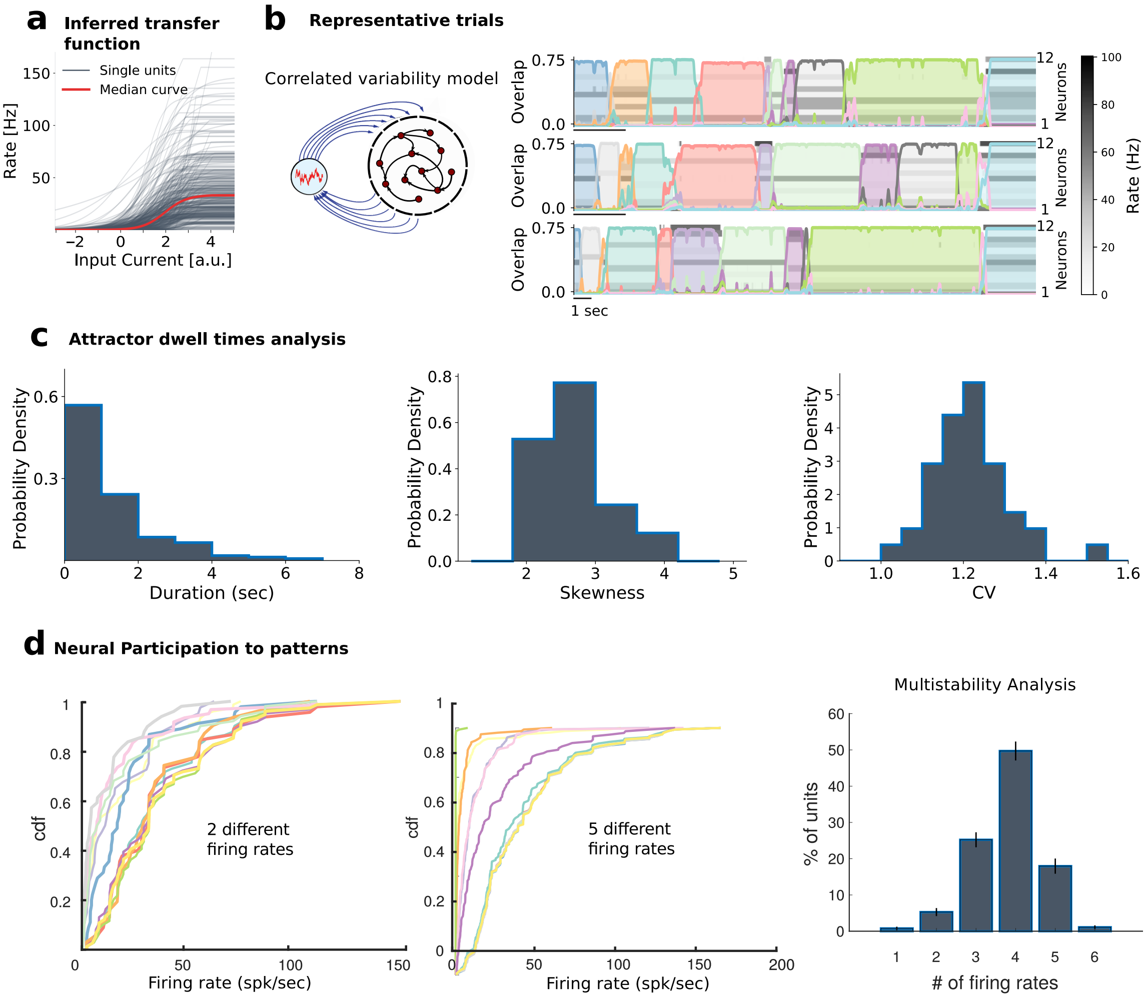

Figure 5. **Attractor model of pattern sequences.** a) Distribution of empirical single-cell current-to-rate transfer functions ϕ_i inferred from the data (366 neurons from 41 sessions), used as transfer functions in the recurrent network model (see Methods). b) The correlated variability model (Eq. 2.2) generates reliable sequences of long-lived attractors with large trial-to-trial variability in attractor dwell times (representative trials: rows represent the activity of 12 neurons randomly sampled from the network; color-coded curves represent time course of overlaps (see Eq. 4.9) between population activity and each attractor; detected attractors are color-shaded). c) Histogram of attractor dwell times across trials in the representative network of b) reveals right-skewed distributions (left, we excluded the first and last pattern in the sequence, whose duration artificially depends on trial interval segmentation). Skewness (center) and coefficient of variability (CV, right) of pattern dwell time distributions reveal large trial-to-trial variability (41 simulated networks). d) Single neuron firing rates are modulated by pattern sequences in the model. Cumulative firing rate distributions conditioned on attractors (color-coded as in b)) for two representative neurons in the model, revealing 2 and 3 significantly different firing rates across attractors, respectively (see Methods section 4.5. Inset: Number of different firing rates per neuron revealed multistable dynamics where 99 ± 1% of neurons had activities modulated by patterns.

driving sequences, then two clear predictions should be borne out in the neural population data. First, the correlated variability term in Eq. (2.1) predicts that population activity fluctuations within a given attractor (color-shaded intervals in Fig. 5b), henceforth referred to as "noise correlations", lie within a subspace whose dimension is much smaller than that expected by chance (Fig. 6a, dimensionality in the model vs. shuffled surrogate dataset, ranksum test, $p < 10^{-15}$). Second, the sequential structure of the correlated variability term in Eq. (2.1) implies that noise correlation directions for attractors that occur in consecutive order within a sequence should be co-aligned. A canonical correlation analysis showed that in the correlated variability model the alignment across attractor (measured using the top K principal components of the noise correlations, where K is its dimensionality) was much larger than expected by chance (Fig. 6b, alignment in the model vs. shuffled surrogate dataset, ranksum test, $p < 10^{-5}$). More specifically, we found that the strongest alignment occurred between consecutive attractors within a sequence, compared to attractors occurring further apart (Fig. 6b, ranksum test, $p < 10^{-20}$).

Having established strong statistical features regarding low-dimensional, aligned noise correlations structure, we tested whether the structure predicted by the model were observed in the M2 neural ensemble data. We defined noise correlations in the empirical data as population activity fluctuations around each neural pattern inferred from the HMM fit (Fig. 6a). Applying the same analyses to the data that were run on the model, we found that indeed empirical noise correlations around each neural pattern had lower dimension than expected by chance, and closely matched the dimensionality predicted by the model (Fig. 6a). CCA further revealed that noise correlations were highly aligned between patterns, significantly above the alignment expected by chance (Fig. 6b, rank-sum test $p = 1.70 \times 10^{-4}$). Finally, directions of variability were more aligned between consecutive patterns, compared to patterns further apart in the sequence, (rank-sum test, $p < 10^{-14}$; see Fig. 6c). Thus, the features of the noise correlations in the neural ensemble data were strongly consistent with the predictions from the correlated variability model.

3 Discussion

Our results establish a correspondence between self-initiated actions and attractor dynamics in secondary motor cortex (M2). We found that population activity in M2 during a self-initiated waiting task unfolded through a sequence of patterns, with each pattern reliably predicting the onset of upcoming actions. We interpreted the observed patterns as metastable attractors emerging from the recurrent dynamics of a two-area neural circuit. The model was capable of robustly generating reliable sequences of metastable attractors recapitulating the properties of the dynamics found in the empirical behavioral and neural data. We propose a neural mechanism explaining the variability in attractor dwell times as originating from correlated variability in a two-area model. The model predicts that population activity fluctuations around each attractor (i.e., "noise correlations") are highly aligned between attractors and constrained to lie on a low-dimensional subspace, and we confirmed these predictions in the empirical neural (M2) data. Our work establishes a

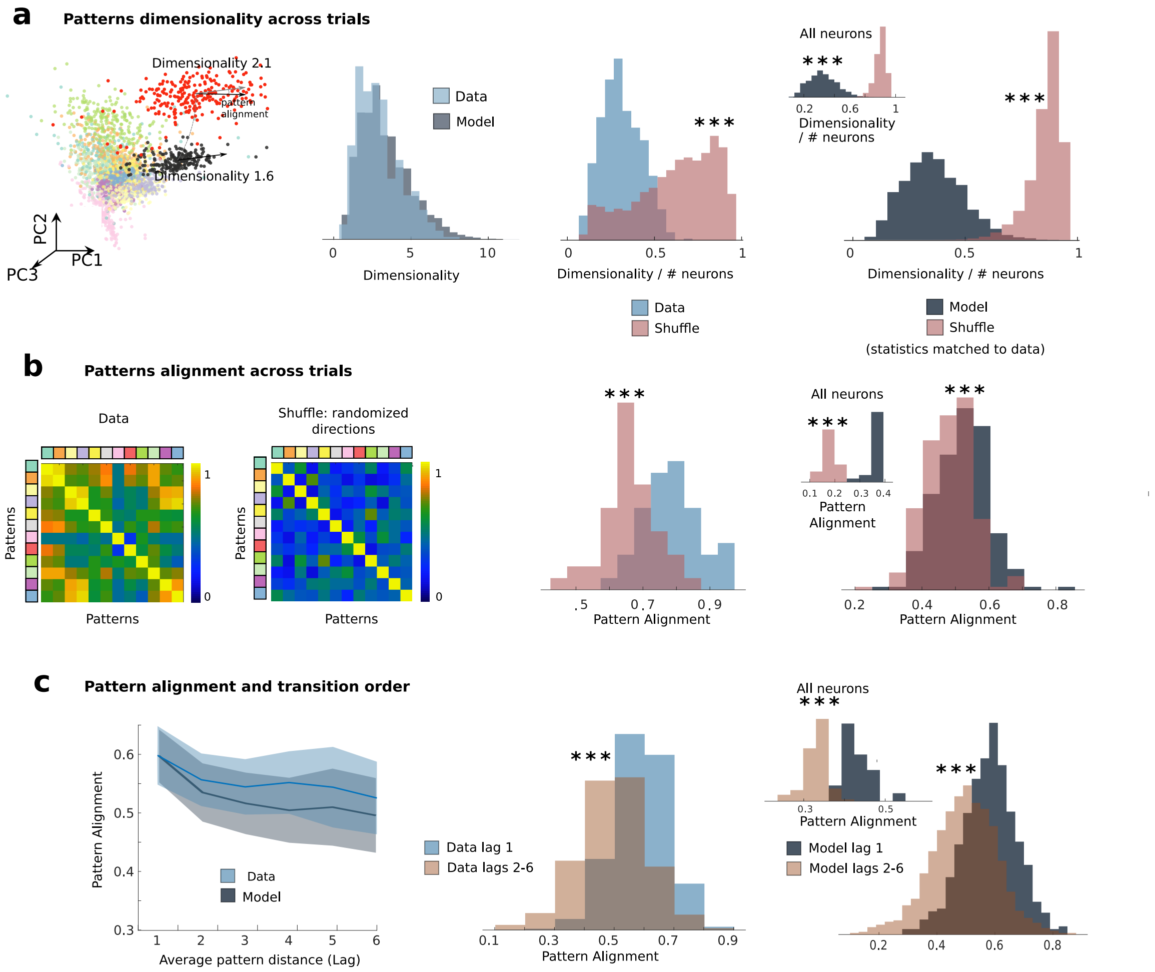

Figure 6. Low-dimensional variability in models and data. a) Comparison of dimensionality of pattern-conditioned noise correlations in the data (blue) and the model (grey) reveals low-dimensional population activity fluctuations, significantly smaller than expected by chance (red, shuffled datasets). From left to right: first panel, representative session as in Fig. 1); second panel, summary across 41 sessions from the data and the model; third panel, fractional dimensionality in the data; fourth panel: model dimensionality estimated by matching ensemble sizes and number trials to data across 41 simulated sessions; inset: dimensionality estimated from N=10000 neurons in 41 simulated sessions. Shuffled datasets are realized with matched statistic by selecting random principal components as compared to top principal components. b) Pattern-conditioned noise correlations are highly aligned between patterns in the data. Alignment between top canonical correlation vectors (blue, data; gray, model) is larger than between random principal component directions (red). c) Left: Alignment of noise correlations between each pattern and patterns occurring at lag n in the sequence (e.g., $n = 1$ represents patterns immediately preceding or following the reference pattern) in the model (grey) and in the data (blue). Pattern alignments are significantly larger for patterns at one lag compared to patterns at longer lags. All panels: $* = p < 0.05$, $*** = p < 0.001$.

mechanistic framework for investigating the neural underpinnings of self-initiated actions and demonstrates a novel link between correlated variability and attractor dynamics.

3.1 Evidence for discrete attractor dynamics in cortex

Evidence from primate and rodent studies supports the hypothesis that cortical circuits generate discrete activity patterns, interpreted as attractors, during working-memory and decision-making [43]. Attractors are characterized by long periods where neural ensembles discharge persistently at approximately constant firing rate (defining a neural pattern) punctuated by relatively abrupt transitions to a different relatively constant pattern. Selective persistent activity during delay periods has been reported in temporal [44, 45] and frontal areas [46, 47] in primates, and frontal areas in rodents [9, 31, 39, 48]. Evidence for attractors encoding sensory stimuli was found in rodent sensory cortex [11, 12, 17]. Optogenetic stimulation of few neurons within an ensemble was shown to drive sustained activation of the entire ensemble [49], persisting for seconds after stimulation offsets [50], compatible with predictions from attractor models [51].

Attractor dynamics may depend on the precise temporal structure of the task: in similar tasks, depending on whether the delay period was of fixed or randomized length, either discrete attractors or ramping activity were observed [9]. Our waiting task differs from these delayed memory tasks in two respects. First, we analyzed the entire p-step action sequence of freely moving animals, rather than a single step of behavior (i.e. delay period). Second, many steps within the action sequence were self-initiated, rather than prompted by experimenter-controlled signals. Consequently, we uncovered a new dynamical regime in which (i) corresponding behavioral and neural states were metastable, with large trial-to-trial variability in dwell times (ranging from hundreds of ms to a few seconds); ii) states were concatenated into a sequence, reliably occurring in most trials, in accordance with the actual action sequence. Experimental evidence for stimulus-specific sequences of metastable attractors was previously found in primate frontal areas [15, 16, 52] and rodent sensory areas [11]. Random sequences were also observed during ongoing periods [12, 18, 26]. In all those cases, and consistent with our results, state dwell times showed large trial-to-trial variability captured by Markovian dynamics (i.e. right-skewed distributions), suggesting an underlying stochastic process driving transitions [12, 21, 25]. A novel feature of our results is that the sequence of attractors is not driven by external stimuli, but rather is internally generated.

3.2 Network models of sequences

The main features of M2 ensemble activity targeted by our two-area mesoscale attractor network (2AMAN) model were the reliable identity and order of long-lived attractors occurring in a sequence, and the large trial-to-trial variability of attractor dwell times, whose distributions are characterized by a positive skewness and large CV. In our model, we showed that both features can be robustly attained when transitions between attractors arise from correlated variability.

Previous network models could achieve either sequence reliability or variability in dwell time distributions, but not both. Models generating reliable pattern sequences include synfire chains [53, 54]. These models rely on a fine tuned connectivity structure producing pattern dwell times with short duration and low variability. While these dynamics are well suited to explain neural activity observed in songbird HVC [55, 56] or mammalian hippocampus [57], their features are not compatible with the observed M2 ensemble activity. Reliable pattern sequences can otherwise be triggered by specific cues in recurrent networks with asymmetric connectivity structure [27, 28, 58–61], trained with unsupervised learning rules [29, 56, 62, 63] or in reservoir networks [64]. However, pattern dwell times in such models are short, set by single-neuron characteristic time constants, and show little trial-to-trial variability. Pattern dwell times could be increased via synaptic delays [27]. However, none of these models is capable of generating large trial-to-trial variability in dwell time distributions and are thus incompatible with the observed M2 data.

Long-lived patterns of neural activity can be sustained by attractor dynamics, where the reverberating activity of neural assemblies is sculpted by the recurrent couplings [51, 65, 66]. Previous attractor networks were shown to generate sequences of long-lived metastable patterns whose features, however, are incompatible with the ones we observed in M2 neural ensembles. In particular, networks with clustered architecture can give rise to metastable attractors with large trial-to-trial variability in dwell time distributions [12, 22, 23, 25]. However, metastable attractors in these models concatenated in random sequences, incompatible with the highly reliable sequences we observed in M2. The reason is that, in these networks, each cluster generates independent fluctuations within activity space, realizing a high-dimensional stochastic process, akin to the private noise model in Fig. S7. These fluctuations drive transitions along random directions in activity space, thus unreliable across trials (when concatenating more than two states in a sequence [67]). To drive a specific transition, independent fluctuations would have to align along a specific direction within the high-dimensional activity space, and the probability of this event occurring vanishes for large network size. Thus it is challenging to generate reliable yet noise-driven sequences in these models. We confirmed this intuition by showing that no regime of parameters allowed transitions with right-skewed attractor dwell times in a private noise model (Fig. S7). One may drive clustered networks with strong time-dependent stimuli to pace activity along stimulus-specific sequences [12, 21, 25, 26, 68]. However, this would merely shift the problem of reliable sequence generation from the local circuit to an upstream area producing the specific input (but see [69]).

By introducing correlated variability to attractor networks, our 2AMAN model provides a circuit mechanism to overcome the curse of dimensionality. By constraining noise correlations onto a low-dimensional manifold, the 2AMAN model attained reliable sequences of long-lived attractors with large trial-to-trial variability in dwell time distributions.

3.3 Neural circuits underlying attractor dynamics

Our 2AMAN mechanistic model naturally captured the essential features of the observed M2 population dynamics. How does the model architecture map onto specific neural circuits? Previous work showed, using inactivation experiments, that the stochastic component in action timing variability originated downstream of the medial prefrontal cortex and presumably in M2 [7]. Our model provides a mechanistic explanation in the form of correlated variability, inherited from synaptic couplings between a large recurrent network representing M2 and a smaller and faster network lacking recurrent couplings. We tentatively hypothesize that this latter network is instantiated by a small subcortical circuit connected to M2, such as the areas that comprise its basal ganglia or thalamic partners. Recent evidence from perturbation experiments in rodents supports a scenario where a mesoscale circuit sustaining attractor dynamics includes the thalamus and motor [31] or prefrontal cortex [39]. This hypothesis could be directly tested via perturbation experiments, and we leave this question for future work. It is plausible that a larger mesoscale network may underlie sequence generation, including cortex, thalamus, and basal ganglia [32–34, 70]. Even though our results suggest that timing variability of multi-step sequences extending on the order of 10 seconds may originate within a cortical-subcortical loop, we did not investigate how even more complex natural behavior could emerge at longer timescales [1, 2]. This may also involve a larger and more distributed mesoscale network [71]. Flexibly switching between different action sequences to adapt to a changing environment may recruit other areas including the anterior cingulate [7] and the basal ganglia [3, 34, 60, 72, 73]. We hope to address these broader issues in future work.

A large amount of evidence implicated preparatory activity in rodent M2, specifically the antero-lateral motor cortex, in action and movement planning both in forced-choice tasks [9, 48, 74–77] as well as self-initiated tasks [6, 7]. The sequences of metastable attractors we uncovered in M2 were consistent with the features of preparatory activity [72]: a precise dictionary linked specific attractors to actions; attractor onset reliably predicted action onset; and action timing variability strongly correlated with attractor onset variability. In particular, we were able to predict actions in incorrect trials using the cross-validated attractor/action dictionary established in correct trials, consistent with recent findings [50]. This interpretation is further supported by the fact that M2 population activity was only modulated by self-initiated actions and M2 neurons were not responsive to other task features such as trial type (patient vs. impatient, or correct vs. incorrect) and reward expectation (large vs. short).

3.4 Correlated variability in sensory vs. motor processing

The main conceptual innovation in our 2AMAN model is the introduction of low-dimensional correlated variability driving reliable sequences with variable timing. Similar “motor noise correlations" have been recently reported during vocal babbling in juvenile songbirds. Correlated variability was present in a motor area (RA) but absent in a premotor area (LMAN) [68], and a mechanism for the emergence of correlated variability via topographic organization of projections from LMAN to RA was proposed. This is in contrast to our observations of correlated variability in M2, a premotor area. In our 2AMAN model, correlated variability arises from a feedback loop between a high- and a low-dimensional recurrent network, mediated by asymmetric couplings in the synaptic connectivity matrix. Asymmetric couplings have been previously used to generate specific temporal dynamics, though in the absence of noise. Examples include models of temporal sequences [27, 28] or recurrent networks within the echo-state/reservoir computing framework [78–81]. All these models are fundamentally different from ours, as their low-rank structure is fixed and time-independent, hence their temporal dynamics entails no trial-to-trial variability. In our model, on the other hand, the low-rank structure is time-varying and generates correlated fluctuations at a fast timescale. We confirmed that low-dimensional correlated fluctuations around each attractor are present in the empirical M2 data. Alternative models where each neuron's fluctuations are independent (i.e., with private noise) failed to capture the observed temporal variability and predicted high-dimensional fluctuations, which were absent in the empirical data.

Low-dimensional correlated variability has been widely reported in sensory cortex, where it may carry information about the animal's state of locomotion [82] or arousal [83, 84], facial and whisker movements [85–87], or attentional state [88, 89]. It has been proposed that low-dimensional correlated variability in sensory cortex, in the form of differential correlations, may be detrimental to sensory processing as it may limit a network's information processing capability [35]. Alternatively, correlated variability may arise from top-down feedback projections carrying task-related information [90]. Here, we found that low-dimensional correlated fluctuations are the crucial mechanism enabling neuronal sequences to unfold with variable timing. It is likely that variable timing is an adaptive feature of motor behavior. Amongst other possible functions, such as avoiding predation or competition, timing variability allows animals to explore the temporal aspects of a given sequence of behavior independently of the choices of actions. We speculate that exploration could allow learning of proper timing by a search in timing space independent of action selection and vice-versa, as may be the case in songbirds [68, 73, 91]. Our results thus suggest that low-dimensional correlations are essential for motor generation.

4 Methods

4.1 Experimental procedures

Behavioral task. Rats were trained on the self-initiated waiting task (Fig. 1a) in a behavioral box containing a Wait port at the center and a Reward port at the side (entry/ exit from ports were detected via infrared photo-beam). Rats self-initiated a trial by poking into the Wait port ("PokeIn"). If the rat stayed in the Wait port for T1 delay (0.4s), the first tone played (tone 1; 6 or 14 kHz tone), signaling availability of reward in the Reward port. If the rat waited in the Wait port after tone 1, then tone 2 was played after a T2 delay (14 or 6 kHz, different from tone 1). If the rat visited the Reward port after tone 2, a large water reward (40 μl) was delivered after a 0.5s delay (patient trial). If the rat poked out after tone 1 but before tone 2, and visited the Reward port, a small water reward (10 μl) was delivered after a 0.5s delay (impatient trial). The rat had to visit the Reward port within 2s after the poke out to collect rewards. These trials were referred to as "correct trials;" trials were the animal performed different action sequences were deemed "incorrect trials." If the rat poked out before tone 1, no rewards were made available. Re-entrance to the Wait port was discourage with a brief noise burst. T2 delay was drawn from an exponential distribution, with minimum value 0.7s and mean adjusted to achieve patient trials in one third of the session. After reward delivery, an inter-trial interval (ITI) started during which white noise played. The time from the PokeIn to the ITI end was held constant, so that the rat could not increase reward collection by leaving the Wait port fast with the goal to start the next trial early. The optimal strategy was thus to always wait for tone 2. To test whether neuronal responses depended on a specific action, 3 rats were trained on two variants of the task. In these experiments, a different behavioral box contained a Reward

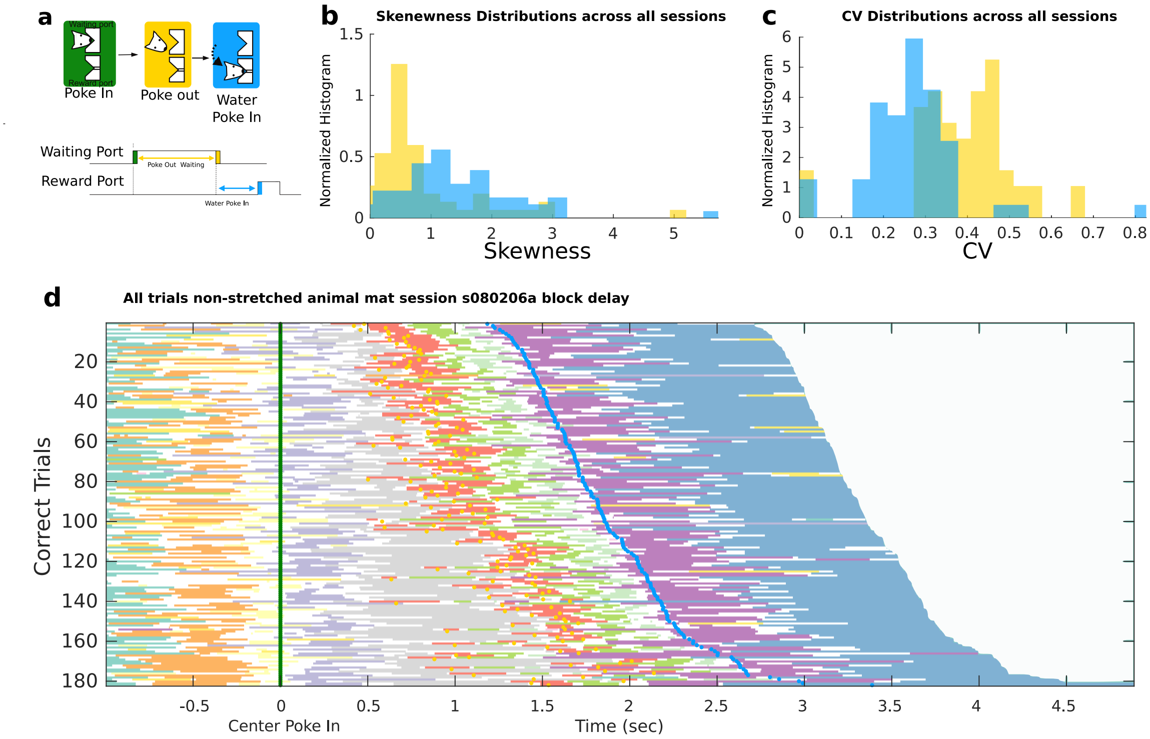

Figure S1. **Interevent interval distributions**. a) Schematic of the interevent time intervals between Poke In and Poke Out (yellow) and between Poke Out and Water Poke In (blue). b) Histogram of skewness of interevent interval distributions across all sessions revealed right-skewed distributions. c) Histogram of coefficients of variability (CV) of interevent intervals across all sessions revealed large trial-to-trial variability variability in interevent times. d) Pattern sequences in all trials, ordered by duration (blue dots represent Water Poke In). As opposed to Fig. 1e where trials time courses were stretched, trials in d represent actual time courses.

port, a nose-poke Wait port, and a lever-press Wait port. Blocks of nose-poke trials and lever-press trials were interleaved in each session. In the nose-poke block, the rat was to perform the same task as above. In the lever-press block, task rules were the same but the rat had to wait for the tones by keeping the lever pressed. The wrong action (nose-poke waiting in the lever-press block and vice versa) was not rewarded and classified as "incorrect trials." Each block last for 70-100 trials. Transitions between the blocks were not signaled. 41 sessions (7 rats) were recorded, see [6] for extensive details.

Electrophysiological data. Rats were implanted with a drive containing 10-24 movable tetrodes targeted to the M2 (3.2-4.7 mm anterior to and 1.5-2.0 mm lateral to Bregma). Electrical signals were amplified and recorded using the NSpike data acquisition system (L.M. Frank, University of California, San Francisco, and J. MacArthur, Harvard University Electronic Instrument Design Lab). Multiple single units were isolated offline by manually clustering spike features derived from the waveforms of recorded putative units using MCLUST software (A.D. Redish, University of Minnesota). Tetrode depths were adjusted before or after each recording session in order to sample an independent population of neurons across sessions. See [6] for details.

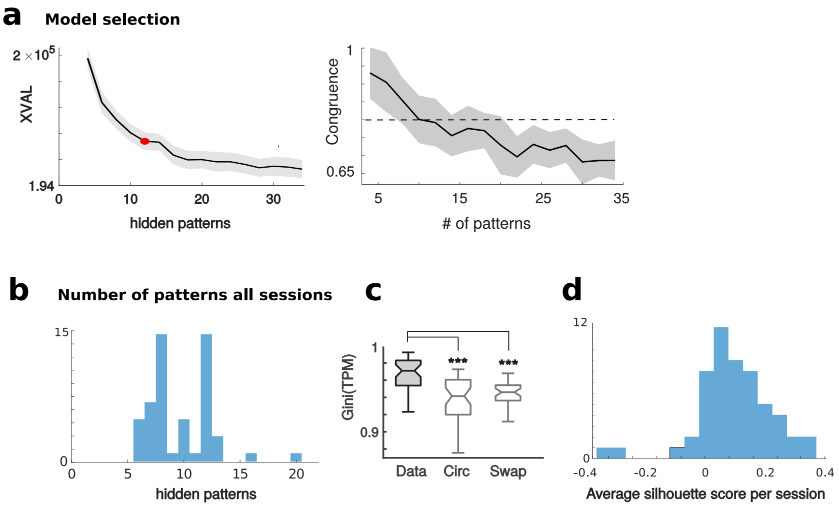

Figure S2. Empirical data fit via hidden Markov models (HMM). a) Model selection in a representative session. Left: The number of patterns was selected via an unsupervised 10-fold cross-validation procedure. The number of patterns yielding the largest curvature in the likelihood trend is selected (XVAL=$-2\times$log-likelihood of held-out trials plateaus for increasing number of patterns). Right: For a fixed number of patterns, the similarity between the HMM parameters (cross-validated congruence, see Methods) optimized in different folds drops below 0.8 (dashed line) when the number of patterns grows beyond the value selected via cross-validation (see [92, 93]). b) Distribution of the number of patterns across all sessions. c) Gini coefficient distribution for the pattern transition probability matrices (TPM) across all sessions, compared to shuffled datasets. Empirical TPMs are sparser than TPMs inferred from surrogate datasets (see Fig. 2, $*** = p < 0.001$). d) Distribution of within- and across-cluster distances between patterns measured with silhouette scores (ranksum test, $p < 2.0 \times 10^{-7}$, see Fig. 3a).

4.2 Neural data analysis

Data analyses was performed with custom-written software using MATLAB (Mathworks) and Python. No statistical methods were used to pre-determine sample sizes, but sample sizes were similar to previous studies [48, 77]. All summary statistics are mean±SD across 41 sessions, unless otherwise stated.

4.3 Pattern sequence estimation

A Poisson-Hidden Markov Model (HMM) analysis was used to detect neural pattern sequences from simultaneously recorded activity of ensemble neurons. Here, we briefly describe the method used and refer to Refs.[12, 25] for details. According to the HMM, the network activity is in one of M hidden "patterns" at each given time. A pattern is a firing rate vector $r_i(m)$ (the "emission matrix", Fig. 1c), where $i = 1, \ldots, N$ is the neuron index and $m = 1, \ldots, M$ identifies the pattern. In each pattern, neurons discharge as stationary Poisson processes (Poisson-HMM) conditional on the pattern's firing rates $r_i(m)$. Stochastic transitions between patterns occur according to a Markov chain with transition matrix (TPM, Fig. 1c) T_{mn}, whose elements represent the probability of transitioning from pattern m to n at each given time. We segmented trials in 5 ms bins, and the observation of either

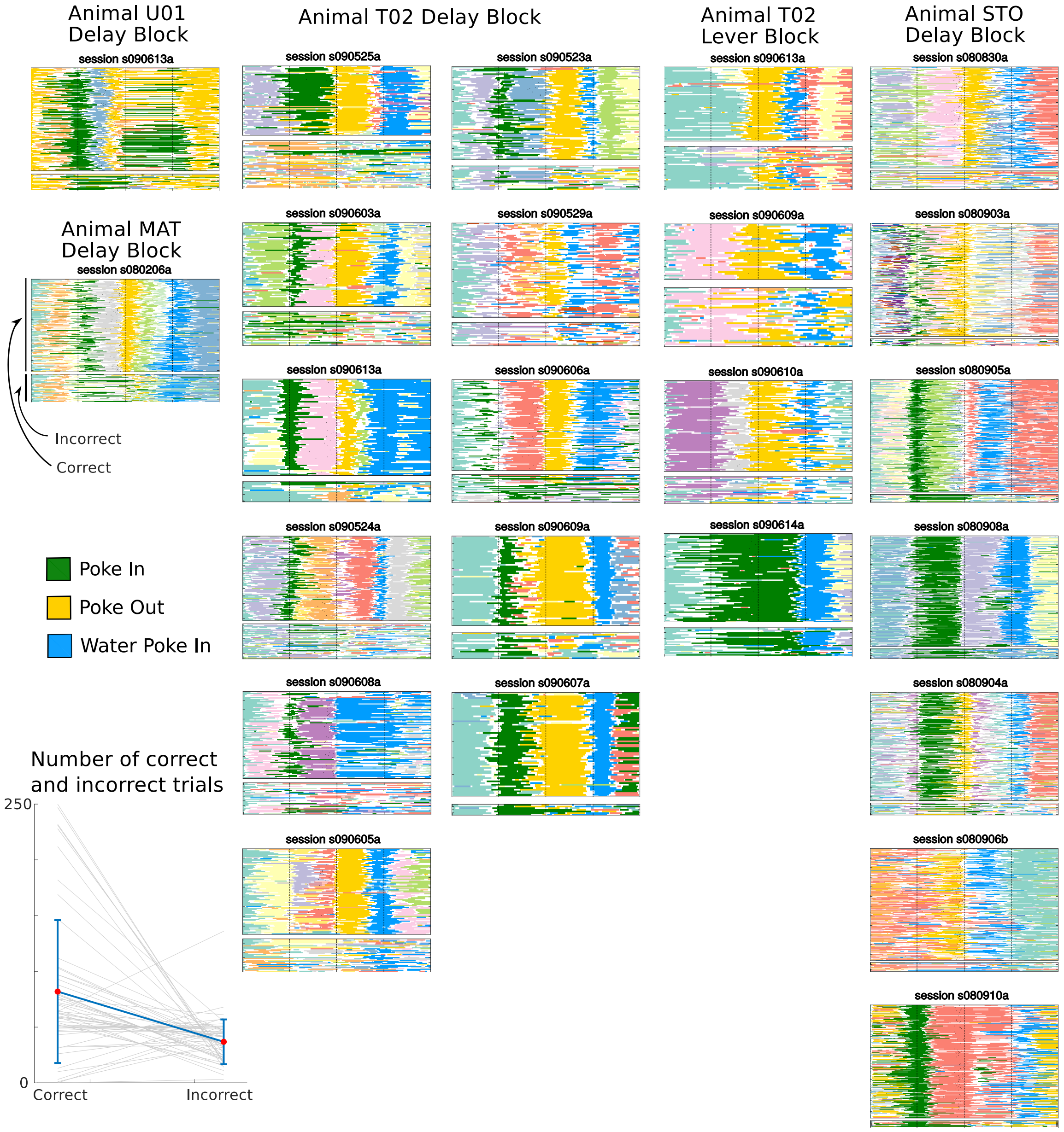

Figure S3. **Stretched and event-aligned pattern sequences for all sessions.** In each session, neural patterns from correct trials (top subplots, same notations as in Fig. 1e) show reliable sequences, as expected from the fixed action sequences to be performed to collect a reward; pattern sequences in incorrect trials are less reliable (bottom subplots, see Fig. 4b), as expected from the inconsistent behavior in those trials. In most sessions, animals only performed the delay task ("Delay block"); in sessions where delay and lever tasks ("Lever block") were interleaved, block trials from the two tasks were plotted separately (only for animal T02). The patterns tagged to one of the three events analyzed (cf. legend and Fig. 4) are consistently colored across session. The remaining patterns do not follow a consistent color code. Inset in bottom left: Number of correct and incorrect trials in each session. Session s090206a is the representative session in Fig. 1.

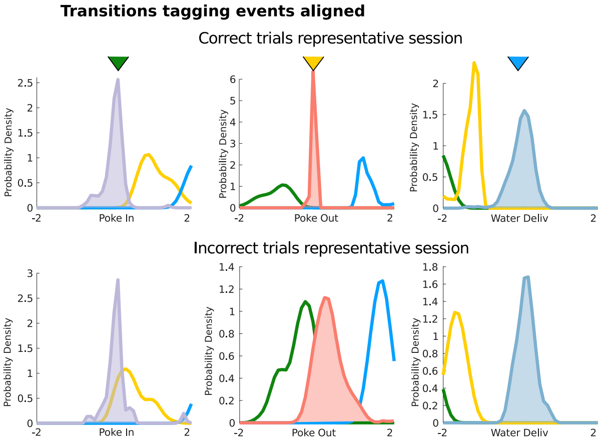

Figure S4. Predicting actions from patterns in incorrect trials. After tagging transitions to events using correct trials (top row, left to right: Poke In, Poke Out, Water Poke In; see Fig. 4), transitions were subsequently aligned to the same events but in incorrect trials (bottom row). In each subplot, the distribution of transition times for the corresponding event (e.g. Poke In in the top left: purple-filled histogram shows distribution of "Poke In" transition times aligned to the Poke In events) is compared to events time distribution for other events (yellow and blue curves).

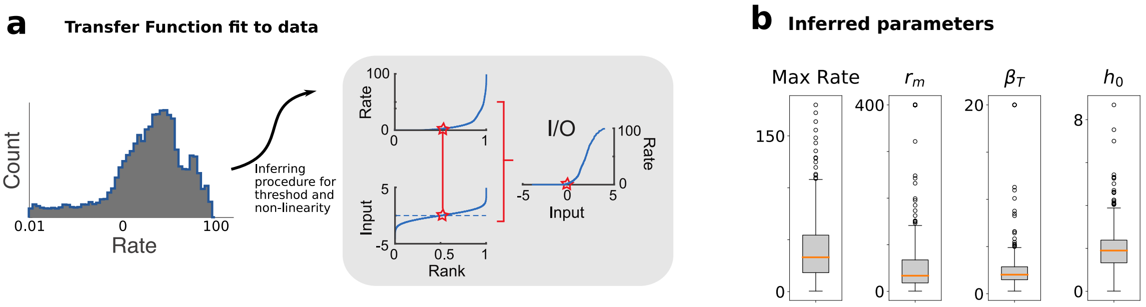

Figure S5. Inferring transfer functions from pattern sequences a) Procedure to infer single-neuron current-to-rate transfer functions from the data. The empirical distribution of firing rates across patterns for a representative neuron (left) was rank-matched to a standardized normal distribution of input currents (top and bottom left panels in grey box), obtaining the current-to-rate function (right). The star in each plot corresponds to the median value. b) Each single-cell current-to-rate function was fit to a sigmoidal function, yielding a distribution of fit parameters (366 neurons from 41 sessions; see Methods, Eq. (4.18)).

$y_i(t) = 1$ (spike) or $y_i(t) = 0$ (no spike) was assigned to a bin at time t for the i-th neuron (Bernoulli approximation); if in a given bin more than one neuron fired, a single spike was randomly assigned to one of the active neurons. A single HMM was fit to all correct trials per session, yielding emission probilities and transition probabilities between patterns, optimized via the Baum-Welch algorithm with a fixed number of hidden patterns M (iterative maximum likelihood estimate of parameters and latent patterns given the observed spike trains).

The number of patterns M is a model hyperparameter, optimized using the following

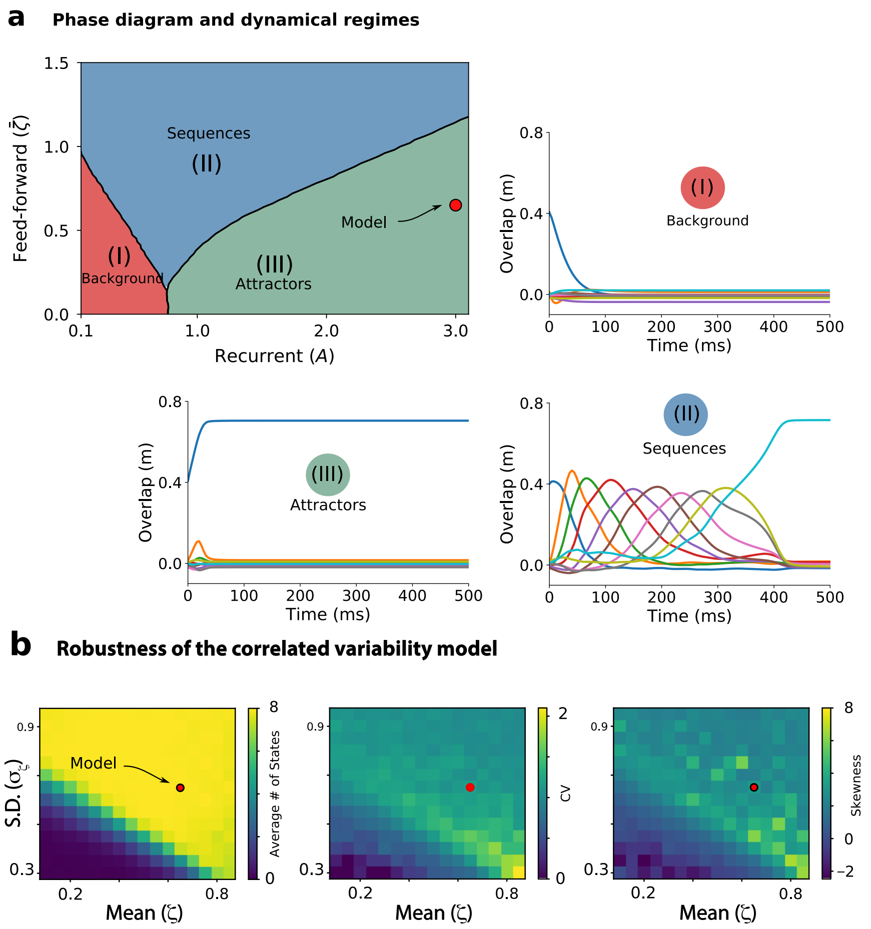

Figure S6. **Phase structure and model robustness.** a) Phase diagram for a recurrent network in the absence of noise (model in Eq. (4.4) with $\sigma_\zeta = 0$; average over 100 network realizations). The number of attractors visited per trial depends on the strength of the average recurrent (x-axis: the parameter A_S represent the strength of J^S in Eq. (4.5)) and feedforward couplings (y-axis: the parameter $\bar{\zeta}$ represents the strength of J^F in Eq. (4.5)). The model activity decays to zero for low values of the couplings (phase I, red); it generates stable attractors for large recurrent weights (phase II, green) or sequences of attractors for large feedforward weights (phase III, blue). In the representative trials, color-coded curves represent the time course of the overlaps (see Eq. (4.9)) between network activity and attractors. The red dot in the phase diagram shows the values of J^S and J^F used in the Results for the correlated variability model (upon adding multiplicative noise for J^F). b) Robustness of the correlated variability model. When sistematically varying the mean $\bar{\zeta}$ and the standard deviation σ_ζ of the noise $\zeta(t)$ in correlated variability model (corresponding to the value of J^S and J^F from panel a), red dot), the network robustly generates long attractor sequences (with an average of $p \approx 8$ attractors in each sequence, left) with large CV (center) and skewness (right) in attractor dwell time distributions over 200 trials.

model selection procedure [18]. In each session, we used K-fold cross-validation (with $K = 20$) to train an HMM on $(K-1)-$folds and estimate the log-likelihood of the held-out trials $LL(M)$ as a function of number of patterns M in the fit (see Fig. S2). The held-out $LL(M)$ increases with M, until reaching a plateau. We selected the number of patterns M^*

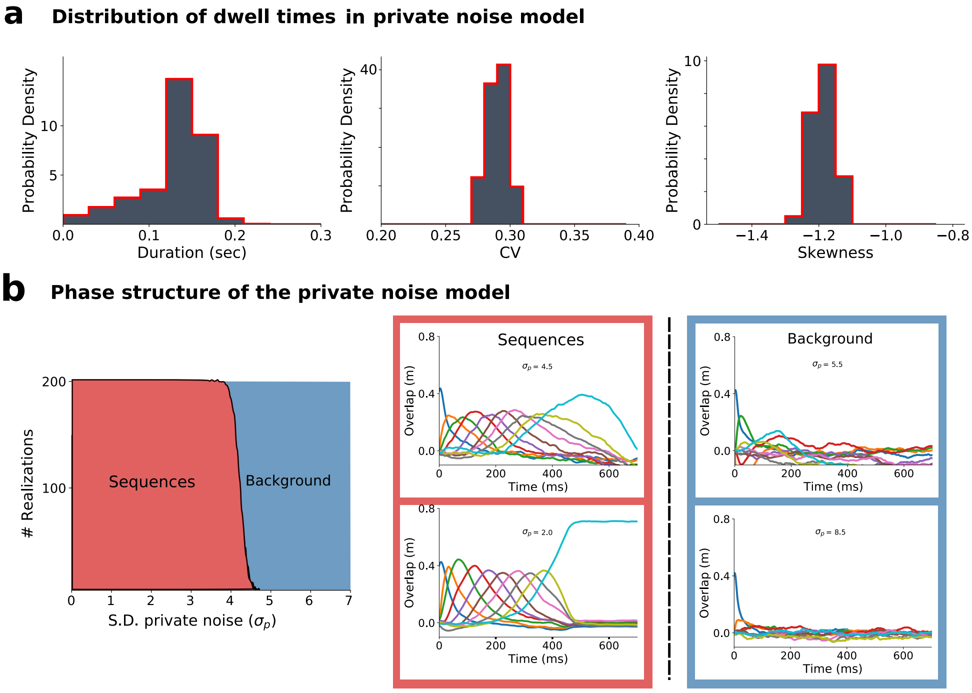

Figure S7. **Phase structure of models with recurrent and feedforward couplings.** a) In a model with recurrent and feedforward couplings, adding private noise with variance (see Methods, Eq. (4.8; network parameters as in Table S1, except for $A_S = 1$, $\bar{\zeta} = 0.65$, $\sigma_\zeta = 0$, $\sigma_p = 2$) introduces a small amount of trial-to-trial variability in pattern dwell time distributions (left), yielding small non-zero CVs (center) and negative skewness (right) across different networks. The dwell time statistics are qualitatively different from the empirical ones (cf. Fig. 1f). b) Left: Beyond a critical value of noise strength σ_p, the private noise model breaks down. For low private noise, networks generate sequential dynamics (red area, networks with sequential dynamics out of 200 network realizations), though with low CV and skewness; for larger values of private noise, activity decays to zero (blue). Right: Representative trials with 4 different values of the noise strength σ_p.

for which the incremental increase $LL(M+1) - LL(M)$ had the largest drop (the point of largest curvature) before the plateau. For control, we performed model selection using an alternative method, the Bayesian Information Criterion [25], obtaining comparable results (not shown).

To gain further insight into the structure of the model selection algorithm, we performed a post-hoc comparison between the parameters optimized on the training set for each value of M (number of patterns), across the cross-validation K-folds. In particular, we estimated the similarity between the optimized features (emission $r_i^{[k_1]}(m)$ and transition matrices $T_{mn}^{[k_1]}$) in the k_1-th fold and the k_2-th fold for given M, according to the following congruence $C(k_1, k_2)$ measure [93]:

$$C(k_1, k_2) = \left(\sum_{m=1}^{M} \sum_{i=1}^{N} \hat{r}_i^{[k_1]}(m) \hat{r}_i^{[k_2]}(m) \right) \cdot \left(\sum_{m,n=1}^{M} \hat{T}_{mn}^{[k_1]} \hat{T}_{mn}^{[k_2]} \right) ,$$

where N is the ensemble size, $\hat{r}_i^{[k]}(m) = r_i^{[k]}(m)/||\overrightarrow{r}^{[k]}(m)||_2$ is the normalized emission for pattern m, and $\hat{T}_{mn}^{[k]}$ is the normalized transition matrix $\hat{T}_{mn}^{[k]} = T_{mn}^{[k]}/||T^{[k]}||_2$. Features were matched across folds using the stable matching algorithm [94]. If the two folds yielded identical parameters, one would find $C(k_1, k_2) = 1$. A congruence above 0.8 signals good quantitative agreement between different folds, whereas congruence below 0.6 suggests a poor similarity among folds [92]. We calculated the average congruence across all fold pairs for given M and verified that the number M^* of patterns selected with the cross-validation procedure above corresponded to the elbow in the congruence curve (see Fig. S2a). For larger number of patterns, average congruence typically fell below 0.8.

The Baum-Welch algorithm only guarantees reaching a local rather than global maximum of the likelihood. Hence, for each session, after selecting the number of pattern M^* as above, we ran 20 independent HMM fits on the whole session, with random initial guesses for emission and transition probabilities, and kept the best fit for all subsequent analyses. The winning HMM model was used to infer the posterior probabilities of the patterns at each given time $p(m, t)$ from the data. Only those patterns with probability exceeding 80% in at least 50 consecutive ms were retained (henceforth denoted simply as patterns, Fig. 1d). This procedure eliminates patterns that appear only very transiently and with low probability, also reducing the chance of over-fitting. Pattern dwell time distributions (Fig. 1f) within each session were estimated from the empirical distribution of interval times where a pattern's probability was above 80%.

4.4 Comparison with surrogate datasets

We compared the HMM analysis of the empirical dataset with two surrogate datasets, obtained with the following shuffled procedure (Fig. 2, [13]). In the "circular" shuffle, each neuron's binned spike counts were circularly shifted within-trial randomly (row-wise circular shift), preserving autocorrelations but destroying pairwise correlations. In the "swap" shuffle, binned population spike counts were randomly permuted in time (column-wise swap), preserving pairwise correlations but destroying autocorrelations. For comparison of the real dataset with shuffled ones, we adopted the same K-fold cross-validation procedure as above, where an HMM was fit on training sets and the posterior probabilities $p(m, t)$ of patterns were inferred from observations in the held-out trials (test set).

From the pattern posterior probabilities inferred on held-outs, we estimated several observables for comparison between real and shuffled datasets. Pattern detection confidence was estimated as the fraction of a trial length where a pattern was detected with high confidence ($p(m, t) > 80\%$). Sparseness of transitions was estimated as the average Gini coefficient of TPMs obtained from the K training sets. We also estimated the across-trials sequence similarity as follows. In a trial where patterns were detected above 80% in a certain consecutive order, we compiled a "symbolic" TPM, whose diagonal element $T_{mm}^{(sym)}$ were set equal to the number of non-consecutive occurrences of pattern m, and off-diagonal element $T_{mn}^{(sym)}$ was set equal to the number of $n \to m$ transitions observed; finally each row was normalized: $T_{mn}^{(sym)} \to T_{mn}^{(sym)}/\sum_{l=1}^{N} T_{ml}$. E.g. the pattern sequence $1, 2, 3, 1, 2$ is in one-to-one correspondence to the symbolic TPM

$$\text{sequence} \quad [1,2,3,1] \leftrightarrow T^{(sym)}_{mn} = \begin{pmatrix} 0.67 & 0.33 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix} .$$

Sequence similarity was defined as the trial-averaged Pearson correlation between $T^{(sym)}$.

In the data, we define the overlaps q between N-dimensional vectors r_i and s_i describing inferred patterns as the correlation coefficient

$$q[r,s] = \frac{1}{N}\sum_{i=1}^{N} \frac{r^i s^i}{\sigma(r)\sigma(s)} ,$$

where $\sigma(r)$ is the standard deviation of r^i.

4.5 Single neuron multistability

To assess how single-neuron activity was modulated across different patterns, local (i.e., single-trial) firing rate estimates for neuron i given a pattern m were obtained from the maximization step of the Baum-Welch algorithm

$$r_i(m) = -\frac{1}{dt}\log\left(1 - \frac{\sum_{t=1}^{T} p(m,t) y_i(t)}{\sum_{t=1}^{T} p(m,t)}\right) , \tag{4.1}$$

where $y_i(t)$ are the neuron's observations in the current trial of length T. To determine whether a neuron's conditional firing rate distributions differed across patterns (Fig. 3c), we performed a non-parametric one-way ANOVA (unbalanced Kruskal-Wallis, p<0.05). A post-hoc multiple-comparison rank analysis (with Bonferroni correction) revealed the smallest number of significantly different firing rate distributions across patterns. Given a p value p_{mn} for the pairwise post hoc comparison between patterns m and n, we considered the symmetric $M \times M$ matrix S with elements $S_{mn} = 0$ if the rates were different ($p_{mn} < 0.05$) and $S_{mn} = 1$ otherwise. For example, consider the case of 3 patterns and the following A matrix, where patterns were sorted by firing rates:

$$S = \begin{pmatrix} \cdot & 1 & 0 \\ \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot \end{pmatrix} .$$

Firing rates of patterns 1 and 2 were not significantly different, but they were different from pattern 3 firing rate. Hence, in this case we classified the neurons as multistable with 2 different firing rates across patterns [12].

4.6 Tagging pattern onsets to self-initiated actions

The HMM analysis yields a posterior probability distribution $p(m,t)$ for the neural pattern m at time t. At any time $\bar{t}$ we identified the active pattern $\bar{m}$ when $p(\bar{m},\bar{t}) \geq 0.8$. In case this criterion was not met by any pattern then no pattern was assigned, cfr Fig. 1d. The onset time of a specific pattern $\bar{m}$ was identified as the first time $\bar{t}$ where $p(\bar{m}, \bar{t}) \geq 0.8$. Transitions of several patterns appeared in close proximity to specific events Fig. 4a, we thus developed a method to tag pattern onsets to specific events. Specifically we tagged onset of a given pattern with one of three actions (Poke In, Poke Out, Water Poke In, respectively, for poking in and out of the Wait port and poking in to the Reward port) with the following procedure. For each session we analyzed all correct trials. We first realigned trials to the specific event recomputing the times of occurrences of all pattern onsets with respect to the event. In each session we analyzed all transitions to patterns which occurred in at least 70% of correct trials. This returned a distribution of times $\mathcal{T}(\bar{m})$ for the onset times of pattern $\bar{m}$. If the average of the distribution $\mu(\mathcal{T}(\bar{m})) \in [-0.5, 0.1]sec$, we tagged the pattern $\bar{m}$ to the event. In case multiple transitions matched our criteria, we selected the one with minimum inter-quartile $iqr(\mathcal{T}(\bar{m}))$. This procedure returned patterns tagged with specific actions for each trial, cfr. Fig. 4c. We name pattern onset times $\{t_{PI}, t_{PO}, t_{WPI}\}$ respectively for the actions Poke In, Poke Out and Water Poke In.

4.7 Decoding actions from pattern onsets

We reversed the pattern tagging procedure to decode actions from pattern onsets. Transitions were tagged to actions using correct trials (training set) using the procedure above, then actions were decoded from pattern onsets using incorrect trials (test set). The decoding procedure follows these steps: for every trial, given an action time $t_{\bar{e}}$ and the tagged pattern onset times $\{t_{PI}, t_{PO}, t_{WPI}\}$, we classified the action according to

$$\text{action} = \underset{e \in \{PI, PO, WPI\}}{\text{argmin}} (t_e - t_{\bar{e}}) \text{ if } (t_{action} - t_{\bar{e}}) \in [-0.5, 0.1] \,\text{sec} \ .$$

In case no patterns passed this criteria the action was not labelled. This procedure labelled 63% of all actions. For each session and all tagged actions we estimated a confusion matrix of our decoding procedure (cfr Fig. 4c) by comparing the true actions (rows of the confusion matrix) with their predicted labels (columns of the confusion matrix). The confusion matrix across all sessions was obtained by averaging confusion matrices for individual sessions.

4.8 Noise correlation analysis

To assess trial-to-trial variability in population activity we measured the neural dimensionality of population activity fluctuations around each pattern. We first estimated the noise covariance $C_{ij}(m)$, namely, the covariance conditioned on intervals where pattern m occurred (the time window with posterior probability $\geq 80\%$ in each trial):

$$C(m)_{ij} = \frac{1}{N_T} \sum_{a}^{N_T} ({r_i}^a(m){r_j}^{a,T}(m) - {r_i}^a(m){r_j}^{a,T}(m)) \ , \qquad (4.2)$$

where N_T is the number of trials in the session and $i, j = 1, \ldots, N$ index neurons. The superscript ^{T} denotes vector transposition. In each trial a and window the average firing rate

${r_i}^a(m)$ in pattern m was computed from Eq. (4.1). We then computed the dimensionality $d(m)$ of population activity fluctuations around pattern m as the participation ratio [26, 95]:

$$d(m) = \frac{Tr[C(m)]^2}{Tr[C(m)^2]} = \frac{(\sum_i^N \lambda_i)^2}{\sum_i^N \lambda_i^2} \tag{4.3}$$

where λ_i are the eigenvectors of the covariance matrix for $i = 1, \ldots, N$ neurons [26, 95]. This measure is bounded by the ensemble size N and captures the number of directions, in neural space, across which variability is spread over.

To test the hypothesis that trial-to-trial variability is constrained within a lower dimensional subspace, we proceeded as follows. For each neural pattern m, we considered the first K Principal Components $\{PC_1, \ldots, PC_K\}_m$ of $C(m)$ in Eq. (4.2), where K is the integer minor or equal to the average of $PR(m)$ across the M patterns within each session: $K = \text{floor}\left(\frac{1}{M}\sum_{m=1}^M PR_m\right)$. This represents the across-patterns average dimensionality of noise correlations within a session. Using a Canonical Correlation Analysis we then estimated the canonical variables between $\{PC_1, \ldots, PC_K\}_{m_1}$ and $\{PC_1, \ldots, PC_K\}_{m_2}$ for pairs of patterns m_1 and m_2, obtaining the respective correlation coefficients ρ_j between the K canonical variables, $j \in \{1..K\}$. Alignment $A(m_1, m_2)$ was then defined as the average correlation coefficient between the canonical variables $A(m_1, m_2) = \frac{1}{K}\sum_{j=K}^N \rho_j$, cf. Fig. 6b.

4.9 Network model

In this section we describe the correlated variability model generating reliable sequences of metastable attractors (see Eq. (2.1)), whose dynamics is ruled by the current-based formulation of the standard rate model [96, 97]:

$$\tau \dot{u}_i(t) = -u_i(t) + \sum_{j=1}^N J_{ij}^S \phi_j(u_j(t)) + \zeta(t) \sum_{j=1}^N J_{ij}^F \phi_j(u_j(t)) \ . \tag{4.4}$$

The firing rates are analog positive variables given by the transformation of synaptic currents to rates by the input-output transfer function $\phi_i(u_i)$. Tranfer functions ϕ_i was inferred from the empirical firing rate distribution of M2 single neurons (see below section 4.11). The parameter τ corresponds to the single neuron time constant. We set the M2 symmetric connectivity to be sparse [98–102]. Our connectivity consists of two terms, traditionally referred to as the *symmetric* term J_{ij}^S and the *asymmetric* or *feedforward* term J_{ij}^F [103]. The symmetric term reads

$$J_{ij}^S = \frac{c_{ij} A_S}{Nc} \sum_{\mu=1}^p f\left[\eta_i^\mu\right] g\left[\eta_j^\mu\right] \ , \tag{4.5}$$

where the variable c_{ij} represents the structural connectivity of the local circuit, modeled as an Erdos-Renyi graph where $c_{ij} = 1$ with probability c. The normalization constant Nc correspond to the average number of connections to a neuron; A_S is the overall strength of the symmetric term. The input to neuron i corresponding to pattern μ was an i.i.d.

standardized normal random variable, $z_i^\mu \overset{i.i.d.}{\sim} N(0,1)$. Firing rate patterns were then distributed as $\eta_i^\mu \overset{i.i.d.}{\sim} \phi(z_i^\mu)$, giving the symmetric term in the connectivity matrix in Eq. (4.5). This term represents a generalization of the covariance rule [104] in which changes in the connectivity by learnig are the product of the nonlinar transformation of pre- and post- synaptic activity, i.e., $\Delta J_{ij}^S \propto f[\eta_i^\mu]g[\eta_j^\mu]$. As is shown in [38] this term may lead to fixed-point attractors in this network. The functions f and g provide the dependence of the learning rule on the post- and pre-synaptic firing rates, respectively, and they control the firing rate statistics of the attractor. While J_{ij}^S is symmetric only if $f = g$, we choose to keep the terminology 'symmetric' for this term for consistency with early work in networks of binary neurons [27, 28, 103, 105]. The functions f and g are given by the step functions

$$f(\eta) = \begin{cases} q_f & \text{if } x_f \leq \eta \\ -(1-q_f) & \text{if } \eta \leq x_f \end{cases} , \qquad g(\eta) = \begin{cases} q_g & \text{if } x_g \leq \eta \\ -(1-q_g) & \text{if } \eta \leq x_g \end{cases} , \tag{4.6}$$

potentiating the post- (pre-) synaptic activity with strength $q_f \in [0,1]$ (q_g), and depressing with strength $1 - q_f$ ($1 - q_g$), respectively. The parameter x_f (x_g) represents the threshold between potentiation and depression for the post- (pre-) synaptic dependence of the learning rule, controlling the sparseness of the nonlinar transformation of the pattern $f(\vec{\eta})$ ($g(\vec{\eta})$) imprinted in the connectivity. We assume that the average synaptic weights changes due to learning one pattern is zero, requiring $\langle g \rangle = 0$, which constrains one of the two parameters of g. The asymmetric term in Eq. (4.4) is the *correlated variability* term term $\zeta(t)\sum_{j=1}^N J_{ij}^F \phi(u_j(t))$, where the rank p of the matrix J_{ij}^F is much lower than the number of neurons N in the network. Hence, this term induces low-dimensional correlated fluctuations across neurons, driven by the Ornstein-Uhlenbeck process $\zeta(t)$:

$$\tau_\zeta \dot{\zeta}(t) = -\zeta(t) + \bar{\zeta} + \sqrt{2\sigma_\zeta^2 \tau_\zeta} x(t) , \tag{4.7}$$

where τ_ζ, $\bar{\zeta}$ and σ_ζ^2 are the timescale, mean and variance of the process, respectively. The variable $x(t)$ represents white noise with zero mean and unit variance. For a derivation of these parameters see the next section.

In Section 2.7, we compared the correlated variability model (see Eq. (4.4)) to a *private noise* model

$$\tau \dot{u}_i(t) = -u_i(t) + \sum_{j=1}^N J_{ij}^S \phi_j(u_j(t)) + \bar{\zeta} \sum_{j=1}^N J_{ij}^F \phi_j(u_j(t)) + \sqrt{2\sigma_p^2 \tau} \chi_i(t) , \tag{4.8}$$

where term $\sqrt{2\sigma_p^2 \tau}\chi_i(t)$ is additive white Gaussian noise with mean zero and variance σ_p representing *private noise*, independently drawn for each neuron. Here, the asymmetric part of the synaptic couplings is constant, proportional to the parameter $\bar{\zeta}$, unlike the time varying asymmetric term in Eq. (4.4).

As a measure of the pattern retrieval (Fig. 5), we used overlaps, defined as the Pearson correlation between the instantaneous firing rate and a given stored pattern $g[\vec{\eta}^l]$ [38, 61]

$$m_l(t) = \frac{Cov\left[g[\vec{\eta}^l]\vec{r}(t)\right]}{\sqrt{Var(g[\vec{\eta}^l])Var(\vec{r}(t))}} . \tag{4.9}$$

4.10 Two-area mesoscale model

In this section, we show how to obtain the network model in Eq. (2.1), starting from the two-area network in Eq. (2.2), whose dynamics are governed by [96, 97]:

$$\begin{aligned} \tau \dot{u}_i(t) &= -u_i(t) + \sum_{j=1}^{N} J^S_{ij}\phi_j(u_j(t)) + \sum_{j=1}^{N_Y} W^{M2\leftarrow Y}_{ij} r^Y_j \ , \\ \tau_Y \dot{r}^Y_i &= -r^Y_i + \sum_{j=1}^{N} W^{Y\leftarrow M2}_{ij}\phi_j(u_j) \ , \end{aligned} \tag{4.10}$$

The first equation describes the local dynamics of N neurons' in area M2, with notations as in Eq. (4.4). The second term represents the activity r^Y_i of N_Y neurons in area Y, where we assume $N_Y \ll N$. We approximate the dynamics of the subcortical area as linear, where $W^{Y\leftarrow M2}_{ij}$ are the projections from M2 to area Y, structured by Hebbian learning (see previous paragraph) as

$$W^{Y\leftarrow M2}_{ij} = s^{Y\leftarrow M2}_{ij}(t)\frac{1}{N}\sum_{l=1}^{p} y^\mu_i g(\eta^\mu_j). \tag{4.11}$$

Here, $g(\eta^\mu)$ is the pre-synaptic dependence of the learning rule; y^μ_i is the post synaptic dependence of the learning rule, which depends on the activity in area Y. Additionally, $s^{Y\leftarrow M2}_{ij}(t)$ represents the synaptic efficacy in the Y to M2 projections due to short term plasticity [40]. We consider a simplified phenomenological model capturing the temporal fluctuations in the synaptic efficacy due to STP given by

$$\dot{s}^{Y\leftarrow M2}_{ij} = \frac{1 - s^{Y\leftarrow M2}_{ij}(t)}{\tau_{STP}} + \sqrt{\frac{2\alpha_{Y\leftarrow M2}}{\tau_{STP}}}\xi^{Y\leftarrow M2}_i(t), \tag{4.12}$$

where τ_{STP} corresponds to the time-scale of the fluctuations in the synaptic efficacy, $\alpha^2_{Y\leftarrow M2}$ is the variance of these fluctuations, and $\xi^{Y\leftarrow M2}_i(t)$ is a Gaussian random variable with mean zero and variance one. Here we assume that changes in the synaptic efficacy depend on post-synaptic fluctuations given by the variable $\xi^{Y\leftarrow M2}_i(t)$. We assume the activity of area Y is fast with respect to M2 ($\tau_Y < \tau$), replacing Eq. (4.10) by its steady state

$$r^Y_i = \sum_{j=1}^{N} W^{Y\leftarrow M2}_{ij}\phi(u_j) \ . \tag{4.13}$$

Feedback projections $\sum_{j=1}^{N_Y} W^{M2\leftarrow Y}_{ij} r^Y_j$ from area Y to M2 in Eq. (4.10) are shaped by Hebbian learning as above, obtaining

$$W^{M2\leftarrow Y}_{ij} = s^{M2\leftarrow Y}_{ij}(t)\frac{1}{N_Y}\sum_{\mu=1}^{p} f(\eta^{\mu+1}_i)y^\mu_j \ . \tag{4.14}$$

Similarly to the M2→Y, we assume that the Y→M2 projections also undergo short-term synaptic plasticity, i.e.,

$$\dot{s}^{M2\leftarrow Y}_{ij} = \frac{1 - s^{M2\leftarrow Y}_{ij}(t)}{\tau_{STP}} + \sqrt{\frac{2\alpha_{M2\leftarrow Y}}{\tau_{STP}}}\xi^{M2\leftarrow Y}_j(t). \tag{4.15}$$

Here, we assume that changes on synaptic efficacy of the Y→M2 projections depend on pre-synaptic fluctuations given by the variable $\xi_j^{M2\leftarrow Y}(t)$. Synaptic delays in the feedback loop are long enough so that, while the pre-synaptic activity of the feedback projections correspond to the pattern η^μ, the post-synaptic activity of the feedback projections is $\eta^{\mu+1}$. Then, the input current due to the feedback loop between M2 and area Y is approximately

$$\begin{aligned}\sum_{l=1}^{N_Y} W_{il}^{Y\to M2} r_l^Y &= \sum_{j=1}^{N}\sum_{l=1}^{N_Y} W_{il}^{M2\leftarrow Y} W_{lj}^{Y\leftarrow M2}\phi(u_j)\\ &= \frac{1}{N}\left(\bar{\zeta} + \frac{\sigma p}{\sqrt{N_Y}}x(t)\right)\sum_{j=1}^{N}\sum_{\mu=1}^{p} f(\eta_i^{\mu+1})g(\eta_j^\mu)\phi(u_j). \qquad (4.16)\end{aligned}$$

Here, we used the fact that $\frac{1}{N_Y}\sum_{l=1}^{N_Y} y_l^\mu y_l^{\mu'} s_{il}^{M2\leftarrow Y}(t) s_{lj}^{Y\leftarrow M2}(t)$ has mean $\bar{\zeta}\delta_{\mu,\mu'}$ and finite variance σ^2, when averaged over an ensemble of $\langle\cdots\rangle_{y,\xi(t)}$ of patterns y and fluctuations of the synaptic efficacies $\xi(t)$ in Eqs. (4.12) and (4.15); $x(t)$ represents white noise with zero mean and unit variance. Therefore, the matrix J_{ij}^F in Eq. (2.1) corresponds to the effective connectivity arising from the feedback loop between M2 and area Y:

$$J_{ij}^F = \frac{\zeta(t)}{N}\sum_{\mu=1}^{p} f(\eta_i^{\mu+1})g(\eta_j^\mu), \qquad (4.17)$$

which has rank $p \ll N$. Assuming $p \sim \mathcal{O}(\sqrt{N_y}) \ll N$, the fluctuations due to STP are order 1. We account for both the strength and the variability in the M2→Y and Y→M2 projections, via the Ornstein-Uhlenbeck process $\zeta(t)$ in Eq. (4.7). Notice that τ_ζ is the effective time-scale of the temporal fluctuations in the sum over Y neurons in Eq. (4.16). Its mean $\bar{\zeta}$ and variance σ_ζ^2 control, respectively, the strength and the variability of the effective feedforward couplings obtained after integrating out the dynamics in area Y. The variance σ_ζ^2 is inversely proportional to the size of the neural population in area Y.

4.11 Inferring the transfer function from data

For inferring the input-output transfer function from *in vivo* recordings we adapted a method proposed in [37] to our hidden Markov model analysis. Briefly, for each session, the empirical distribution of mean firing rates across patterns and neurons is constructed. As in [37], we assumed normally distributed synaptic input currents. By rank-matching the firing rates to a standardized normal distribution we obtained the empirical current-to-rate transfer function (see S5). Similarly to [38], for each recorded unit we fit this curve with a sigmoidal function

$$\phi(u) = \frac{R_0}{1 + e^{-\beta(u-h_0)}}. \qquad (4.18)$$

If input currents produced firing rates in Eq. (4.18) larger than a neuron's maximal firing rate R_{max}, then the correspding firing rates were set to R_{max}. Using the above procedure we inferred a distribution of parameters $\{(R_{max}^{(i)}, R_0^{(i)}, \beta^{(i)}, h_0^{(i)})\}_{i=1}^{366}$, one from each recorded unit (Fig. 5a). For conveying the diversity in the transfer functions inferred from data, in our model we randomly sampled with replacement 10000 samples from the parameter distribution above.

Model parameters		
Parameter	Value	comment
c	0.1	connectivity sparsity
N	10000	network size
q_f	.65	potentiation offset for f
x_f	1.7	potentiation/depression threshold for f
x_g	1.7	potentiation/depression threshold for g
A_S	3	strength of the symmetric connectivity
$\bar{\zeta}$	0.65	mean of the synaptic noise
σ_ζ	0.65	standard deviation of the synaptic noise
τ_ζ	20ms	synaptic noise time constant
σ_p	0	standard deviation of the private noise
τ	20ms	single neuron time constant

Table S1. Network parameters.

4.12 Network simulations

For the network simulations of the correlated variability model in Figs. 5 and 6, the parameter values used are listed in Table S1. The number of sessions and trials per session are matched to those in the empirical data. The number of attractors in each session, e.g., p (see Eq. (4.5)) are taken to be the same as the number of patterns inferred in each empirical session using the HMM. An attractor was detected in the model when the overlap between network activity and the attractor is larger than 0.4.

For the network simulations of the private noise model in Fig. S6 the parameter are the same as in Table S1 except $A_S = 1$, $\sigma_\zeta = 0$, and $\sigma_p = 0.2$. An attractor was detected in the model when the overlap between network activity and the attractor is larger than 0.2.

The simulations where performed using custom Python scripts. The code is available upon request.

Acknowledgments

We would like to thank F. Cazettes, N. Steinmetz for discussions. We thank in particular L. Logiaco, J. M. Murray and N. Brunel for suggestions on the network model. This work was supported by a National Institute of Deafness and Other Communication Disorders Grant K25-DC013557 (L.M.), by the Swartz Foundation Award 66438 (L.M.).

Author contributions L.M. and Z.M. designed the project; M.M. and Z.M. performed the experiment; S.R. and L.M. analyzed the data; U.P. and S.R. developed the theoretical model with help from L.M.; L.M., S.R., and U.P. wrote the manuscript with input from M.M. and Z.M.; L.M. supervised the project.

References

[1] Gordon J Berman, William Bialek, and Joshua W Shaevitz. Predictability and hierarchy in drosophila behavior. *Proceedings of the National Academy of Sciences*, 113(42):11943–11948, 2016.

[2] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

[3] Jeffrey E Markowitz, Winthrop F Gillis, Celia C Beron, Shay Q Neufeld, Keiramarie Robertson, Neha D Bhagat, Ralph E Peterson, Emalee Peterson, Minsuk Hyun, Scott W Linderman, et al. The striatum organizes 3d behavior via moment-to-moment action selection. *Cell*, 2018.

[4] Peter R Killeen and J Gregor Fetterman. A behavioral theory of timing. *Psychological review*, 95(2):274, 1988.

[5] Scott W Linderman, Annika LA Nichols, David M Blei, Manuel Zimmer, and Liam Paninski. Hierarchical recurrent state space models reveal discrete and continuous dynamics of neural activity in c. elegans. *bioRxiv*, page 621540, 2019.

[6] Masayoshi Murakami, M Inês Vicente, Gil M Costa, and Zachary F Mainen. Neural antecedents of self-initiated actions in secondary motor cortex. *Nature neuroscience*, 17(11):1574, 2014.

[7] Masayoshi Murakami, Hanan Shteingart, Yonatan Loewenstein, and Zachary F Mainen. Distinct sources of deterministic and stochastic components of action timing decisions in rodent frontal cortex. *Neuron*, 94(4):908–919, 2017.

[8] Nuo Li, Kayvon Daie, Karel Svoboda, and Shaul Druckmann. Robust neuronal dynamics in premotor cortex during motor planning. *Nature*, 532(7600):459, 2016.

[9] Hidehiko K Inagaki, Lorenzo Fontolan, Sandro Romani, and Karel Svoboda. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature*, 566(7743):212, 2019.

[10] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[11] L. M. Jones, A. Fontanini, B. F. Sadacca, P. Miller, and D. B. Katz. Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proc Natl Acad Sci U S A*, 104(47):18772–7, 2007.

[12] Luca Mazzucato, Alfredo Fontanini, and Giancarlo La Camera. Dynamics of multistable states during ongoing and evoked cortical activity. *The Journal of Neuroscience*, 35(21):8214–8231, 2015.

[13] Kourosh Maboudi, Etienne Ackermann, Laurel Watkins de Jong, Brad E Pfeiffer, David Foster, Kamran Diba, and Caleb Kemere. Uncovering temporal structure in hippocampal output patterns. *eLife*, 7:e34467, 2018.

[14] Giancarlo La Camera, Alfredo Fontanini, and Luca Mazzucato. Cortical computations via metastable activity. *arXiv preprint arXiv:1906.07777*, 2019.

[15] Itay Gat and Naftali Tishby. Statistical modeling of cell assemblies activities in associative

cortex of behaving monkeys. In *Advances in neural information processing systems*, pages 945–952, 1993.

[16] M. Abeles, H. Bergman, I. Gat, I. Meilijson, E. Seidemann, N. Tishby, and E. Vaadia. Cortical activity flips among quasi-stationary states. *Proc Natl Acad Sci USA*, 92:8616–8620, 1995.

[17] A. Ponce-Alvarez, V. Nacher, R. Luna, A. Riehle, and R. Romo. Dynamics of cortical neuronal ensembles transit from decision making to storage for later report. *J Neurosci*, 32(35):11956–69, 2012.

[18] Tatiana A Engel, Nicholas A Steinmetz, Marc A Gieselmann, Alexander Thiele, Tirin Moore, and Kwabena Boahen. Selective modulation of cortical state during spatial attention. *Science*, 354(6316):1140–1144, 2016.

[19] Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721, 2017.

[20] Jalil Taghia, Weidong Cai, Srikanth Ryali, John Kochalka, Jonathan Nicholas, Tianwen Chen, and Vinod Menon. Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nature communications*, 9(1):2505, 2018.

[21] P. Miller and D. B. Katz. Stochastic transitions between neural states in taste processing and decision-making. *J Neurosci*, 30(7):2559–70, 2010.

[22] G. Deco and E. Hugues. Neural network mechanisms underlying stimulus driven variability reduction. *PLoS Comput Biol*, 8(3):e1002395, 2012.

[23] A. Litwin-Kumar and B. Doiron. Slow dynamics and high variability in balanced cortical networks with clustered connections. *Nat Neurosci*, 15(11):1498–505, 2012.

[24] D. J. Amit and N. Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb Cortex*, 7(3):237–52, 1997.

[25] Luca Mazzucato, Giancarlo La Camera, and Alfredo Fontanini. Expectation-induced modulation of metastable activity underlies faster coding of sensory stimuli. *Nature neuroscience*, page 1, 2019.

[26] Luca Mazzucato, Alfredo Fontanini, and Giancarlo La Camera. Stimuli reduce the dimensionality of cortical activity. *Frontiers in systems neuroscience*, 10:11, 2016.

[27] Haim Sompolinsky and I Kanter. Temporal association in asymmetric neural networks. *Physical review letters*, 57(22):2861, 1986.

[28] David Kleinfeld. Sequential state generation by model neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 83(24):9469, 1986.

[29] Ulises Pereira and Nicolas Brunel. Unsupervised learning of persistent and sequential activity. *Frontiers in Computational Neuroscience*, 13:97, 2020.

[30] KuangHua Guo, Naoki Yamawaki, Karel Svoboda, and Gordon MG Shepherd. Anterolateral motor cortex connects with a medial subdivision of ventromedial thalamus through cell type-specific circuits, forming an excitatory thalamo-cortico-thalamic loop via layer 1 apical tuft dendrites of layer 5b pyramidal tract type neurons. *Journal of Neuroscience*, 38(41):8787–8797, 2018.

[31] Zengcai V Guo, Hidehiko K Inagaki, Kayvon Daie, Shaul Druckmann, Charles R Gerfen, and Karel Svoboda. Maintenance of persistent activity in a frontal thalamocortical loop. *Nature*, 545(7653):181, 2017.

[32] Sébastien Hélie, Shawn W Ell, and F Gregory Ashby. Learning robust cortico-cortical associations with the basal ganglia: an integrative review. *Cortex*, 64:123–135, 2015.

[33] Michel Desmurget and Robert S Turner. Motor sequences and the basal ganglia: kinematics, not habits. *Journal of Neuroscience*, 30(22):7685–7690, 2010.

[34] Miho Nakajima, L Ian Schmitt, and Michael M Halassa. Prefrontal cortex regulates sensory filtering through a basal ganglia-to-thalamus pathway. *Neuron*, 2019.

[35] Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature neuroscience*, 17(10):1410, 2014.

[36] Crispin W Gardiner et al. *Handbook of stochastic methods*, volume 3. springer Berlin, 1985.

[37] Sukbin Lim, Jillian L McKee, Luke Woloszyn, Yali Amit, David J Freedman, David L Sheinberg, and Nicolas Brunel. Inferring learning rules from distributions of firing rates in cortical neurons. *Nature neuroscience*, 18(12):1804, 2015.

[38] Ulises Pereira and Nicolas Brunel. Attractor dynamics in networks with learning rules inferred from in vivo data. *Neuron*, 99(1):227–238, 2018.

[39] L Ian Schmitt, Ralf D Wimmer, Miho Nakajima, Michael Happ, Sima Mofakham, and Michael M Halassa. Thalamic amplification of cortical connectivity sustains attentional control. *Nature*, 545(7653):219, 2017.

[40] Misha Tsodyks, Klaus Pawelzik, and Henry Markram. Neural networks with dynamic synapses. *Neural computation*, 10(4):821–835, 1998.

[41] Kimberly Reinhold, Anthony D Lien, and Massimo Scanziani. Distinct recurrent versus afferent dynamics in cortical visual processing. *Nature neuroscience*, 18(12):1789, 2015.

[42] Jorge Jaramillo, Jorge F Mejias, and Xiao-Jing Wang. Engagement of pulvino-cortical feedforward and feedback pathways in cognitive computations. *Neuron*, 101(2):321–336, 2019.

[43] Christopher J Cueva, Encarni Marcos, Alex Saez, Aldo Genovesio, Mehrdad Jazayeri, Ranulfo Romo, C Daniel Salzman, Michael N Shadlen, and Stefano Fusi. Delay activity dynamics: task dependent time encoding and low dimensional trajectories. *bioRxiv*, page 504936, 2018.

[44] Joaquin M Fuster and John P Jervey. Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science*, 212(4497):952–955, 1981.

[45] Yasushi Miyashita. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335(6193):817, 1988.

[46] Joaquin M Fuster and Garrett E Alexander. Neuron activity related to short-term memory. *Science*, 173(3997):652–654, 1971.

[47] Shintaro Funahashi, Charles J Bruce, and Patricia S Goldman-Rakic. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of neurophysiology*, 61(2):331–349, 1989.

[48] Jeffrey C Erlich, Max Bialek, and Carlos D Brody. A cortical substrate for memory-guided orienting in the rat. *Neuron*, 72(2):330–343, 2011.

[49] James H Marshel, Yoon Seok Kim, Timothy A Machado, Sean Quirin, Brandon Benson, Jonathan Kadmon, Cephra Raja, Adelaida Chibukhchyan, Charu Ramakrishnan, Masatoshi Inoue, et al. Cortical layer–specific critical dynamics triggering perception. *Science*, 365(6453):eaaw5202, 2019.

[50] Ziqiang Wei, Hidehiko Inagaki, Nuo Li, Karel Svoboda, and Shaul Druckmann. An orderly single-trial organization of population dynamics in premotor cortex predicts behavioral variability. *Nature communications*, 10(1):216, 2019.

[51] D. J. Amit and N. Brunel. Dynamics of a recurrent network of spiking neurons before and following learning. *Network: Comput. Neural Syst.*, 8:373–404, 1997.

[52] E. Seidemann, I. Meilijson, M. Abeles, H. Bergman, and E. Vaadia. Simultaneously recorded single units in the frontal cortex go through sequences of discrete and stable states in monkeys performing a delayed localization task. *J Neurosci*, 16(2):752–68, 1996.

[53] M. Abeles. *Corticonics*. New York: Cambridge University Press, 1991.

[54] M. Diesmann, M. O. Gewaltig, and A. Aertsen. Stable propagation of synchronous spiking in cortical neural networks. *Nature*, 402:529–533, 1999.

[55] Richard HR Hahnloser, Alexay A Kozhevnikov, and Michale S Fee. An ultra-sparse code underliesthe generation of neural sequences in a songbird. *Nature*, 419(6902):65, 2002.

[56] Ila R Fiete, Walter Senn, Claude ZH Wang, and Richard HR Hahnloser. Spike-time-dependent plasticity and heterosynaptic competition organize networks to produce long scale-free sequences of neural activity. *Neuron*, 65(4):563–576, 2010.

[57] Zoltán Nádasdy, Hajime Hirase, András Czurkó, Jozsef Csicsvari, and György Buzsáki. Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience*, 19(21):9497–9507, 1999.

[58] Stanislas Dehaene, Jean-Pierre Changeux, and Jean-Pierre Nadal. Neural networks that learn temporal sequences by selection. *Proceedings of the National Academy of Sciences*, 84(9):2727–2731, 1987.

[59] Alessandro Treves. Frontal latching networks: a possible neural basis for infinite recursion. *Cognitive neuropsychology*, 22(3-4):276–291, 2005.

[60] James M Murray et al. Learning multiple variable-speed sequences in striatum via cortical tutoring. *Elife*, 6:e26084, 2017.

[61] Maxwell Gillett, Ulises Pereira, and Nicolas Brunel. Characteristics of sequential activity in networks with temporally asymmetric hebbian learning. *bioRxiv*, page 818773, 2019.

[62] Joseph K Jun and Dezhe Z Jin. Development of neural circuitry for precise temporal sequences through spontaneous activity, axon remodeling, and synaptic plasticity. *PloS one*, 2(8):e723, 2007.

[63] Jian K Liu and Dean V Buonomano. Embedding multiple trajectories in simulated recurrent neural networks in a self-organizing manner. *Journal of Neuroscience*, 29(42):13172–13181, 2009.

[64] Kanaka Rajan, Christopher D Harvey, and David W Tank. Recurrent network models of sequence generation and memory. *Neuron*, 90(1):128–142, 2016.

[65] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[66] Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Spin-glass models of neural networks. *Physical Review A*, 32(2):1007, 1985.

[67] Gianluigi Mongillo, Daniel J Amit, and Nicolas Brunel. Retrospective and prospective persistent activity induced by hebbian learning in a recurrent cortical network. *European Journal of Neuroscience*, 18(7):2011–2024, 2003.

[68] Ran Darshan, William E Wood, Susan Peters, Arthur Leblois, and David Hansel. A canonical neural mechanism for behavioral variability. *Nature communications*, 8:15415, 2017.

[69] Jeremy Bernstein, Ishita Dasgupta, David Rolnick, and Haim Sompolinsky. Markov transitions between attractor states in a recurrent neural network. In *2017 AAAI Spring Symposium Series*, 2017.

[70] Risa Kawai, Timothy Markman, Rajesh Poddar, Raymond Ko, Antoniu L Fantana, Ashesh K Dhawale, Adam R Kampff, and Bence P Ölveczky. Motor cortex is required for learning but not for executing a motor skill. *Neuron*, 86(3):800–812, 2015.

[71] Karel Svoboda and Nuo Li. Neural mechanisms of movement planning: motor cortex and beyond. *Current opinion in neurobiology*, 49:33–41, 2018.

[72] Xin Jin and Rui M Costa. Shaping action sequences in basal ganglia circuits. *Current opinion in neurobiology*, 33:188–196, 2015.

[73] Mimi H Kao, Allison J Doupe, and Michael S Brainard. Contributions of an avian basal ganglia–forebrain circuit to real-time modulation of song. *Nature*, 433(7026):638, 2005.

[74] Nuo Li, Tsai-Wen Chen, Zengcai V Guo, Charles R Gerfen, and Karel Svoboda. A motor cortex circuit for motor planning and movement. *Nature*, 519(7541):51, 2015.

[75] Tsai-Wen Chen, Nuo Li, Kayvon Daie, and Karel Svoboda. A map of anticipatory activity in mouse motor cortex. *Neuron*, 94(4):866–879, 2017.

[76] Jung Hoon Sul, Suhyun Jo, Daeyeol Lee, and Min Whan Jung. Role of rodent secondary motor cortex in value-based action selection. *Nature neuroscience*, 14(9):1202, 2011.

[77] Zengcai V Guo, Nuo Li, Daniel Huber, Eran Ophir, Diego Gutnisky, Jonathan T Ting, Guoping Feng, and Karel Svoboda. Flow of cortical activity underlying a tactile decision in mice. *Neuron*, 81(1):179–194, 2014.

[78] Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560, 2002.

[79] Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *science*, 304(5667):78–80, 2004.

[80] David Sussillo and Larry F Abbott. Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544–557, 2009.

[81] Francesca Mastrogiuseppe and Srdjan Ostojic. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, 2018.

[82] Cristopher M Niell and Michael P Stryker. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, 65(4):472–479, 2010.

[83] Pierre-Olivier Polack, Jonathan Friedman, and Peyman Golshani. Cellular mechanisms of brain state–dependent gain modulation in visual cortex. *Nature neuroscience*, 16(9):1331, 2013.

[84] Matthew J McGinley, Martin Vinck, Jacob Reimer, Renata Batista-Brito, Edward Zagha, Cathryn R Cadwell, Andreas S Tolias, Jessica A Cardin, and David A McCormick. Waking state: rapid variations modulate neural and behavioral responses. *Neuron*, 87(6):1143–1161, 2015.

[85] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D Harris. Spontaneous behaviors drive multidimensional, brain-wide population activity. *BioRxiv*, page 306019, 2018.

[86] Simon Musall, Matthew T Kaufman, Ashley L Juavinett, Steven Gluf, and Anne K Churchland. Single-trial neural dynamics are dominated by richly varied movements. *bioRxiv*, page 308288, 2019.

[87] David B Salkoff, Edward Zagha, Erin McCarthy, and David A McCormick. Movement and performance predict widespread cortical activity in a visual detection task. *bioRxiv*, page 709642, 2019.

[88] Marlene R Cohen and John HR Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594, 2009.

[89] Chengcheng Huang, Douglas A Ruff, Ryan Pyle, Robert Rosenbaum, Marlene R Cohen, and Brent Doiron. Circuit models of low-dimensional shared variability in cortical networks. *Neuron*, 101(2):337–348, 2019.

[90] Adrian G Bondy, Ralf M Haefner, and Bruce G Cumming. Feedback determines the structure of correlated variability in primary visual cortex. *Nature neuroscience*, 21(4):598, 2018.

[91] Jesse H Goldberg and Michale S Fee. Vocal babbling in songbirds requires the basal ganglia-recipient motor thalamus but not the basal ganglia. *Journal of neurophysiology*, 105(6):2729–2739, 2011.

[92] Alex H Williams, Tony Hyun Kim, Forea Wang, Saurabh Vyas, Stephen I Ryu, Krishna V Shenoy, Mark Schnitzer, Tamara G Kolda, and Surya Ganguli. Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron*, 98(6):1099–1115, 2018.

[93] Giorgio Tomasi and Rasmus Bro. A comparison of algorithms for fitting the parafac model. *Computational Statistics & Data Analysis*, 50(7):1700–1734, 2006.

[94] David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 120(5):386–391, 2013.

[95] L. F. Abbott, K. Rajan, and H. Sompolinsky. *Interactions between Intrinsic and Stimulus-Evoked Activity in Recurrent Neural Networks*, chapter 4. Oxford University Press, 2011.

[96] Stephen Grossberg. On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks. *journal of Statistical Physics*, 1(2):319–350, 1969.

[97] Kenneth D Miller and Francesco Fumarola. Mathematical equivalence of two common forms of firing rate models of neural networks. *Neural computation*, 24(1):25–31, 2012.

[98] Adrian Mason, Andrew Nicoll, and Ken Stratford. Synaptic transmission between individual pyramidal neurons of the rat visual cortex in vitro. *Journal of Neuroscience*, 11(1):72–84, 1991.

[99] Henry Markram, Joachim Lübke, Michael Frotscher, Arnd Roth, and Bert Sakmann. Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex. *The Journal of physiology*, 500(2):409–440, 1997.

[100] Carl Holmgren, Tibor Harkany, Björn Svennenfors, and Yuri Zilberter. Pyramidal cell communication within local networks in layer 2/3 of rat neocortex. *The Journal of physiology*, 551(1):139–153, 2003.

[101] Alex M Thomson and Christophe Lamy. Functional maps of neocortical local circuitry. *Frontiers in neuroscience*, 1:2, 2007.

[102] Sandrine Lefort, Christian Tomm, J-C Floyd Sarria, and Carl CH Petersen. The excitatory neuronal network of the c2 barrel column in mouse primary somatosensory cortex. *Neuron*, 61(2):301–316, 2009.

[103] E. Domany, J.Leo van. Hemmen, and K. Schulten. *Models of Neural Networks I*. Springer, 1995.

[104] Terrence J Sejnowski. Storing covariance with nonlinearly interacting neurons. *Journal of mathematical biology*, 4(4):303–321, 1977.

[105] A Herz, B Sulzer, R Kühn, and JL Van Hemmen. Hebbian learning reconsidered: Representation of static and dynamic objects in associative neural nets. *Biological cybernetics*, 60(6):457–467, 1989.